UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação semanticamente enriquecida por expressões do domínio

Ricardo Brigato Scheicher

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Ricardo Brigato Scheicher

Classificação semanticamente enriquecida por expressões do domínio

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
Outubro de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S318c

Scheicher, Ricardo Brigato
Classificação semanticamente enriquecida por expressões do domínio / Ricardo Brigato Scheicher; orientador Solange Oliveira Rezende. -- São Carlos, 2022.
259 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Classificação semântica de textos. 2. Representação semântica de textos. 3. Extração de termos. 4. Mineração de textos. I. Rezende, Solange Oliveira, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Ricardo Brigato Scheicher

Semantically enriched classification by domain expressions

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
October 2022

AGRADECIMENTOS

Agradeço acima de tudo à Deus, por me permitir estar aqui nesse momento e ser o suporte de todos os passos da minha vida.

Agradeço aos meus pais, minha mãe Magda Scheicher e ao meu falecido pai Jorge Scheicher, que me deram toda a base para chegar até aqui.

Agradeço à minha esposa Tânia Scheicher, companheira de todos os meus momentos, dores, angústias, cansaço, alegrias e comemorações. Meu braço direito de uma longa jornada.

Agradeço à minha orientadora, Profa. Dra. Solange Oliveira Rezende, por todos os ensinamentos, pela sua amizade e generosidade, pela oportunidade de pesquisa e pela confiança depositada em meu trabalho.

Agradeço também aos diversos colegas do curso de doutorado, que estiveram presentes em minha vida e, de alguma forma, fizeram parte do meu crescimento profissional. Dentre eles, o agradecimento especial é para Roberta Akemi Sinoara, Dildre Georgiana Vasques e Flávia Lemos Sampaio Xavier.

Agradeço a todos os professores e pesquisadores por todo empenho e dedicação nos ensinamentos.

Agradecimentos especiais são direcionados à Universidade de São Paulo, aos professores e funcionários do ICMC-USP, aos membros da banca do exame de qualificação e de defesa.

Agradeço ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, que apoiou financeiramente a realização desta pesquisa.

RESUMO

SCHEICHER, R. B. **Classificação semanticamente enriquecida por expressões do domínio**. 2022. 259 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Com o aumento da quantidade e variedade de textos em formato digital, seja em redes sociais, web ou internamente nas organizações, as técnicas de Mineração de Textos tornam-se essenciais no apoio à extração e organização desse conhecimento. Em tarefas de classificação de documentos, a representação dos dados tem um alto impacto na qualidade da solução final e modelos tradicionais de representação textual, como a *Bag-of-Words* (BoW), limitam-se apenas ao léxico impossibilitando a distinção de documentos com vocabulário semelhante e ideias diferentes sobre um mesmo assunto. Problemas de diferentes níveis de complexidade semântica possuem determinadas características que influenciam diretamente no desempenho de tarefas de classificação. Nesse sentido, o uso de informações emanticamente mais ricas em conjunto com a representação tradicional BoW permite atingir resultados mais eficazes em tarefas de Mineração de Textos. Expressões do domínio são consideradas informações enriquecidas que carregam consigo um certo nível semântico. A representação *generalized of Expressions of Domain (BoED)* é construída a partir e listas de termos do domínio e identificadores de classe, que geram as expressões do domínio e pode ser aplicada em diversas áreas do conhecimento como forma de informação semanticamente enriquecida. Com o propósito geral de avançar as pesquisas na área de Mineração de Textos e melhorar resultados de classificação de nível semâmantico usando informações enriquecidas, nesta tese de doutorado foram desenvolvidas e a avaliadas as seguintes abordagens: (i) proposta de três diferentes versões das representações enriquecidas semanticamente gBoED, (ii) método de classificação semanticamente enriquecida por expressões do domínio, (iii) método semiautomático de extração de termos e construção de representação semântica baseado em regras morfossintáticas, (iv) método semiautomático de extração de termos baseado em modelos de linguagem BERT, (v) estudo de caso de classificação semântica em pedidos de acesso à informação. Os métodos foram desenvolvidos e avaliados em dez coleções de documentos diferentes, em idioma português e inglês, juntamente com as diferentes versões de representações semanticamente enriquecidas. Os resultados indicam que os métodos propostos são promissores, possibilitando melhorar a acurácia de tarefas de classificação semântica em domínio restrito, quando comparada aos resultados com o método tradicional BoW.

Palavras-chave: Classificação semântica de textos, Representação semântica de textos, Extração de termos, Mineração de textos.

ABSTRACT

SCHEICHER, R. B. **Semantically enriched classification by domain expressions**. 2022. 259 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

With the increase in the amount and variety of texts in digital format, even on social networks, web, or in organizations, Text Mining techniques become essential to support the extraction of knowledge. In document classification tasks, data representation has a high impact on the quality of the final solution, and traditional models of textual representation, such as Bag-of-Words (BoW), are limited only to the lexicon making it impossible to the definition of documents with risk and different ideas on the same subject. Problems of different levels of semantic complexity have certain characteristics that directly influence the classification tasks' performance. In this sense, the semantically richer use of information in conjunction with a BoW representation allows for achieving more effective results in Text Mining tasks. Domain expressions are a type of enriched information that carries with it a certain semantic level. The *generalized of Expressions of Domain (BoED)* representation is built from domain terms lists and class identifiers lists, which generate domain expressions and can be applied in several areas of knowledge as a form of semantically enriched information. With the general purpose of advancing semantic-level Text Mining research and improving semantic-level classification results, this thesis has been developed and evaluated the following approaches: (i) Purpose of different versions of the semantically enriched representations gBoED, (ii) semantically enriched classification method by domain expressions, (iii) semiautomatic method of terms extraction and semantic representation construction based on morphosyntactic rules, (iv) semiautomatic method of terms extraction based on BERT language models, (v) case study of semantic classification in requests of access to information. The methods were developed and included in ten different document collections, in Portuguese and English, which can be presented as different versions of the semantically enriched representations. The results indicate that the purposed method is promising, improving accuracy results in semantic classification tasks when compared to the traditional method BoW.

Keywords: Text semantic classification, Text semantic representation, Terms extraction, Text mining.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	*Bidirectional Encoder Representation from Transformers*
BiLSTM	Rede Neural LSTM Bidirecional
BoED	*Bag of Expressions of Domain*
BoW	*Bag of Words*
CBOW	*Continuous bag-of-words*
CGU	Controladoria-Geral da União
e-SIC	Sistema Eletrônico de Serviço de Informação ao Cidadão
EAT	Extração automática de termos
gBoED	*generalized Bag of Expressions of Domain*
GDPR	*General Data Protection Regulation*
GloVe	*Global Vectors for Word Representation*
IMBHN	*Inductive Model based on Bipartite Heterogeneous Networks*
KNN	*k-Nearest Neighbors*
LAI	Lei de Acesso à Informação
LGPD	Lei Geral de Proteção de Dados
LSTM	*Long Short-Term Memory*
MLM	*Masked Language Model*
MRC	*Machine Reading Comprehension*
MT	Mineração de Textos
NB	*Naïve Bayes*
NSP	*Next Sentence Prediction*
PLN	Processamento de Linguagem Natural
REN	Reconhecimento de Entidades Nomeadas
RNC	Rede Neural Convolucional
RNR	Rede Neural Recorrente
RRC	*Review Reading Comprehension*
SVM	*Support Vector Machine*
TCU	Tribunal de Contas da União

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

O fenômeno *Big Data*, juntamente com a computação em nuvem é o que permitiu nos últimos anos um aumento constante na quantidade de dispositivos conectados à Internet e serviços disponíveis na rede. A maioria dos cidadãos em todo o mundo está ativa em várias mídias sociais, muitos deles com múltiplas contas em vários sites sociais, como Facebook, LinkedIn, Twitter e Instagram e realizam postagens de textos, fotos, áudios e vídeos diariamente de qualquer canto do mundo. YouTube e Netflix permite assistir a centenas de filmes em seu celular ou TV. Todos esses recursos têm gerado nos últimos anos um aumento expressivo na quantidade de dados criada e armazenada em todo o mundo (GUNDU; ANURADHA, 2020).

Em 2020, uma projeção realizada por Mishra *et al.* (2020) indica que em 2025 serão gerados 165 zettabytes de dados. Diariamente o site de busca da empresa Google realiza por volta de 1 bilhão de buscas, mais de 800 milhões de atualizações são feitas no Facebook e o Youtube recebe em torno de 4 bilhões de visualizações. Todo esse conteúdo é criado e acessado por milhares de pessoas em todo o mundo e, segundo Negi (2021), cerca de 80% desta massa de dados está concentrada na forma de dados não estruturados, mais especificamente em formato de artigos, notícias, documentos, publicações. A partir de toda essa quantidade de dados, faz-se necessário métodos para organização e obtenção de conhecimento.

A área de Mineração de Textos (MT) vem se destacando por desenvolver e aplicar técnicas que utilizam processos para descobrir conhecimento em coleções de dados textuais de forma automatizada (NEGI, 2021; AGGARWAL; ZHAI, 2012; REZENDE *et al.*, 2003). Nesse sentido, a MT concentra-se na busca por padrões, tendências e regularidades em documentos escritos em linguagem natural. De maneira geral, a MT visa a extração de conhecimentos relevantes em um determinado domínio.

O processo de MT pode ser dividido em uma sequência de etapas, descritas com mais

detalhes na Seção 2.1, que permitem a extração de conhecimentos acerca dos dados analisados. As possibilidades de aplicações dentro do processo de MT são bastante variadas. Elas vão desde a gestão e organização de conhecimentos, passando por exemplo pela classificação automática de documentos, artigos ou notícias, até a extração de conhecimentos em um domínio específico, como bases biomédicas, notícias, textos de redes sociais e artigos. Deste modo, o processo de MT deve ser instanciado de acordo com a necessidade de cada aplicação. Por exemplo, na classificação automática de documentos são utilizadas técnicas com o objetivo de obter classificadores eficientes. Já aplicações que têm como objetivo uma eficiente organização da informação textual geralmente utilizam métodos de agrupamento de dados.

1.2 Motivação e lacunas

Um desafio que tem recebido cada vez mais importância na comunidade que trabalha com MT é o estudo e o tratamento da semântica ou do contexto semântico existente nos textos. O assunto semântica dos textos possui diversos níveis de entendimento. De forma geral, os textos são tratados como um conjunto de palavras não ordenadas. Neste formato, eles podem ser facilmente representados em uma matriz do tipo atributo-valor, denominada *Bag of Words* (BoW), aceita como entrada por grande parte dos algoritmos de aprendizado e extração de padrões. No entanto, em muitas aplicações o uso das informações semânticas contidas nos textos pode levar a análises mais significativas (AGGARWAL; ZHAI, 2012). Nesta tese de doutorado o termo semântica é usado em um sentido mais geral, considerando o significado de itens linguísticos, sejam eles palavras, expressões ou documentos completos.

A incorporação de informações semanticamente mais ricas às representações textuais permite atingir resultados mais eficazes em tarefas de MT, como classificação de documentos, agrupamento, análise de sentimentos e sistemas de recomendação (NIKISHINA *et al.*, 2022; YAN *et al.*, 2020; LI *et al.*, 2018; SINOARA; ANTUNES; REZENDE, 2017). O uso de informações enriquecidas possibilita uma melhora na representação de relacionamentos entre palavras, conceitos de afirmação e negação, bom ou ruim, verdadeiro ou falso, entre outros. No uso de representações tradicionais que consideram apenas a ocorrência das palavras de forma individual, como a BoW, esse nível de informação é perdida.

Nas tarefas de MT, ao se analisar a semântica dos textos é possível, por exemplo, diferenciar não apenas assuntos de determinada coleção de notícias, mas também diferenciar o posicionamento (a favor ou contra) daquele texto em relação ao assunto. Em um exemplo mais prático, na área de análise de sentimentos o principal objetivo é a identificação e classificação de documentos relacionados à opinião de usuários sobre uma determinada entidade (produto, filme, serviço, etc). Nesse tipo de aplicação o contexto o qual esses produtos estão inseridos faz grande diferença em relação ao resultado da classificação. Neste caso, pode-se classificar os textos de usuários que avaliam um produto de forma positiva ou negativa, ou seja, se falaram bem ou

mal daquele item. Este pode ser considerado um uso mais efetivo da semântica nas técnicas de mineração de textos. Portanto, dependendo do objetivo que se deseja atingir com os resultados do processo de Mineração de Textos, a análise do contexto semântico dos textos passa a ser de fundamental importância.

Em Sinoara (2018) são apresentados dois níveis de complexidade semântica no que diz respeito às tarefas de organização e classificação de textos. O primeiro nível, denominado organização por tópicos, consiste em problemas de classificação que dependem basicamente do vocabulário (léxico). Em tarefas de classificação inseridas neste primeiro nível, uma representação de textos tradicional BoW já é suficiente para gerar um classificador com bons resultados. O segundo nível, denominado organização semântica, a organização dos documentos necessita de mais informações para ser resolvida do que apenas o léxico. De maneira geral, a organização dos documentos neste nível de complexidade, apenas o léxico não é suficiente, são necessárias informações adicionais, para a construção de classificadores que atinjam resultados satisfatórios. Como exemplo de classificação no segundo nível de complexidade semântica, no domínio de opiniões sobre produtos é interessante classificar os textos por polaridade (positivo, negativo, ou neutro). Neste caso, são necessárias informações adicionais ao vocabulário utilizado que identifiquem este nível de informação.

Diversas técnicas de representação de textos que incorporam características semânticas vem sendo propostas na literatura com o objetivo de agregar informações mais ricas e permitir melhores resultados em tarefas de classificação de documentos (SINOARA, 2018). A técnica conhecida como **Expressões do Domínio** agrega um tipo informação que incorpora características semânticas relacionadas ao domínio do problema. Cada expressão do domínio é composta pela associação de um termo relacionado ao domínio com um identificador de classe. Na prática, as expressões do domínio são conjuntos de informações computacionalmente mais caras de serem obtidas pois necessitam do esforço de especialistas do domínio para serem produzidas. As representações de textos tradicionais como a BoW não representam de forma adequada as informações semânticas contidas nos textos, trazendo resultados insatisfatórios em tarefas de classificação de documentos nos problemas de diferentes níveis de complexidade semântica. Informações enriquecidas semanticamente, como as expressões do domínio, quando associadas às representações de textos, podem contribuir para a melhoria dos resultados.

Além disso, nos últimos anos diversos governos e instituições em todo o mundo, vêem dando cada vez mais atenção na segurança dos dados e nas decisões tomadas por sistemas baseados em modelos de aprendizado de máquina (DUTKIEWICZ, 2021; WACHTER, 2019). No Brasil, com a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), e também na Europa, com a *General Data Protection Regulation* (GDPR) (COUNCIL, 2016). As regulamentações trazem maior atenção pra o uso e proteção dos dados, além da transparência e a explicabilidade de sistemas que fazem uso de modelos inteligentes e as soluções precisam atender a essas necessidades (HAMON *et al.*, 2022; MEURISCH; MüHLHäUSER, 2021). As expressões do

domínio, além de contribuirem com a semântica das representações, também possibilitam maior nível de explicabilidade e transparência aos modelos de aprendizado de máquina.

Este trabalho é motivado pela capacidade de melhorar resultados de classificação de nível semântico utilizando informações semanticamente enriquecidades associadas à representação tradicional modelo espaço-vetorial BoW. Assim, nesta tese, é proposto o desenvolvimento de representações semanticamente enriquecidados por expressões do domínio, bem como um método de classificação que utiliza informações enriquecidades para solução de problemas nos diferentes níveis de complexidade semântica e melhorando a explicabilidade dos resultados. Além disso, é proposta a melhoria da construção das representações semanticamente enriquecidas por meio de técnicas semiautomatizadas de extração e construção de listas de termos do domínio e identificadores de classe, que irão compor as expressões do domínio.

1.3 Problema, questões de pesquisa e objetivos

O problema a ser tratado nesta tese refere-se principalmente à melhoria de resultados de classificiação de nível semântico utilizando-se para isso informações enriquecidades por meio de expressões do domínio.

Este trabalho de doutorado tem sido desenvolvido com o propósito geral de avançar as pesquisas da área de Mineração de Textos, em especial na classificação automática de textos, em relação à incorporação da semântica na representação de coleções de documentos para resolver problemas de diferentes níveis de complexidade semântica. Assim, o desenvolvimento deste trabalho está guiado por questões de pesquisa referentes ao estado atual das pesquisas de Mineração de Textos com foco na semântica e também referentes ao impacto da semântica tanto na representação de documentos quanto na tarefa de classificação automática.

As questões de pesquisa que direcionam este trabalho são apresentadas a seguir:

Q1 A utilização e a combinação de representações semanticamente enriquecidas com a *Bag of Words* pode levar a melhores resultados de classificação?

Q2 Qual o impacto da utilização de expressões do domínio geradas a partir de listas construídas de forma semiautomática, nas tarefas de classificação para resolverem problemas de diferentes níveis de complexidade semântica?

Neste contexto, o **objetivo geral** deste trabalho é desenvolver diferentes abordagens de representações semanticamente enriquecidas e um método de classificação que utilize as representações enriquecidades e que proporcione melhores resultados em cenários de classificação de nível semântico.

Esse objetivo geral está organizado nos seguintes objetivos específicos:

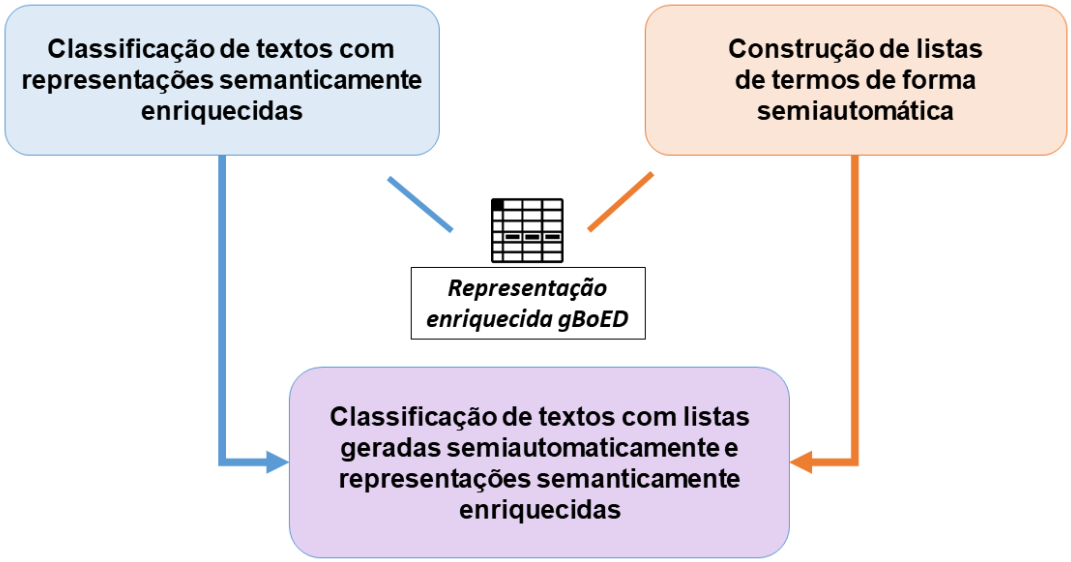
O1 Propor e desenvolver diferentes soluções de representações semanticamente enriquecidas por expressões do domínio que possam ser aplicadas em diferentes domínios e idiomas.

O2 Propor, desenvolver e analisar o impacto de um método de classificação que combine a representação BoW às representações semanticamente enriquecidas por expressões do domínio em problemas de diferentes níveis de complexidade semântica.

O3 Aplicar e analisar o impacto de soluções de extração de termos para a construção de listas termos do domínio e identificadores de classe, usando técnicas baseadas em análise sintática e BERT, a fim de tornar o processo de construção das diferentes versões de representações enriquecidas por expressões do domínio mais automatizado.

A Figura 1 apresenta um diagrama que resume como as soluções geradas para cada objetivo desta tese interagem entre si.

Figura 1 – Diagrama de interação entre as soluções desta tese.



Fonte: Elaborada pelo autor.

Na Seção 1.4 são apresentados os principais resultados obtidos no desenvolvimento desta tese de doutorado.

1.4 Principais resultados

- **Processo de generalização da representação semanticamente enriquecida Bag of Expressions of Domain (BoED)**. Foi realizado um processo de generalização da construção da representação *Bag of Expressions of Domain* (BoED). A representação passou a ser chamada *generalized Bag of Expressions of Domain* (gBoED) e futuramente, após a elaborações da representação com métrica de distância, esta passou a ser chamada de gBoED_Freq pois possui métrica por frequência. Esse resultado está relacionado à questão de pesquisa **Q1** e ao objetivo específico **O1**.

- **Método de Classificação Semanticamente Enriquecida por Expressões do Domínio**. Desenvolvimento de um método de classificação de nível semântico que associa modelos gerados por meio da representação tradicional BoW com a representação baseada em Expressões do Domínio. Foram realizados experimentos aplicados em diferentes coleções de documentos, variando um conjunto de algoritmos, configurações e diferentes cenários, para análise dos impactos do uso do método e suas limitações. Com o desenvolvimento, análise e avaliação experimental do método de classificação, foi tratada a questão de pesquisa **Q1**, atendendo ao objtivo específico **O2**.

- **Método de extração de termos e representação enriquecida baseados em regras morfossintáticas**. Desenvolvimento, aplicação e análise de método de extração de termos baseado em regras sintáticas e análise de impacto utilizando o Método de classificação semanticamente enriquecida por Expressões do Domínio. Esse resultado contempla a questão de pesquisa **Q2** e ao objetivo específico **O3**.

- **Método de extração de termos baseado em modelo de linguagem BERT**. Adaptação, aplicação e análise de impacto de um método de semiautomático de extração de termos baseado em modelo de Linguagem BERT, de modo a melhorar o processo de extração de termos e construção das listas de termos do domínio e identificadores de classe. Esse resultado também contempla a questão de pesquisa **Q2** e ao objetivo específico **O3**.

- **Representações Semanticamente Enriquecidas por Expressões do Domínio**. Ao longo do desenvolvimento das soluções geradas ao longo desta tese de doutorado, foram criadas de diferentes abordagens da Representação Semanticamente Enriquecidas por Expressões do Domínio. A primeira delas criada a partirdo processo de generalização da representação BoED, foi desenvolvida a representação gBOED_Freq, que utiliza métrica de frequência das expressões em cada documento. A segunda representação baseada-se na métrica de distância entre os termos da expressão em uma sentença. Esta representação foi denominada gBoED_Dist. Por último, a partir das regras morfossintática desenvolvidas para extração de termos, foi proposta a representação gBoED_Syntax construída com base em regras morfossintáticas. Esse resultado atende ao objetivo específico **O1**, que serve de base para a questão de pesquisa **Q1**.

- **Análise de impacto do uso das listas extraídas pelos métodos semiautomáticos no Método de Classificação Semanticamente Enriquecida por Expressões do Domínio**. Aplicação e análise do impacto da utilização das representações semanticamente enriquecidades por expressões do domínio construídas por listas de termos extraídas pelos métodos semiautomáticos, baseados em regras morfossintáticas e que utilizam modelo de linguagem BERT. Esse resultado atende às questões de pesquisa **Q1** e **Q2**, além de atender ao objetivo específico **O2** e **O3**

1.5 Organização do texto

O restante desta tese está organizado como se segue.

Capítulo 2 - Fundamentação Teórica. Nesse capítulo são apresentados conceitos essenciais para o entendimento deste trabalho e o posicionamento do mesmo na área de pesquisa na qual ele está inserido. Assim, primeiramente é apresentada uma visão geral sobre o processo de Mineração de Textos, com especial atenção para a representação de documentos e a tarefa de classificação automática de textos. Outro destaque é o assunto que envolve a semântica textual, expressões do domínio e, a definição do principal problema que será tratado neste que trabalho que é definido como níveis de complexidade semântica que permeiam as tarefas de classificação e o uso de representações semanticamente enriquecidas. Ainda no Capítulo 2 são apresentados conceitos relacionados a modelos de linguagem, extração de termos e aspectos, bem como classificação usando BERT.

Capítulo 3 - Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. Nesse capítulo é apresentado o principal método desenvolvido nesta tese de doutorado: o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. No Capítulo 3 é apresentado o trabalho de generalização da representação enriquecida por expressões do domínio, denominada gBoED, bem como as duas versões da representação que diferenciam-se pela formação das expressões e da métrica associada: por Frequência (gBoED_Freq) e por Distância (gBoED_Dist). Ainda neste capítulo são apresentados os trabalhos relacionados com esta tese e algumas lacunas. Por fim, é realizado uma intensa validação do método e das qualidade das representações formadas e aplicadas a 10 coleções de documentos distintas. O processo de validação visa identificar a qualidade método perante características como idioma (Português e Inglês), domínio ao qual a coleção está inserida e a performance do métodos em diferentes cenários.

Capítulo 4 - Extração de Termos e Classificação Semântica Baseados em Regras Morfossintáticas. Nesse capítulo é apresentado o desenvolvimento de um método de extração de termos baseado em regras morfossintáticas. Os conjuntos de termos extraídos formam as listas de termos do domínio e identificadores de classe. Com as mesmas regras foi gerada uma nova abordagem da representação semanticamente enriquecida por expressões do domíni. Tanto

as listas de termos quanto a nova representação enriquecida foram aplicados ao Método de Classificação Semanticamente Enriquecida por Expressões do Domínio e validados em diversos algortimos e cenários.

Capítulo 5 - Extração de Termos e Classificação Semântica Usando Modelo de Linguagem BERT. Nesse capítulo é apresentada a aplicação de um método baseado em modelo de linguagem BERT para extração automática de termos e adaptado para o cenário de expressões do domínio. Os conjuntos de termos extraídos são aplicados e validados pelo Método de Classificação Semanticamente Enriquecida por Expressões do Domínio.

Capítulo 6 - Estudo de Caso: Uso do Método de Classificação Semântica em Pedidos de Informação. Nesse capítulo é apresentado um estudo de caso realizado em parceria com uma aluna do Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI) de 2020. Nesse estudo de caso foi realizada a aplicação e validação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio em dados reais, bem como das utilizadade e qualidade das representações gBoED_Freq e gBoED_Dist, em dados de pedidos de acesso à informação da Controladoria Geral da União, Governo Federal.

Capítulo 7 - Conclusões. Esse capítulo traz as considerações finais e as contribuições desta tese. Também são apresentadas as publicações originadas deste trabalho, as limitações e os direcionamentos para trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de textos

Devido à grande quantidade de documentos em formato textual sendo produzidos em grande velocidade e armazenados diariamente em meio digital, a análise manual desse conteúdo torna-se inviável. Contudo, existe a necessidade e o interesse em organizar, classificar e, de modo geral, extrair conhecimento desses dados (KUMAR; KAR; ILAVARASAN, 2021). Com esse objetivo, a área de Mineração de Textos (MT) tem como objetivo a aplicação de técnicas e algoritmos para dar suporte à extração de conhecimento de grandes coleções de documentos (HICKMAN *et al.*, 2022; KUMAR; KAR; ILAVARASAN, 2021).

O processo de Mineração de Textos, apresentado na Figura 2, pode ser dividido em cinco etapas distintas: Identificação do Problema, Pré-processamento, Extração de Padrões, Pós-processamento e Utilização do Conhecimento (HASSANI *et al.*, 2020; JO, 2018; GAIKWAD; CHAUGULE; PATIL, 2014; REZENDE *et al.*, 2003).

- **Identificação do Problema:** nesta etapa o especialista em MT deve especificar os objetivos da mineração e juntamente com um especialista do domínio deve delimitar o escopo do problema a ser tratado. É realizada a captação de todo o conjunto de textos, que podem ser encontrados em diferentes locais, como bibliotecas de documentos impressos ou mídias digitais, computador em arquivos armazenados em discos rígidos e, em geral, na Internet. Neste último, podem ser encontrados diversos repositórios espalhados em servidores por todo o mundo. As especificações definidas na etapa de Identificação do Problema servem como base para as próximas etapas do processo de MT, as quais podem ser executadas de forma cíclica.

- **Pré-processamento:** a etapa de Pré-processamento é considerada a fase mais onerosa do processo. É responsável pela preparação e estruturação dos dados para serem submetidos

Figura 2 – Processo de Mineração de Textos.

Fonte: Rezende *et al.* (2003).

aos algoritmos de extração de padrões (GAIKWAD; CHAUGULE; PATIL, 2014). É nessa etapa que os documentos são representados de modo a torná-los processáveis pelos algoritmos usados para extração de padrões.

Para que seja possível extrair padrões com qualidade, técnicas podem ser aplicadas durante a etapa de pré-processamento de modo a preservar os padrões ocultos nos documentos (JO, 2018). É aconselhável forte análise sobre o texto para avaliar quais as melhores técnicas a serem utilizadas.

Após a preparação do texto, um bom modelo de representação deve ser adotado para permitir a extração dos padrões que atendam os objetivos definidos no início do processo. Na Subseção 2.1.1 é apresentada uma discussão mais detalhada sobre as principais técnicas de pré-processamento e preparação de textos, além das formas como os dados textuais podem ser representados na Mineração de Textos e as iniciativas em direção ao enriquecimento da representação utilizando aspectos semânticos.

- **Extração de Padrões:** após a preparação e limpeza dos dados, a etapa de "Extração de Padrões" visa a aplicação de técnicas para extração de conhecimentos. Para isso, são utilizadas combinações de algoritmos e técnicas de Mineração de Dados provenientes de diversas áreas do conhecimento, tais como: aprendizado de máquina, estatística e processamento de língua natural. A escolha do algoritmo é feita com base nos dados disponíveis e no tipo de conhecimento que se deseja descobrir.

  Segundo Gaikwad, Chaugule e Patil (2014), as principais tarefas de Mineração de Dados que podem ser aplicadas nesta etapa são:

  - **Classificação:** visa identificar a qual classe um determinado registro pertence. Utilizam-se algoritmos de aprendizado de máquina supervisionado ou semi-supervisionado. Neste trabalho a tarefa de classificação será utilizada com o objetivo de avaliar os métodos propostos. Na Subseção 2.1.2 serão apresentados maiores detalhes sobre classificação de documentos.
  - **Agrupamento:** também utilizada na organização de documentos, essa tarefa visa identificar e aproximar os documentos mais similares. Um grupo (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais grupos. Esta tarefa difere da classificação pois não necessita que os documentos estejam previamente categorizados. Por exemplo, em uma auditoria, poderiam ser separados os comportamentos suspeitos.
  - **Regressão:** similar à classificação, a regressão é usada quando o registro é identificado por um valor numérico e não por uma categoria. Assim, é possível estimar o valor de uma determinada variável analisando-se os valores das demais variáveis. Por exemplo, estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas.
  - **Associação:** a tarefa de associação consiste em identificar atributos que estão relacionados. Por exemplo, identificar quais produtos são levados em conjunto pelos consumidores de um supermercado, ou seja, se um consumidor levou um produto a, será que ele leva b ou c ?

- **Pós-Processamento e Utilização do Conhecimento:** após a extração de padrões, nessa fase o foco é avaliar se os resultados obtidos atingem os objetivos definidos no início do processo. Isso pode ser feito com o apoio de um especialista do domínio ou por meio da aplicação de métricas objetivas de avaliação. Por fim, após a validação e cumprimento dos objetivos estipulados, o conhecimento é distribuído aos usuários. Na Subseção 2.1.3 são apresentadas as formas de avaliação dos resultados.

Em seguida, serão apresentados de forma mais detalhada conceitos que envolvem a etapa de **pré-processamento e representação de textos**.

2.1.1 Pré-processamento e representação de textos

Como dito anteriormente, a etapa de Pré-processamento é a fase mais onerosa do processo de Mineração. A Figura 3 apresenta um conjunto com as principais técnicas que podem ser aplicadas nesta etapa. Ela é dividida em quatro importantes tarefas: 1. Preparação dos Textos, 2. Extração de Termos, 3. Enriquecimento e 4. Representação. Elas são aplicadas de acordo com a necessidade de cada problema e sem uma ordem definida.

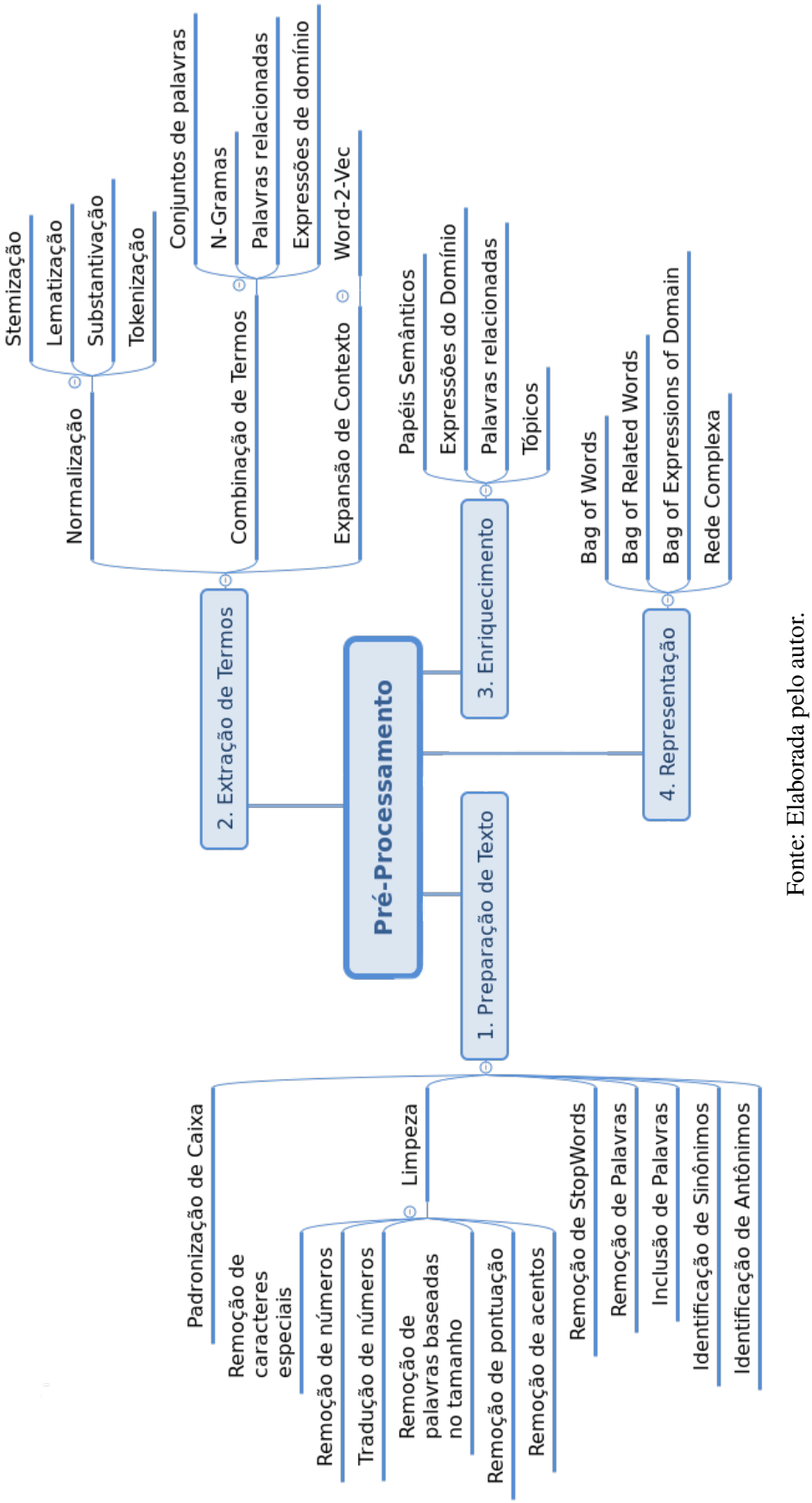


Figura 3 – Tarefas e técnicas de Pré-processamento.

Fonte: Elaborada pelo autor.

Na tarefa de **Preparação dos dados** é possível aplicar as seguintes técnicas:

- **Padronização de caixa:** Padronização do texto em caracteres maiúsculos ou minúsculos. Exemplo: cantar ⇒ CANTAR e vice-versa;
- **Limpeza de Dados:** Várias tarefas correspondem à limpeza de dados:
  - Remoção de caracteres especiais: remoção de caracteres como *, @, $, #, %, &, _, { ;
  - Remoção de números;
  - Tradução de números: Conversão de números ordinais para uma versão por extenso. Exemplo: 45 ⇒ quarenta e cinco;
  - Remoção de palavras baseadas no tamanho: pode-se definir um tamanho mínimo ou máximo de caracteres para as palavras. Exemplo: Para um limite mínimo de 2 caracteres, elimina-se palavras como se, ao, ai;
  - Remoção de pontuação: remoção de pontuações como ponto (.), vírgula (,), travessão (-), dois-pontos (:), parênteses (), ponto de exclamação (!);
  - Remoção de acentos: remoção de acentos em palavras como em país ⇒ pais, mínimo ⇒ mínimo;
- **Remoção de *Stopwords*:** Remoção de palavras ou *tokens* que aparecem repetidamente e acabam por prejudicar a análise textual. Exemplo: conjunções ⇒ e, mas, ou, logo, pois, que, como, porque;
- **Remoção de Palavras:** Remoção de palavras que prejudicam a análise do texto e que não pertencem exatamente ao conjunto de *Stopwords*;
- **Inclusão de Palavras:** Inclusão de palavras que auxiliam a análise do texto;
- **Identificação de Sinônimos ou Antônimos:** Identificação de palavras que possuem o mesmo significado de outras, ou significados opostos. Exemplo: Apresentar ⇒ divulgar, difundir, veicular, publicar;

Na tarefa de **Extração de Termos**, as seguintes técnicas são possíveis:

- **Normalização:** Possui as seguintes sub-tarefas:
  - **Stemização (ou Radicalização):** Reduz cada palavra do texto ao seu radical. Exemplo: cantar, cantarolar, cantou ⇒ cant;
  - **Lematização:** Substitui uma palavra flexionada pela forma eliminando número e gênero. Exemplo: cantaremos ⇒ cantar;

  - **Substantivação:** Substitui uma palavra flexionada pelo seu substantivo Exemplo: cantar, cantaremos ⇒ canto;
  - **Tokenização:** Dividir o texto em unidades básicas conhecidas como *tokens*, utilizando delimitadores como espaços em branco ou pontuação.

- **Combinação de Termos:**
  - **Conjuntos de Palavras:** Palavras que ocorrem juntas de forma não sequencial. Exemplo: na sentença *"O cachorro comeu o chinelo outra vez."* pode-se extrair *cachorro_comeu*, *cachorro_comeu_chinelo*, *cachorro_chinelo_outra*;
  - **N-Gramas:** Agrupamento de itens por palavras, fonemas, sílabas ou letras. Exemplo: na sentença *"O cachorro comeu o chinelo outra vez."* pode-se extrair bigramas (*cachorro_comeu*, *comeu_o*, *chinelo_outra*), trigramas (*O_cachorro_comeu*, *cachorro_comeu_o*, *comeu_o_chinelo*);
  - **Palavras Relacionadas:** Relacionamento de palavras por um determinado contexto. Exemplo: No contexto de supermercado, pode-se relacionar palavras como produtos, compras, pagamento, itens;
  - **Expressões de Domínio:** Conjuntos de palavras ligadas por um domínio específico. Exemplo: No domínio de esportes, pode-se identificar expressões do tipo *Guga_venceu*, *Hamilton_na_pole*, *Barrichello_abandona_prova*.

- **Expansão de Contexto:** Agregação de informações relacionadas ao contexto de uma determinada palavra. Exemplo: Na sentença *"O cachorro comeu o chinelo outra vez."*, cachorro ⇒ mamífero, quadrúpede; chinelo ⇒ calçado;

Na tarefa de **Enriquecimento**, as seguintes técnicas podem ser aplicadas:

- **Papéis Semânticos:** Identificação dos papéis pertencentes a cada parte de uma sentença. Os papéis podem ser de agente (quem realiza uma ação), ação (verbo que exprime a ação), instrumento (por meio do qual uma ação ocorre), paciente (sofre o efeito da ação), beneficiário (receptor do resultado de uma ação), caminho (percurso no qual a ação é realizada), tempo (instante de uma ação), etc. Como exemplo, a Tabela 1 apresenta os papéis semânticos referentes a cada parte da senteça "O usuário limpou a impressora com um detergente.".

- **Expressões do domínio:** Possui o mesmo significado do item pertencente à Combinação de Termos. As expressões do domínio podem ser utilizadas tanto como combinação de termos quanto enriquecimento de dados;

- **Palavras Relacionadas:** Possui o mesmo significado do item pertencente à Combinação de Termos. Também podem ser utilizadas como técnica de enriquecimento de dados;

Tabela 1 – Papéis semânticos aplicados à uma sentença.

Papel	*Palavra*
Agente	usuário
Ação	limpar
Objeto	impressora
Instrumento	detergente

- **Tópicos:** Extração de itens que possuem relação com o contexto dos documentos.

Por último, na tarefa de **Representação** é gerada a representação dos documentos utilizando as possíveis estruturas:

- ***Bag of Words***(BoW): Representação em forma de matriz atributo-valor do tipo documento-termo;
- ***Bag of Related Words*:** Matriz atributo-valor de palavras relacionadas;
- ***Bag of Expressions of Domain*:** Matriz atributo-valor de expressões do domínio;
- ***Redes Complexas*:** Representação no formato de grafo, cujos nós e arestas permitem representar textos em várias dimensões, como, termo, documento, relacionamentos, direções, etc.

Todas as técnicas descritas anteriormente estão relacionadas à forma como as informações contidas nos textos serão representadas, ou seja, como os dados presentes nos textos serão transformados de um formato não estruturado para um formato estruturado. Este projeto de pesquisa irá atuar nesta tarefa do processo de Mineração de Textos, a representação de textos. O objetivo é incorporar aspectos semânticos de modo a enriquecer as representações.

A técnica de representação mais tradicional existente na literatura para textos é conhecida como *Bag of Words* (BoW). Nela, o conjunto de documentos são representados em um modelo espaço-vetorial (TAN; STEINBACH; KUMAR, 2005), formando uma matriz do tipo atributo-valor, em que cada instância (exemplo) corresponde a uma linha e seus atributos (características) correspondem às colunas. Nessa representação, os termos são considerados independentes entre si, formando um conjunto desconectado de palavras.

Na Tabela 2 é apresentado um exemplo de BoW, na qual a métrica que relaciona os termos com os documentos corresponde à presença daquele termo no documento. Na representação em formato binário, caso o termo esteja presente no documento, a posição dos termos no documento recebe valor 1, caso contrário 0. Nessa tabela de exemplo os documentos correspondem às sentenças definidas na Figura 4.

Figura 4 – Sentenças representadas em uma BoW.

$D1$	*Pedro atacou o homem.*
$D2$	*O homem atacou Pedro.*
$D3$	*Pedro atravessou a rua.*

Fonte: Elaborada pelo autor.

Tabela 2 – Exemplo de matriz documento-termo *Bag of Words (BoW)*

	Pedro	**atacou**	**o**	**homem**	**atravessou**	**a**	**rua**
D1	1	1	1	1	0	0	0
D2	1	1	1	1	0	0	0
D3	1	0	0	0	1	1	1

Nota – As sentenças utilizadas neste exemplo não foram submetidas a nenhuma técnica de Pré-processamento definida nesta Seção. Por isso, os artigos (a, o) não foram removidos como *StopWords*.

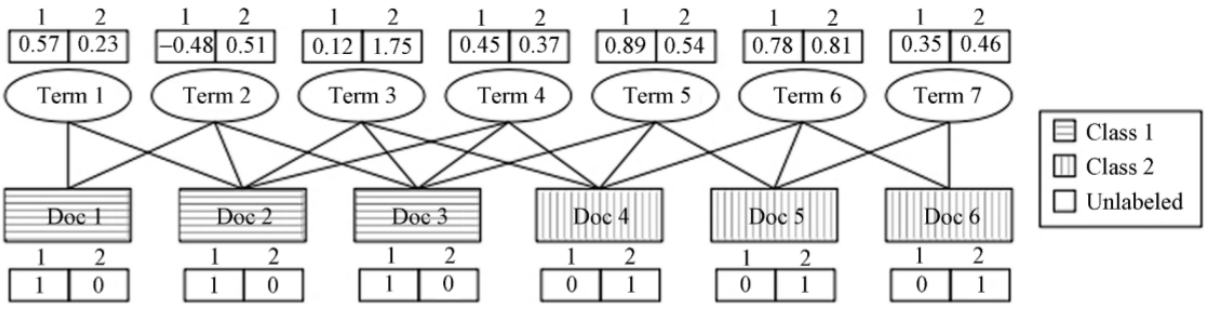
Ainda na Tabela 2, pode-se observar que, por considerar as palavras de forma individual a representação BoW define igualmente as sentenças dos documentos $D1$ e $D2$, não considerando a semântica contida nas sentenças.

Outros tipos de representações são propostas na literatura para representar textos e documentos. Diversos autores propõem soluções de representação de textos em documentos no formato de redes complexas. Em Rossi *et al.* (2014), Rossi, Lopes e Rezende (2016) por exemplo, são propostas duas soluções de representações de textos em formato de redes complexas heterogêneas bipartidas para classificação de documentos supervisionada e semissupervisionada em diferentes domínios. Em Mujtaba *et al.* (2018), Janz, Kdzia e Piasecki (2018), Schlutter e Vogelsang (2018) são propostas diferentes soluções usando redes para classificação de textos em diversos idiomas. É importante ressaltar que apesar de diversos autores terem buscado resolver problemas de representação usando diferentes abordagens, como é o caso das redes complexas, em nenhum dos trabalhos encontrados foram observadas soluções para problemas de classificação com níveis mais altos de complexidade semântica. Da mesma maneira, em todos os trabalhos os termos ainda são tratados de forma individual, o que causa grande perda semântica às representações dos documentos.

Na Figura 5 é apresentado um exemplo da abordagem proposta por Rossi *et al.* (2014) usando redes bipartidas. Nessa abordagem, o autor utiliza uma camada de nós para representar os documentos e outra camada de nós para representar os termos mais relevantes para esses documentos. As arestas conectam os documentos com os termos que a eles pertencem. Pesos são associados aos termos de modo a refletir o grau de relevância que um termo possui em relação

a um documento. Valores negativos para pesos indicam que um determinado termo inibe ou diminui o nível de associação de um documento com uma classe. As classes são associadas aos documentos por meio de valores binários.

Figura 5 – Exemplo de rede complexa heterogênea bipartida.



Fonte: Rossi *et al.* (2014).

Segundo Lu e Getoor (2003), principalmente em problemas que envolvem semântica de textos, aplicar algoritmos assumindo que as instâncias e os atributos são independentes pode levar a resultados incorretos. Além disso, as representações que utilizam a técnica *Bag of Words* em grandes coleções de documentos, geralmente apresentam alta dimensionalidade e alta esparsidade, causando a baixa eficiência de muitos algoritmos de extração de padrões ao lidar com representações deste tipo (BREVE, 2010).

Um dos focos deste trabalho está na construção de um método de classificação de problemas de diferentes níveis de complexidade semântica que utiliza representações de textos enriquecidas semanticamente. As expressões do domínio serão utilizadas dentro desta abordagem como forma de enriquecimento das representações textuais. A descrição do desenvolvimento desta proposta e os resultados serão descritos com detalhes no Capítulo 3.

Portanto, em cenários que consideram a semântica dos textos e as relações que os termos possuem entre si, a perda dessas informações pode prejudicar os resultados dos métodos de extração de padrões. Para isso, torna-se necessário o desenvolvimento e a utilização de técnicas complementares ou auxiliares que enriqueçam as representações com informações de nível semântico. Na Seção 2.2 a semântica existente nos textos será explicada de forma mais detalhada. Na Subseção 2.1.2 serão apresentados os conceitos relacionados à extração de padrões por meio de classificadores e como serão utilizados no desenvolvimento deste trabalho.

2.1.2 Extração de padrões

A etapa de extração de padrões tem por objetivo identificar as correlações entre elementos, de modo a extrair conhecimentos compreensíveis, válidos, novos e potencialmente úteis dentro de uma grande massa de dados (AGGARWAL; ZHAI, 2012). Em MT a entrada de dados para a etapa de extração de padrões é a representação da coleção de documentos.

Como apresentado na seção anterior, diversas abordagens de representação de textos tem sido explorados com o objetivo de agregar o maior conjunto de informações possível enriquecendo seu conteúdo. A representação mais tradicional *Bag of Words* baseia-se no modelo espaço-vetorial para a construção de uma matriz do tipo documento-termo que representa os documentos com uma coleção de palavras independentes. Outras abordagens buscam a construção de representação que garantem algum nível de relacionamentos, como é o caso das representações por grafos (SUN; HAN, 2012; HAN, 2012).

Um problema de extração de padrões consiste de uma tarefa de classificação ou agrupamento. A principal área de estudo sobre extração de padrões é a área de Aprendizado de Máquina (AM). Nesta área de conhecimento são utilizados algoritmos que aprendem o conhecimento e identificam padrões entre os dados. As técnicas de AM são divididas em três grandes frentes:

Aprendizado supervisionado: Os exemplos são previamente rotulados por um especialista do domínio e o algoritmo aprende com os exemplos dados.

Aprendizado não-supervisionado: Os exemplos dados não são rotulados e os algoritmos buscam características no conjunto de dados de como a separar os conjuntos por medidas de similaridade.

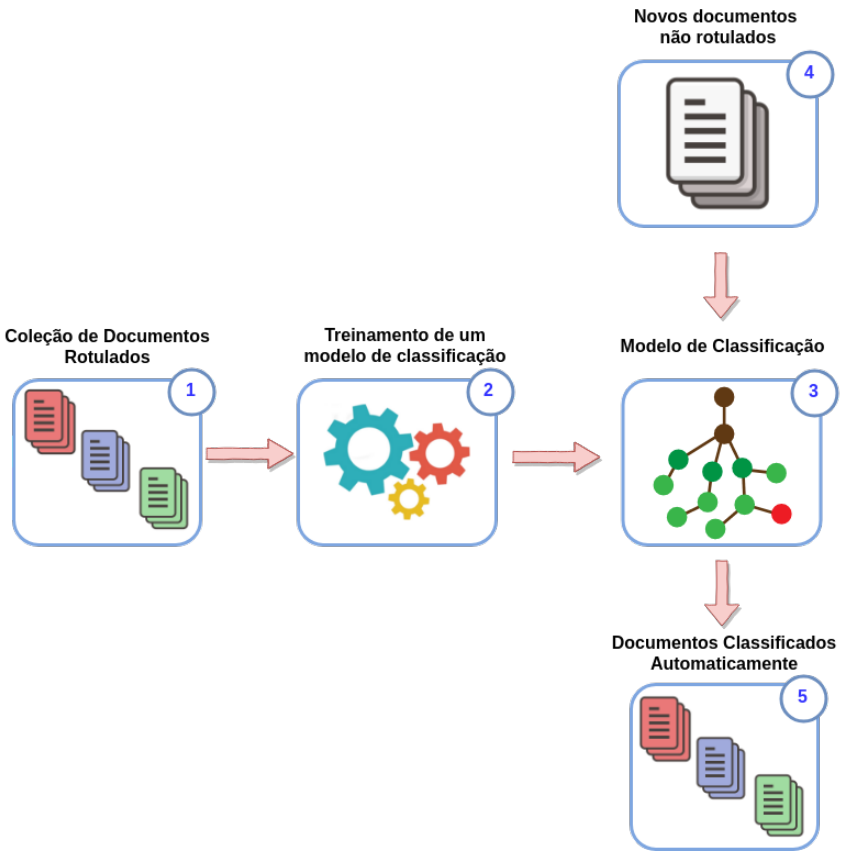
Aprendizado semissupervisionado: Derivada das técnicas de aprendizado supervisionado e não-supervisionado, alguns dados rotulados são oferecidos aos algoritmos juntamente com dados não rotulados.

O interesse na área de extração de padrões tem crescido muito devido as aplicações que, além de serem desafiantes, são também computacionalmente cada vez mais exigentes. Áreas como bioinformática e análise de sequências genéticas, classificação e organização de documentos, sensoriamento remoto e prognóstico da produção de colheitas, além de reconhecimento de voz são alguns exemplos de domínios de aplicações de aprendizado de máquina e extração de padrões.

Considerando-se o uso de algoritmos de aprendizado supervisionads, o problema de classificação automática de textos é definido como se segue. Dado um conjunto de classes (C) e uma coleção de documentos rotulados (D), documentos cuja classe é conhecida, um algoritmo supervisionado induz uma função F que mapeia os documentos de D às classes de C ($F : D \mapsto C$). A função F é chamada de modelo de classificação (ou classificador) e é utilizada para predizer a classe de novos documentos. A Figura 6 ilustra as etapas do processo de construção de um classificador.

Na Figura 6, a classificação automática de textos é apresentada em 5 etapas descritas a seguir:

Figura 6 – Esquema ilustrativo da classificação automática de textos por meio de aprendizado supervisionado



Fonte: Adaptada de Rossi (2016).

1. **Coleção de documentos rotulados:** Nesta etapa a preparação dos dados ocorre por meio da entrada de dados e do pré-processamento dos textos. Nesta etapa uma coleção de documentos é preparada utilizando as técnicas apresentadas na Subseção 2.1.1. Na classificação supervisionada, os documentos necessitam estar rotulados, e.g., anotados com as respectivas classes. Nesta etapa a coleção de documentos é representada no modelo espaço-vetorial. O modelo espaço-vetorial foi adotado nas soluções apresentadas nesta tese de doutorado, tanto no treinamento de classificadores tradicionais quanto na proposta do método de classifica e da representação enriquecida por expressões do domínio

2. **Treinamento de um modelo de classificação:** Nesta fase ocorre a construção de um modelo de classificação utilizando os documentos rotulados e representados em um modelo espaço-vetorial. No treinamento do classificador, os seguintes algoritmos estão entre os mais utilizados e recomendados para a classificação de textos (ROSSI, 2016; AGGARWAL; ZHAI, 2012; SEBASTIANI, 2002):

   i. **C4.5:** Método baseado em busca via árvore de decisão;

   ii. ***k-Nearest Neighbors* (KNN):** Método baseado em distâncias entre os vizinhos mais próximos;

iii. ***Naïve Bayes* (NB) e variantes**: Método do tipo probabilístico;

iv. ***Support Vector Machine* (SVM):** Método baseado em otimização, por meio de vetores de suporte;

v. ***Perceptron* e *Multi-layer Perceptron*:** Método baseado em Redes Neurais Artificiais.

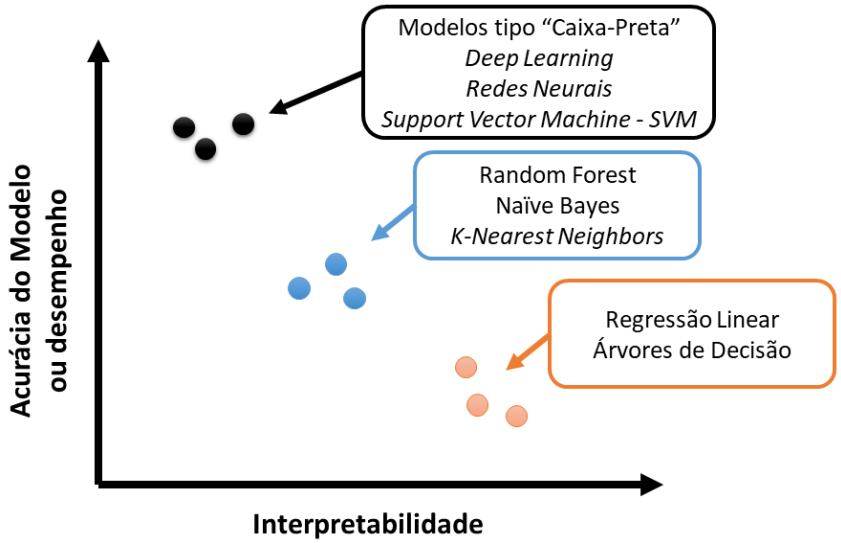
Ainda considerando as representações no modelo espaço-vetorial, além dos algoritmos tradicionais que utilizam como entrada uma matriz, também é possível aplicar algoritmos de aprendizado baseados em redes bipartidas, visto que essas redes podem ser diretamente derivadas de representações no modelo espaço-vetorial. Esse é o caso dos algoritmos *Inductive Model based on Bipartite Heterogeneous Networks* (IMBHN): $IMBHN^C$ (ROSSI *et al.*, 2014) e $IMBHN^R$ (ROSSI, 2016). Por apresentarem uma boa performance de classificação, esses algoritmos serão considerados junto com os tradicionais algoritmos de aprendizado supervisionado no desenvolvimento das soluções para as propostas apresentadas neste trabalho.

3. **Modelo de classificação:** Essa etapa representa o modelo de classificação definido e realiza as tarefas de classificação para novos exemplos de entrada, representados pelo passo 4. No passo 5, os documentos são classificados pelo modelo gerado, utilizando o aprendizado do conjunto de dados de treinamento.

A classificação automática de textos pode ser aplicada em diversos problemas, como organização e recuperação de notícias, artigos e outros tipos de documentos, filtragem de *e-mails* e detecção de *spam*, desambiguação lexical de sentidos, análise de sentimentos e mineração de opinião (MATSUNO *et al.*, 2016; ROSSI, 2016; AGGARWAL; ZHAI, 2012; SEBASTIANI, 2002). Dada a grande aplicabilidade da tarefa de classificação automática de textos, diversas pesquisas têm sido desenvolvidas. Tais pesquisas buscam melhorar cada vez mais a qualidade dos classificadores e aproveitar as diferentes particularidades de cada problema. Uma das diversas abordagens alternativas para se realizar classificação é a combinação de classificadores, técnica conhecida como *ensemble* de classificadores.

Na classificação automática de textos, a avaliação experimental de classificadores é normalmente realizada medindo-se o desempenho de classificação (SEBASTIANI, 2002). Segundo Gunning (2017), Linardatos, Papastefanopoulos e Kotsiantis (2020), a análise comparativa do desempenho dos diferentes tipos de classificadores é importante visto que interpretabilidade ou explicabilidade versus desempenho dos modelos é um *trade-off* comum no aprendizado. Em geral, modelos mais complexos (como SVM e Redes neurais) tendem a ter desempenho melhor do que os modelos mais transparentes (como as árvores de decisão – C4.5), no entanto o maior desempenho vem acompanhado de menor explicabilidade (DOSHI-VELEZ; KIM, 2017; SHAH; JESHWANI; BHATT, 2020). Na Figura 7 é possível visualizar o gráfico que mostra a relação entre desempenho e interpretabilidade dos algortimos tradicionais de aprendizado de máquina.

Figura 7 – Relação entre desempenho e interpretabilidade.



Fonte: Adaptada de Shah, Jeshwani e Bhatt (2020).

Diversas medidas de desempenho podem ser calculadas com base nos valores de uma matriz de confusão, que apresenta o número de exemplos classificados corretamente e incorretamente por um determinado classificador. Na Figura 8 é ilustrada uma matriz considerando uma classe $c_i \in C$.

Figura 8 – Matriz de confusão para a classe c_i.

		Classe predita	
		c_i	$c_j\ (j \neq i)$
Classe real	c_i	TP_{c_i}	FN_{c_i}
	$c_j\ (j \neq i)$	FP_{c_i}	TN_{c_i}

Fonte: Elaborada pelo autor.

A Acurácia de um classificador é dada pela porcentagem de documentos corretamente classificados, conforme a Equação 2.1.

$$\text{Acurácia} = \frac{\sum_{i=1}^{|C|} TP_{c_i}}{N} \tag{2.1}$$

Outras duas medidas que são comumente utilizadas na avaliação da classificação automática de textos são as medidas Precisão (Equação 2.2) e Revocação (Equação 2.3). A Precisão estima a probabilidade condicional de um documento ser da classe c_i, dado que o classificador o rotulou com c_i. Já a Revocação estima a probabilidade condicional do classificador rotular um documento com a classe c_i, dado que o documento realmente é da classe c_i.

$$\text{Precisão}_{c_i} = \frac{TP_{c_i}}{(TP_{c_i} + FP_{c_i})} \tag{2.2}$$

$$\text{Revocação}_{c_i} = \frac{TP_{c_i}}{(TP_{c_i} + FN_{c_i})} \tag{2.3}$$

Para problemas multiclasse, nos quais existem duas ou mais classes e cada documento é rotulado com apenas uma classe, as medidas Precisão e Revocação de cada classe individualmente podem ser sumarizadas considerando-se duas abordagens (SEBASTIANI, 2002): *micro-averaging* (Equações 2.4 e 2.5) e *macro-averaging* (Equações 2.6 e 2.7). As duas abordagens dão ênfases diferentes para a distribuição dos exemplos nas diferentes classes. Como pode ser observado nas Equações 2.4 e 2.5, por realizar a soma dos termos individualmente para cada classe, a abordagem *micro-averaging* corresponde à medida Acurácia (Equação 2.1). Portanto, as medidas Precisão$^{\mu}$, Revocação$^{\mu}$ e Acurácia resultam em um mesmo valor. Já na abordagem *macro-averaging* (Equações 2.6 e 2.7), são dados pesos iguais às classes, independentemente do número de exemplos. Com isso, essa abordagem tende a enfatizar mais a classificação correta de exemplos de classes minoritárias, que possuem uma quantidade muito pequena de exemplos em relação a outras classes, do que a abordagem *micro-averaging*.

$$\text{Precisão}^{\mu} = \frac{\sum_{i=1}^{|C|} TP_{c_i}}{\sum_{i=1}^{|C|} (TP_{c_i} + FP_{c_i})} \tag{2.4}$$

$$\text{Revocação}^{\mu} = \frac{\sum_{i=1}^{|C|} TP_{c_i}}{\sum_{i=1}^{|C|} (TP_{c_i} + FN_{c_i})} \tag{2.5}$$

$$\text{Precisão}^{M} = \frac{\sum_{i=1}^{|C|} \text{Precisão}_{c_i}}{|C|} \tag{2.6}$$

$$\text{Revocação}^{M} = \frac{\sum_{i=1}^{|C|} \text{Revocação}_{c_i}}{|C|} \tag{2.7}$$

As medidas Precisão e Revocação são complementares e podem ser combinadas em outras medidas, como a medida $F1$ (Equação 2.8), que corresponde a uma média harmônica simples das medidas Precisão e Revocação. Ao se utilizar as medidas Precisão e Revocação considerando as abordagens *micro-averaging* ou *macro-averaging*, tem-se as medidas F1$^{\mu}$ ou F1^{M}. Vale notar que, como Precisão$^{\mu}$ e Revocação$^{\mu}$ resultam em um mesmo valor, a medida F1$^{\mu}$ também será equivalente ao valor de Acurácia.

$$F1 = 2 * \frac{\text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.8}$$

Nessa seção foi apresentada uma revisão sobre a classificação automática de textos por meio de algoritmos de AM. Neste trabalho, serão utilizados classificadores para construção de um método para classificação de problemas de diferente complexidade semântica, utilizando

enriquecimento por meio de expressões do domínio. A construção de representações de documentos também será avaliada por meio de classificação de textos e as métricas apresentadas. Informações mais detalhadas sobre esse extenso assunto podem ser encontradas em diversos livros e artigos de revisão da literatura, como os trabalhos de Faceli Ana Carolina Lorena e Carvalho (2021), Aggarwal (2014), Sebastiani (2002) e Mitchell (1997).

2.1.3 Pós-processamento

Na etapa de pós-processamento, as soluções desenvolvidas no processo de Mineração de Textos são avaliadas. Este trabalho de doutorado se divide em três abordagens principais que visam a melhoria do desempenho de classificadores na solução de problemas de diferentes níveis de complexidade semântica. Tais níveis de complexidade semântica são detalhados na Subseção 2.2.1. Cada uma das abordagens será avaliada individualmente de acordo com seu desempenho e comparada com métodos de classificação que utilizam abordagens tradicionais, como a representação *Bag of Words* (BoW). Como base para a avaliação de desempenho serão utilizadas as métricas padrão de classificação automática, apresentadas na Subseção 2.1.2.

Na seção seguinte será apresentado com maiores detalhes do assunto que envolve a semântica de textos. A semântica de textos é o principal tema que envolve esta pesquisa de doutorado.

2.2 Aspectos semânticos dos textos

A semântica é um ramo da linguística que compreende o estudo do significado das palavras (OLIVEIRA, 2012). No estudo da semântica existem diversas vertentes que vão desde a relação entre as palavras e os seres ou coisas do mundo real, até as mudanças de sentido que as palavras sofrem ao longo do tempo. Segundo Riemer (2010) a semântica é um tema bastante abrangente e não existe um consenso sobre os seus limites de estudo. Portanto, essa seção não visa cobrir o tema de forma detalhada. O objetivo é apresentar alguns conceitos da básicos da semântica que possam contextualizar esta monografia de qualificação. O principal interesse deste trabalho está na extração e utilização de informações semânticas que possam enriquecer representações de texto e, com isso, melhorar o desempenho em tarefas de classificação de documentos escritos em língua natural.

De modo geral, as palavras apresentam diversas relações entre si. Essas relações têm impacto direto no significado dos textos e, consequentemente, em algumas tarefas computacionais de Mineração de Textos e de Processamento de Língua Natural, que lidam com a organização de documentos. Conforme apresentado por Oliveira (2018) e Pietroforte (2010), entre essas relações estão:

Sinonímia - Termos que podem se substituir em determinado contexto. Por exemplo: as

palavras “novo” e “jovem” são sinônimas quando se trata da característica de um ser vivo, como em “homem novo”/“homem jovem”, porém “jovem” não pode substituir “novo” quando se trata da característica de um objeto, como em “livro novo” (OLIVEIRA, 2018; PIETROFORTE, 2010).

Antonímia - Termos que possuem significados opostos. Assim como os sinônimos, os antônimos também dependem do contexto. Palavras diferentes podem ter o mesmo antônimo desde que possuam ao menos um sentido em comum. Por exemplo: “velho” pode ser antônimo tanto de “fresco” quanto de “novo”, como em “queijo velho”, “queijo novo” e “queijo fresco” (OLIVEIRA, 2018; PIETROFORTE, 2010).

Hiperonímia - Um termo é hiperônimo de outro quando existe uma relação de englobamento entre eles em uma hierarquia de classificação. Hiperônimo é o termo englobante. Por exemplo: “esporte” é hiperônimo de “futebol” (OLIVEIRA, 2018).

Hiponímia - A hiponímia caracteriza-se pela mesma relação de hiperonímia, porém o hipônimo corresponde ao termo que é englobado. Exemplo: “futebol” é hipônimo de “esporte”.

Homonímia - Termos homônimos são termos com origens distintas e significados distintos, mas que apresentam a mesma forma gráfica (termos homógrafos), fonética (termos homófonos) ou ambas (homônimos perfeitos). Exemplo: o termo “cobra” pode ser tanto o substantivo, nome de um animal, quanto o verbo cobrar conjugado no presente para a terceira pessoa do singular.

Holonímia - Relação parte - todo entre termos. Holônimo é o termo que corresponde ao todo na relação parte-todo. Exemplo: “carro” (todo) é holônimo de “freio” (parte) (OLIVEIRA, 2018).

Meronímia - Corresponde à mesma relação que em holonímia, porém o termo merônimo é aquele que corresponde à parte. Por exemplo: “freio” é parte de “carro”, portanto “freio” é merônimo de “carro” (OLIVEIRA, 2018).

Polissemia - Relação de um termo que possui mais de um significado. Exemplo: o termo “banco” pode tanto se referir ao banco de se sentar, quanto ao banco onde se guarda dinheiro.

Ao consider a semântica, além do significado das palavras, o significado de uma frase (ou sentença) também depende da sua estrutura gramatical. Nas frases “João matou o bandido” e “O bandido matou João” pode-se verificar o mesmo conjunto de palavras, porém, apresentam significados distintos devido às estruturas sintáticas diferentes (alternância entre sujeito e objeto) (MüLLER, 2010). Algumas das relações semânticas presentes em sentenças são:

Paráfrase - Corresponde à relação de sinonímia estendia para sentenças. Por exemplo, na sentença "O sinal estava vermelho" no contexto do trânsito de uma rua poderia ser parafraseado para "O semáforo estava fechado". As duas sentenças possuem o mesmo sentido, porém são escritas utilizando um conjunto de palavras diferente.

Acarretamento - Corresponde à relação de hiponímia estendida para sentenças. As sentenças "Dexter é um assassino" e "Dexter é americano" acarretam a situação presente na sentença "Dexter é um assassino americano".

Contradição - Duas sentenças são contraditórias quando elas não podem ser simultaneamente verdadeiras. Nesta situação não se pode ter as duas sentenças verdadeiras. Por exemplo, "Nesta caixa todas as bolas são vermelhas" e "Alguma Bola não é Vermelha". Neste caso uma afirmação é falsa em relação à outra.

Ambiguidade - Uma sentença é ambígua quando ela pode ter mais de um sentido. A ambiguidade de uma sentença pode ser causada por uma palavra ambígua, por diferentes estruturas sintáticas possíveis, ou por uma ambiguidade semântica (causada por relações anafóricas, dêiticas ou de escopo, relações descritas na sequência). Ambiguidade é também uma característica que nomeia a falta de clareza em uma expressão. No exemplo "Pedro disse ao amigo que havia chegado". Nesta sentença não é possível afirmar se quem havia chegado era Pedro ou o amigo. Em "O guarda deteve o suspeito em sua casa", não é possível determinar de quem era a casa, se era do guarda ou do suspeito.

Relação Anafórica - Ocorre quando um pronome presente na sentença se refere a um nome citado anteriormente na mesma sentença. Na sentença "Ana comprou um cão. O animal já conhece todos os cantos da casa", o termo "o animal" faz referência ao termo antecedente "o cão".

Relação Dêitica - Ocorre quando um pronome presente na sentença se refere a um ente que existe no contexto. Por exemplo, na sentença "Esta cadeira está quebrada" o pronome "esta" faz referência à "cadeira que está aqui perto de mim".

Relação de Escopo - Ocorre quando a interpretação de uma expressão da sentença depende da interpretação de outra. Exemplo: "Cada aluno leu dois livros" pode significar que cada aluno leu quaisquer dois livros ou que dois determinados livros foram lidos pelos alunos.

No exemplo do Quadro 1 é possível observar algumas características importantes que impactam de forma direta no processo de descoberta de conhecimento.

No exemplo apresentado, além do significado das palavras estão envolvidas as seguintes características:

Quadro 1 – Exemplos de sentenças com diferentes características semânticas.

$D1$	*A Empresa Alfa adquiriu a Empresa Beta.*
$D2$	*A Empresa Beta adquiriu a Empresa Alfa.*
$D3$	*A Empresa Beta foi adquirida pela Empresa Alfa.*
$D4$	*A Empresa Alfa comprou a Empresa Beta.*

Fonte: Adaptada de Sinoara (2018).

Relação sujeito x objeto. As sentenças $D1$ e $D2$ são diferenciadas pela sintaxe, pela ordem com que as mesmas palavras aparecem. Ambas possuem as mesmas palavras, porém o sujeito de $D1$ é o objeto de $D2$ e vice-versa.

Voz ativa x voz passiva. As sentenças $D1$ e $D3$ reportam o mesmo fato, apesar de terem sujeitos e objeto/agente da passiva opostos.

Sinonímia. As sentenças $D1$ e $D3$ reportam o mesmo fato, apesar de usarem verbos diferentes.

Como é possível observar a partir do exemplo apresentado, as diversas relações semânticas entre palavras e sentenças impactam diretamente na forma como o texto é entendido e no processo de descoberta de conhecimento. Na subseção seguinte são apresentados os diferentes níveis de complexidade semântica que podem ser encontrados e que serão explorados neste trabalhos de doutorado dentro do processo de descoberta de conhecimento.

2.2.1 Níveis de complexidade semântica

O entendimento da língua natural é um processo complexo que envolve a compreensão de diversos fatores como o vocabulário utilizado, a gramática do idioma, as relações semânticas entre os itens linguísticos e o conhecimento de mundo. Tais fatores estão ligados, respectivamente, ao significado das palavras, às regras que definem como as palavras são utilizadas, ligações entre os itens e o contexto no qual os textos estão escritos.

Sabendo que textos são uma fonte rica de conhecimento, a área de MT e PLN, entre outras tarefas, busca extrair e representar os conhecimentos intrínsecos aos textos para que eles possam ser utilizados em outras tarefas como a classificação de textos. Segundo Riemer (2010), o principal desafio ligado à descoberta automática de conhecimento em textos está relacionado diretamente ao não estruturado e a fenômenos que podem alterar o significado do que está sendo dito, como sarcasmo, ironia e ambiguidade.

Buscando analisar o conhecimento intrínseco aos textos de forma mais profunda, o exemplo do Quadro 2 apresenta como tais fatores podem afetar o problema de classificação automática de textos. Neste exemplo, são apresentados dois documentos extraídos de uma coleção de notícias de esportes e que precisa ser classificada.

Quadro 2 – Exemplos de sentenças de esportes.

$D1$	*Guga é o campeão do Tennis Masters Cup. Ele venceu Agassi por três sets a zero no jogo final.*
$D2$	*Hamilton larga na pole position e vence o Grande Prêmio do Canadá. Após colisão, Massa abandona a prova.*

Fonte: Adaptada de Sinoara (2018).

No cenário deste exemplo, a questão que envolve a classificação desta coleção seria relacionada à questões do tipo "Qual é o assunto do documento?" ou "Sobre o que é cada documento?". Ao observar o documento $D1$ verifica-se que ele possui os termos "Guga", "Tennis Masters Cup", "sets", "Agassi" e "jogo". O documento $D2$ possui os termos "Hamilton", "pole position", "Grande Prêmio", "Massa" e "prova". Esses termos são bem característicos de seus esportes, o que permite distingui-los em dois grupos (ou classes) distintas: $D1$ está relacionado com o esporte Tênis e $D2$ com o esporte Fórmula 1. Verifica-se que cada esporte possui termos (ou palavras-chave) específicos e que documentos relacionados a um mesmo esporte podem ter palavras similares. Logo, cada classe (ou grupo esperado) pode ser determinada em grande parte pelo vocabulário utilizado.

No entanto, usuários diferentes possuem necessidades diferentes de classificação de documentos. Considerando um cenário em que pode ser interessante classificar a mesma coleção de notícias pelo desempenho dos atletas brasileiros, verifica-se que a questão que envolve a organização da coleção de documentos seria "Esse documento refere-se a vitória de um atleta brasileiro?". Considerando novamente os documentos $D1$ e $D2$ do Quadro 2, para este caso, as informações importantes são "Guga é o campeão" e "Massa abandona a prova". E para organizar corretamente esses documentos é necessário saber que Guga e Massa são atletas brasileiros. Com isso, pode-se dizer que $D1$ refere-se a uma vitória de brasileiro e $D2$ refere-se a uma derrota.

Nesse contexto, neste trabalho de doutorado os problemas de classificação de documentos são divididos em dois níveis de complexidade semântica. O primeiro nível, que é chamado de organização por tópicos e consiste em problemas de classificação que dependem basicamente do vocabulário. Nesse problema, cada classe possui termos bastante característicos, e, portanto, o léxico (vocabulário) possui grande relevância para representar o conteúdo dos documentos. Pode-se dizer que os documentos podem ser diferenciados em grande parte pelas palavras utilizadas.

O segundo nível de complexidade semântica engloba os demais problemas de classificação de documentos. Esse segundo nível é chamado de organização semântica. Neste nível, a organização dos documentos necessita de mais informações para ser resolvida do que apenas o léxico. Tais problemas requerem uma análise mais profunda, além apenas das palavras, visto que os documentos de classes distintas podem usar o mesmo vocabulário.

Na Subseção 2.1.1 foram apresentadas algumas técnicas responsáveis pela representação das informações contidas nos documentos. Além da classificação de documentos, esta proposta de doutorado irá explorar formas de enriquecimento das representações de textos com informações semânticas com o objetivo de auxiliar o desempenho de classificadores tanto em problemas do primeiro quanto do segundo nível de complexidade semântica. Na Subseção seguinte serão abordadas os conceitos e técnicas que serão exploradas nesta pesquisa de doutorado sobre enriquecimento de representações por meio de expressões do domínio.

2.2.2 Enriquecimento semântico de representações de textos

Os problemas de organização de documentos tratados na área de MT são tradicionalmente problemas do primeiro nível de complexidade semântica, a organização por tópico. Sebastiani (2002) apresenta uma revisão completa sobre classificação automática de textos por meio de Aprendizado de Máquina. Em sua revisão, o autor apresenta a tarefa de classificação automática de textos como sendo uma tarefa de detecção de tópicos, com a rotulação de textos escritos em língua natural por meio da atribuição de uma categoria temática presente em uma lista de categorias pré-definidas.

Sinoara (2018) também realiza uma análise bastante abrangente levantando diversas coleções de textos de *benchmarking* utilizadas em pesquisas de MT. Nelas também é possível verificar que existe uma predominância por tópicos nos problemas de classificação. Nas coleções elencadas, apenas três de 45 são classificadas como organização semântica. Essas três coleções são coleções de análise de sentimentos e visam a classificação do sentimento (polaridade) no nível de documento. A classificação do sentimento é um caso particular de problemas do segundo nível de complexidade semântica. Nesse tipo de classificação, palavras e expressões de sentimento são indicadores importantes do sentimento manifestado no documento. Palavras de sentimento são palavras utilizadas para expressar sentimentos positivos ou negativos, tais como “bom”, “péssimo”, “incrível”. Dada a importância de tais palavras na análise de sentimentos, várias pesquisas têm o foco na criação de listas de palavras de sentimento (*Sentiment Lexicon*) (LIU, 2012), sendo a *SentiWordNet*[1] um exemplo de recurso léxico bastante utilizado em aplicações de análise de sentimentos.

Como apresentado na Subseção 2.1.1, a tradicional representação *Bag of Words* utiliza vetores para representar os documentos cujas dimensões são os termos presentes na coleção de documentos. Essa representação carrega a informação de que a frequência das palavras nos

[1] SentiWordNet: <http://sentiwordnet.isti.cnr.it/>

documentos indica a relevância do documento para uma determinada consulta (*query*). Essa ideia costuma funcionar bem para os problemas do primeiro nível de complexidade semântica discutidos anteriormente. Uma justificativa intuitiva para esse fato é que o tópico de um documento será influenciado pela escolha de palavras do autor ao escrever o documento (TURNEY; PANTEL, 2010). Assim, para se resolver problemas de maior complexidade semântica, que vão além da classificação por tópico, a representação *Bag of Words*, em geral, não é suficiente.

Buscando representações mais ricas de documentos escritos em língua natural, é possível utilizar recursos de algumas tarefas da área de PLN. Algumas dessas tarefas são apresentadas brevemente a seguir.

Reconhecimento de Entidades Nomeadas - Tarefa de extração de informação que busca identificar as ocorrências de palavras ou expressões que correspondem a nomes próprios, nome de organizações, locais (AMARAL; VIEIRA, 2014; GRISHMAN; SUNDHEIM, 1996). Além das entidades identificadas por nome próprio, também é comum o reconhecimento de expressões temporais e numéricas.

Anotação de Papéis Semânticos - Tarefa que busca identificar o predicado de uma oração e atribuir papéis semânticos a seus argumentos (FONSECA *et al.*, 2016; PALMER; GILDEA; XUE, 2010). Com os papéis semânticos obtém-se informações do tipo "quem fez o que para quem", além de "como fez" e "quando fez". Como exemplos de papéis semânticos tem-se Agente (aquele que inicia a ação), Paciente (aquele afetado pela ação), Instrumento (algo ou meio utilizado para efetuar a ação) e Local (lugar de um objeto ou ação).

Desambiguação Lexical de Sentidos - Tarefa que busca determinar qual sentido uma palavra apresenta quando é utilizada em determinado contexto (NÓBREGA; PARDO, 2014; MORO; RAGANATO; NAVIGLI, 2014; AGIRRE, 2007). Essa tarefa é realizada com o apoio de recursos léxicos, como a *WordNet*[2], que agrupa as palavras em conjuntos de sinônimos e apresenta relacionamentos entre esses conjuntos e seus membros.

Tratamento de sinônimos - Tarefa relacionada ao tratamento das relações de sinonímia que as palavras podem apresentar. Diferentes palavras podem se substituir em uma sentença sem que o significado expresso seja alterado. A *WordNet* é um recurso bastante utilizado para a identificação de sinônimos, visto que apresenta uma lista de sinônimos para cada sentido de uma palavra.

Resolução de correferências - Tarefa que busca identificar todas as expressões que se referem a uma mesma entidade no texto. Uma expressão anafórica, que se refere a uma entidade que foi apresentada anteriormente no texto, pode ser pronomial (como ele, ela, meu) ou definida (a aluna, o presidente) (VIEIRA; GONçALVES; SOUZA, 2008).

[2] WordNet: <http://wordnet.princeton.edu/>

2.2.3 Expressões do domínio

As expressões do domínio são uma das formas de representação de conhecimento sobre textos de uma determinada literatura específica, seja ela textos científicos, notícias, opiniões de produtos ou textos de redes sociais. O enriquecimento de representações de documentos textuais está diretamente ligado à inserção de algum tipo de conteúdo que agregue à essa representação determinado nível de informação privilegiada que uma *Bag of Words*(BoW) não possui.

O trabalho de Marques *et al.* (2015) apresenta uma representação de textos, denominada *Bag of Expressions of Domain* (BoED) e é aplicada especificamente para o domínio da área de Desenvolvimento de Produtos e Serviços, área específica da Engenharia de Produção. Neste trabalho o autor considera aspectos semânticos denominados Expressões do Domínio. A nova representação é aplicada na classificação de documentos do mesmo domínio e são necessárias listas de termos e expressões que pertencem à área do domínio. Tais listas são descritas a seguir.

1. *Lista de métodos e ferramentas:* a primeira lista é composta por nomes de métodos e ferramentas de Sistemas Produto-Serviço e que são de interesse do usuário. Essa lista é definida como sendo o conjunto M que contém os nomes dos k métodos ou ferramentas e seus respectivos sinônimos (s_i):

   $M = \{m_1(s_{11},\ldots,s_{1i}), m_2(s_{21},\ldots,s_{2i}),\ldots,m_k(s_{k1},\ldots,s_{ki})\}$.

2. *Lista de palavras de aplicação:* a segunda lista é composta por palavras ou expressões que os autores utilizam para indicar que um método ou ferramenta foi aplicado. Essa lista é definida como sendo o conjunto A que contém as p expressões que indicam a aplicação de um método ou ferramenta e seus respectivos sinônimos (s_i):

   $A = \{a_1(s_{11},\ldots,s_{1i}), a_{21}(s_{21},\ldots,s_{2i}),\ldots,a_p(s_{p1},\ldots,s_{pi})\}$.

3. *Lista de palavras de desenvolvimento teórico:* a terceira lista é composta por palavras ou expressões que os autores utilizam para apresentar o desenvolvimento teórico de um método ou ferramenta em particular. Essa lista é definida como sendo o conjunto T que contém as q expressões que indicam o desenvolvimento teórico de um método ou ferramenta e seus respectivos sinônimos (s_i):

   $T = \{t_1(s_{11},\ldots,s_{1i}), t_{21}(s_{21},\ldots,s_{2i}),\ldots,t_q(s_{q1},\ldots,s_{qi})\}$.

Segundo Marques *et al.* (2015), a tarefa de geração das três listas pode ser feita de forma manual ou utilizando técnicas de Mineração de Textos. Quando realizada de forma manual, deve-se selecionar um conjunto de artigos de referência e gerar as três listas. Por meio de Mineração de Textos, considera-se o uso de técnicas de reconhecimento de entidades nomeadas e regras de associação. No entanto, apesar dos resultados serem promissores, algumas deficiências foram identificadas na geração automática dessas listas.

2.3 Extração de termos

A tarefa de extração de termos é muito importante e bastante pesquisada nos dias de hoje. Porém, não existe uma única definição formal sobre o que é "*termo*". Segundo a norma internacional ISO1087 (1990), "*termo*" corresponde à "***designação**, por meio de uma unidade linguística, de um conceito definido em uma língua de especialidade*" (ISO1087, 1990, p. 5). O verbo *"designar"*, segundo o dicionário Michaelis [3] *online* significa *"1 Apontar, indicar, nomear. 2 Assinalar, marcar. 3 Denominar, qualificar. 4 Escolher, nomear. 5 Determinar, fixar. 6 Ser o símbolo de, significar."*

Castellví, Bagot e Palatresi (2001) define "*termo*" como: "***unidade terminológica** obtida a partir de domínio especializado*". Palatresi (2002) também define "*termo*" como "***unidade terminológica** que é utilizada para designar conceitos em um âmbito tematicamente restrito*". Tais definições são as mais aceitas na literatura. Barros (2004, p. 40) define "*termo*" como sendo "***unidade lexical** com um conteúdo específico dentro de um domínio específico*", sendo que "*unidade lexical*" corresponde ao "*símbolo linguístico, composto de expressão e de conteúdo, que pertence a uma das grandes classes gramaticais (substantivo, verbo, adjetivo ou advérbio)"*. Com base nessas definições, nesta tese é considerado que *"termo"* (ou unidade terminológica) é uma unidade lexical utilizada para designar conceito em um cenário tematicamente restrito, ou seja, um conceito dentro de um domínio.

Na maioria das pesquisas encontradas na literatura sobre termos, os autores afirmam que estes geralmente são unidades nominais, uma vez que designam conceitos (por exemplo, para denominar/dar um nome a um conceito) (SAGER, 1990; BARROS, 2004; BATISTA, 2011). Outros autores consideram termos como sendo estruturas sintáticas como `substantivos` ou `adjetivos` (ALMEIDA; VALE, 2008), podendo ser também `verbos` (FRANTZI; ANANIADOU; TSUJII, 1998; CASTELLVÍ, 1999; ZAVAGLIA *et al.*, 2007), por exemplo, no domínio de Informáticao o verbo "*formatar*" pode ser considerado um termo. Apesar de diversas estruturas sintáticas poderem consideradas como termos, de maneira geral, os autores também consideram que a quantidade de `substantivos` presentes em domínios específicos é extremamente desproporcional em relação aos `adjetivos` e `verbos`.

Devido à dificuldade de definir de forma única o que é um "*termo*", a extração de termos torna-se uma tarefa não trivial em Mineração de Textos (ESTOPÀ; BURGOS *et al.*, 2006). Outra dificuldade associada a esta tarefa é o fato de que os termos dependem do domínio o qual estão inseridos. Para ilustrar as principais dificuldades e características da extração de termos são apresentados algum exemplos de sentenças, destacando os termos, relacionadas a diferentes domínios como Educação à Distancia (EaD[4], nanociência e nanotecnologia (Nano[5] e, por último,

[3] Dicionário Michaelis *Online*: https://michaelis.uol.com.br
[4] Exemplos retirados do trabalho de (SOUZA; FELIPPO, 2010)
[5] Exemplos retirados do trabalho de (ALMEIDA; VALE, 2008)

revestimentos cerâmicos (RC[6]):

(a) **Ensino à Distância:** "Um dos principais pontos que consideramos bem relevante para configuração de [[**ambientes virtuais**] **de aprendizagem**] é o design simples e fácil. (...) O [**ciberespaço**] é muito mais que um [**meio de comunicação**] ou mídia. (...) criar e produzir [**material didático**] **impresso**] para EAD é uma alternativa necessária".

(b) **Nanociência e Nanotecnologia:** "Um [**nanômetro**] equivale à bilionésima parte de um metro e qualquer medida nessa escala é invisível a olho nu."

(c) **Revestimentos Cerâmicos:** "As [**fritas**] são utilizadas na atualidade como principais constituintes dos [**esmaltes**] empregados na fabricação nacional de revestimentos cerâmicos."

Nos exemplos apresentados, as principais características das unidades em destaque estão relacionadas aos significados em cada domínio do conhecimento em que elas são aplicadas. Em revestimento cerâmico (subdomínio da Engenharia de Materiais), por exemplo, os termos "**frita**" e "**esmalte**" tem um significado muito específico no discurso técnico. Nesse domínio, "frita" significa "vidro moído obtido a partir da fusão e da mistura 'de diferentes ingredientes, como boratos, potássio, soda, cal, alumina, etc."[7] e "esmalte" significa "cobertura de aspecto semelhante ao vidro, impermeável, branca, colorida, transparente ou opaca, que é aplicada sobre a placa cerâmica como decoração e/ou proteção"[8]. Quanto à estrutura mórfica os termos podem ser: (i) lexias simples, isto é, formadas por apenas um radical, com ou sem afixos, ou (ii) lexias complexas, isto é, formadas por mais de um radical (ISO1087, 1990, p. 7). No Quadro 3, são mostrados alguns exemplos de termos de diferentes domínios.

Quadro 3 – Exemplos de termos simples e complexos de diferentes domínios.

		Domínios
Termos Simples	frita esmalte	revestimento cerâmico
	ciberespaço interatividade	educação à distância
	nanômetro ácido	nanociência e nanotecnologia
Termos Complexos	revestimento cerâmico resistência mecânica	revestimento cerâmico
	meio de comunicação ambiente virtual (de aprendizagem)	educação à distância
	escala nanométrica potência óptica	nanociência e nanotecnologia

Fonte: Elaborada pelo autor.

[6] Exemplos retirados do trabalho de (ALMEIDA *et al.*, 2011)
[7] Definição retirada de Almeida *et al.* (2011, p. 32)
[8] Definição retirada de Almeida *et al.* (2011, p. 27)

Os termos que constituem lexias simples podem apresentar alguma característica morfológica que os diferencia das unidades lexicais que são usadas na língua geral ou em outro domínio especializado. No termo *"ciberespaço"* (a), por exemplo, identifica-se *"cíber-"*, abreviatura de "Cibernética", utilizada na denominação de vários conceitos da EAD que se relacionam à Internet, ou seja, mundo ou espaço virtual (GIANOTI, 2012). O termo "*nanômetro*" (b) é outro exemplo paradigmático de lexia simples cuja estrutura morfológica revela seu estatuto termiIiológico. Esse estatuto se deve à presença do prefixo "*nano-*"[9], empregado para indicar a escala 10^{-9} da medida indicada (metro) (ALMEIDA; VALE, 2008). Já os termos da medicina, por sua vez, caracterizam-se pela presença de morfemas prefixais (p. ex.: "*artri/o*") e sufixais (p. em.: "*-patia*") de origem grega (ou latina), como em "*artrite*" e "*cardiopatia*", respectivamente.

A denominação dos conceitos, no entanto, nem sempre é feita por unidades lexicais que apresentam alguma marca morfológica que caracteriza o domínio a que pertence. Muitas vezes, utilizam-se palavras com significado especializado que não manifestam particularidades mórficas. Nessa categoria, enquadram-se, por exemplo, os termos "*frita*" e "*esmalte*" (c). Apesar de não possuírem elementos formais de especificidade (sejam eles prefixos, sufixos, etc.). essas unidades são, do ponto de vista conceitual, portadoras de significados especializados muito precisos, como destacado no início deste texto.

Na língua geral, essas unidades são, comumente, genéricas e polissêmicas ("*esmalte*", por exemplo, é definido como "substância transparente que, aplicada em estado líquido, se transforma em película dura e brilhante depois da secagem" e tem três acepções em média nos dicionários de língua portuguesa). Por outro lado, nos domínios de especialidade, essas mesmas unidades apresentam significado específico e comumente não são polissêmicas.

Lexias complexas são as estruturas mais frequentemente utilizadas na denominação dos conceitos em domínios especializados e são formadas por diferentes estruturas formais denominados padrões morfossintáticos. Para o idioma português, a estrutura mais frequente é `substantivo+adjetivo`, como em (a), "ambiente virtual" e "material didático" e em (c), "revestimento cerâmico" . Em inglês, considera-se `adjetivo+substantivo`. Outros padrões morfossintáticos que comumente caracterizam os termos complexos são: (i) `substantivo+adjetivo+preposição+substantivo` (p. ex.: "*ambiente virtual de de aprendizagem*" em (a)), (ii) `substantivo+preposição+substantivo` (p. ex.: "*meio de comunicação*" em (a)), e (iii) `substantivo+adjetivo+adjetivo` (p. ex.: "*material didático impresso*" em (a)), etc.

Outra característica das lexias complexas é a especialização do termo genérico. Como exemplo, cita-se o termo "*ambiente virtual*", que se aplica aos sistemas computacionais. A partir da especialização desse termo, gerou-se o novo termo "*ambiente virtual de aprendizagem*". No que diz respeito às lexias simples, as dificuldades recaem principalmente na identificação de

[9] "*Nano-*" prefixo grego que remete a "*nánnos*", ou seja, de expressiva pequenez, ou "*nânos*", que equivale a um multiplicador 10^{-9} da unidade de medida indicada. Assim, um "*nanômetro*" corresponde a 10^{-9} metros (nm 10^{-9} m (HOUAISS; VILLAR; FRANCO, 2001)

candidatos que não possuem marcas morfológicas de especificidade que indicam o seu potencial terminológico, ou seja, de candidatos que são utilizados também na língua geral por um não-especialista, como "*esmalte*". A utilização de critérios como a frequência, ou seja, o fato de se reencontrar sempre a unidade, podem ser úteis para definir um termo como candidato dentro em textos de domínio. Entretanto, o uso da frequência não obtém um resultado totalmente satisfatório, uma vez que, por exemplo, há alguns candidatos bastantes frequentes mas que não pertencem ao domínio em questão. Por exemplo, "*se refere*", "*definido como*" e "*nós*" podem aparecer frequentemente no *corpus*, mas provavelmente não são termos. Ou, ainda, há candidatos raros no *corpus* mas que pertencem ao domínio. Por exemplo, a sigla "*LMS*" (Ambiente Virtual de Aprendizagem, do ingles *Learning Management System*) aparece raramente no *corpus* do domínio de EAD e é considerada termo desse domínio.

Os critérios propostos para a identificação e delimitação dos termos complexos baseiam-se no grau de lexicalização, que, por sua vez, determina os limites da unidade sintagmática. Segundo Barros (2004, p. 103), para a seleção dos termos a partir de um conjunto de candidatos, algumas características são observadas pelos especialistas para a identificação e delimitação dos termos.

(a) "Não autonomia de um componente em relação aos outros que compõem a unidade léxico-semântica sem que haja modificação de sentido; Exemplo: "*quinta*" e "*feira*" em "*quinta-feira*";

(b) Impossibilidade de comutação de um componente sem acarretar mudança de sentido. Exemplo: "*mesa-redonda*" (tipo de debate) / "*mesa quadrada*";

(c) Não-separabilidade dos componentes. Exemplo: "*terra fina*" / "*esta terra é fina*";

(d) Particularidade da estrutura interna. Exemplo: ausência de determinação significa integração dos elementos constitutivos: "*ter medo*", "*fazer justiça*"."

Além dessas características, outros critérios podem ser aplicados, como a comutação sinonímica. Segundo esse critério, a possibilidade de comutar "*estrada de ferro*" por "*ferrovia*" indica que "*estrada de ferro*" é uma unidade terminológica em potencial. Outro critério importante para a verificação do grau de lexicalização de um sintagma e a frequência de coocorrências, ou seja, o fato de se reencontrar sempre a mesma associação de palavras no domínio de estudo pode ser considerado uma sugestão de lexicalização do sintagma. Por fim, os candidatos identificados com base em ao menos um dos critérios mencionados são selecionados e enviados ao especialista juntamente com os candidatos a termos simples.

A extração de termos é tradicionalmente baseada em três abordagens: estatística, linguística e híbrida. As abordagens caracterizam-se pelo tipo primordial de conhecimento utilizado na referida tarefa. Na Subseção 2.3.1, é apresentada com maiores detalhes cada uma das abordagens (GAIKWAD; CHAUGULE; PATIL, 2014; PAZIENZA; PENNACCHIOTTI; ZANZOTTO, 2005).

2.3.1 Abordagem estatística

Abordagens estatísticas de extração de termos utilizam o conhecimento obtido por meio da aplicação de medidas estatísticas. De maneira geral, é aplicado um pré-processamento envolvendo conceitos apresentados na Subseção 2.1.1, como a *tokenização*, remoção das *stopwords* e a representação dos textos em uma matriz do tipo *Bag of Words* (BoW). Para a tarefa de extração de termos, a representação que pode ser utilizada é baseada em n-grama[10]. Na matriz cada linha representa um documento (D_i) e cada coluna representa um n-grama do documento (N_j). Como verificado em seções anteriores, a aplicação de medidas estatísticas por meio de uma BoW ignora qualquer informação estrutural sobre as sentenças dos textos, como a ordem em que os n-gramas ocorrem. A partir dos valores obtidos pela medida escolhida, os candidatos a termos são ranqueados e aqueles com pontuação mais elevada têm maior probabilidade de serem termos do domínio (GAIKWAD; CHAUGULE; PATIL, 2014; PAZIENZA; PENNACCHIOTTI; ZANZOTTO, 2005).

As medidas comumente utilizadas no desenvolvimento de extratores (semi)automáticos segundo a abordagem estatística são independentes de idioma. A independência de idioma é uma característica vantajosa do ponto de vista computacional, pois a aplicação das medidas não requer a especificação (manual ou automática) de qualquer tipo de conhecimento (p. ex.: morfológico, sintático, etc.) sobre o idioma dos textos em processamento, o que torna a extração (semi)automática mais simples e rápida. Em comparação à extração humana, a independência de idioma não reflete o processo realizado pelos especialistas do domínio, já que estes utilizam o conhecimento linguístico para identificar termos. Um tipo de conhecimento linguístico é o morfológico, utilizado, por exemplo, para identificar termos compostos por morfemas greco-latinos (p. ex.: *artr(i/o)*) "artrite" e "osteoartrite").

A principal limitação de extratores de termos que utilizam abordagem estatística é o "silêncio", ou seja, a não identificação e seleção de termos reais em um texto ou *corpus*. Um exemplo desse problema é quando a medida escolhida utiliza a frequência de cada termo no *corpus* com base para a extração de termos e, por isso, um determinado termo (p. ex., "polinização", do domínio de ecologia) não é extraído por possuir baixa frequência.

Segundo Kageura e Umino (1996), as medidas estatísticas buscam identificar duas propriedades terminológicas: *unithood* e *termhood*. As medidas que expressam *unithood* revelam a força ou estabilidade de expressões complexas (isto é, formadas por dois ou mais elementos separados por espaços em branco). As medidas que expressam *termhood* revelam, por sua vez, o grau de relação entre uma expressão linguística e um domínio do conhecimento. Em outras palavras, *termhood* expressa o quanto uma expressão linguística (seja ela simples, como "*polaridade*", ou complexa, como "*molécula orgânica*" e "*molécula de água*") está relacionada a um domínio.

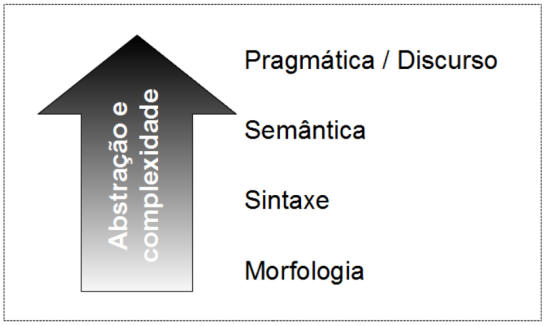
[10] Na abordagem estatística, um n-grama é uma sequência de *n tokens* (por exemplo, unigrama, bigrama, trigrama, etc.).

2.3.2 Abordagem linguística

Na abordagem linguística, os candidatos a termos são identificados e extraídos de um corpus com base em suas características ou propriedades linguísticas, as quais podem ser de diferentes tipos ou níveis.

Conforme ilustado na Figura 9, a área de Processamento de Linguagem Natural (PLN) baseia-se em uma hierarquia de tipos de conhecimento linguístico, elaborada com base em uma escala de abstração e complexidade, ou seja, quanto mais alto for o nível nessa escala, mais complexos será a modelagem e o tratamento computacional do conhecimento. No nível mais inferior dessa escala está o conhecimento morfológico, seguido pelo conhecimento sintático, semântico e pragmático-discursivo (MARTIN, 2009).

Figura 9 – Níveis linguísticos de conhecimento.



Fonte: Adaptada de Martin (2009).

O conhecimento de nível morfológico, baseia-se na existência de morfemas que indicam a ocorrência de um possível termo. Em medicina e domínios correlatos, por exemplo, um possível morfema, de origem grega ou latina é *artr(i/o)*, como em "*artrite*" (VIVALDI; RODRÍGUEZ, 2007). Em textos do domínio da nanociência e nanotecnologia, é possível detectar candidatos a termos com base na identificação de morfemas como *nano-*. Esse prefixo compõe inúmeros termos simples, como "*nanotubo*", e complexos, como "*nanotubo de carbono*", desse domínio (ALMEIDA; VALE, 2008).

O conhecimento de nível sintático refere-se à ordem e função dos termos nas sentenças. Comumente é feito por meio da identificação da estrutura sintagmática das sentenças, a partir da qual os sintagmas nominais são selecionados como candidatos a termos. Por exemplo, na sentença "*Moléculas de ácido silícico condensam com formação de água.*", "*molécula*", por ser núcleo de um sintagma nomina1, é identificada como um candidato a termo.

Extração de candidatos de acordo com o nível semântico refere-se ao significado ou conceito subjacente aos termos, por exemplo "*mundo*" é refenrenciado por <concreto.lugar>. Na literatura geral, há poucos trabalhos que utilizam conhecimento mais abstrato, como o de nível semântico (PALATRESI, 2002). Quanto ao conhecimento de nível pragmático, desconhecem-se trabalhos que buscam identificar candidatos a termos por meio de propriedades referentes ao uso.

De um modo geral, a extração de termos é realizada segundo a abordagem linguística baseada em conhecimento de nível morfossintático (PAZIENZA; PENNACCHIOTTI; ZANZOTTO, 2005). No caso, busca-se realizar a Extração automática de termos (EAT) por meio de (i) a categoria sintática dos n-gramas presentes no *corpus* (por exemplo: `verbo`, `substantivo`, `adjetivo`, etc.) (ii) padrões moforssintáticos (por exemplo: `substantivo+adjetivo`, `substantivo+preposição+substantivo`).

Os padrões morfossintáticos são recursos bastante utilizados. Termos de domínios específicos apresentam uma estrutura interna que seguem determinados padrões. No domínio da nanociência e nanotecnologia é possível verificar estruturas morfossintáticas compostas por `substantivo+adjetivo`. Esse padrão, pode fazer referência a estruturas termos como em "*material nanoestruturado*". Outro padrão morfossintático encontrado é `substantivo+preposição+substantivo`. Esse padrão corresponde a termos como "*nanotubo de carbono*".

Em abordagens linguísticas, assim como a abordagem estatística, o pré-processamento também envolve a sequência de tarefas *tokenização* e remoção das *stopwords*. Além disso, outras tarefas são aplicadas como etiquetação morfossintática (do inglês, *part-of-speech tagging*), que consiste na associação, a cada palavra do *corpus*, de uma etiqueta que indica a sua categoria sintática (p. ex.: "*nanotubo*_`substantivo` de_`preposição` *carbono*_`substantivo`) (VOUTILAINEN, 2003) e normalização das palavras dos textos, que consiste em uniformizá-las por meio da redução de suas variações. Na literatura, são encontradas três técnicas de normalização, a saber: radicalização, lematização e substantivação. Tais técnicas são apresentadas na Subseção 2.1.1.

Quando a utilização de conhecimento linguístico é baseada na identificação de sintagmas, os extratores comumente realizam o reconhecimento da estrutura sintática (do inglês, *parsing*) das sentenças, atribuindo funções sintéticas aos constituintes reconhecidos (MITKOV, 2022), por exemplo. sujeito e predicado, sintagma nominal e sintagma verbal. Em resumo, os extratores linguísticos podem fazer uso de ferramentas de processamento de língua natural como os sentenciadores, tokenizadores, etiquetadores morfossintáticos (em inglês, *taggers*), lematizadores, radicalizadores (do inglês, *stemmer*), substantivadores e analisadores sintéticos (do inglês, *parsers*).

Independentemente do tipo de conhecimento empregado, os resultados obtidos pelos extratores baseados nessa abordagem são em geral melhores do que os obtidos pela abordagem estatística. No entanto, a abordagem linguística também não está livre de problemas. No caso, a extração é dependente de língua, pois a identificação dos candidatos requer a especificação de algum tipo de conhecimento linguístico, como por exemplo a a categoria sintática das palavras. O uso de ferramentas como *taggers*, *parsers* ou lematizadores estão sujeitos a gerar erros frequentes que afetam as tarefas de identificação e extração de candidatos a termos. No processo feito de forma manual, a especificação linguística torna a extração de candidatos mais cara e lenta.

2.3.3 Abordagem híbrida

A abordagem híbrida pauta-se em propriedades estatísticas e linguísticas para a identificação e extração de candidatos a termos. Nos extratores que seguem essa abordagem, a ordem de utilização do conhecimento pode variar. Em alguns extratores, o conhecimento estatístico é utilizado antes do linguístico e, em outros, o conhecimento estatístico é utilizado depois do linguístico.

Segundo Teline *et al.* (2003) e Pazienza, Pennacchiotti e Zanzotto (2005), os melhores resultados são obtidos quando as medidas estatísticas são aplicadas sobre uma lista de candidatos que foram previamente extraídos com base em alguma propriedade linguística, pois a confiabilidade das medidas estatísticas é maior quando estas são aplicadas a candidatos a termo linguisticamente "justificados". Uma das razões para esse fato é que os termos normalmente seguem um padrão pré-definido para cada domínio (substantivos, principalmente). Esse padrão é identificado ao analisar morfossintaticamente o candidato a termo, porem o padrão pode ser diferente dependendo do contexto que o candidato aparece. Por exemplo, a palavra "*segundo*" pode ser um substantivo ("*Alguns segundos são suficientes...*") ou um numeral ordinal ("*O segundo ano...*"). Os métodos estatísticos não consideram esse contexto e é essa uma das razões pelas quais se aconselha primeiramente identificar as propriedades linguísticas dos candidatos e, em seguida, aplicar métodos estatísticos. Além disso, os métodos estatísticos são mais rígidos e podem eliminar os termos que tem baixa frequência, mas que são importantes para o domínio.

2.4 Modelos de linguagem

Modelos de linguagem estatísticos, ou somente modelos de linguagem, descrevem a distribuição de probabilidade de uma sequência de palavras, associando uma probabilidade para cada sentença em um idioma (DAUPHIN *et al.*, 2017). Tais modelos possuem diversos usos como reconhecimento de fala (*speech recognition*), *Part of Speech* (PoS) *tagging*, análise sintática (*parsing*), tradução (*machine translation*), reconhecimento de texto manuscrito (*handwriting recognition*) e recuperação de informação (*information retrieval*).

Por exemplo, considere as seguintes sentenças: s_1 = *"o gato está dormindo"* e s_2 = *"está gato dormindo o"*. Ao comparar a probabilidade das duas sentenças ocorrem na língua portuguesa, tem-se $p(s_1) > p(s_2)$.

Considerando a sentença s_1 definida anteriormente, a distribuição de probabilidade *p(o gato está dormindo) = q(o) q(gato | o) q(está | o gato) q(dormindo | o gato está)*.

Os modelos de linguagem baseados em redes neurais obtiveram performance superior para estimar essa distribuição de probabilidades (BENGIO; DUCHARME; VINCENT, 2000; SCHWENK, 2007; DEORAS; KOMBRINK *et al.*, 2011). Esses trabalhos motivaram o aprendizado de representações de palavras em um espaço de baixa dimensão, também conhecidos como

vetores distributivos (*distributional vectors*) ou *word embeddings*, o qual seguem a hipótese de que palavras similares ocorrem em contextos similares. Dessa forma, os vetores também tentam capturar características da vizinhança de uma palavra e similaridade entre palavras. Esse método se popularizou por meio do trabalho de Mikolov *et al.* (2013), que propôs o *Continuous bag-of-words* (CBOW) e o *Skip-Gram* que permitiam construir *word embeddings* de forma eficiente e com grande qualidade de textos. A seguir serão apresentados alguns métodos para aprendizado de modelos de representação.

- **Word2vec:** em Mikolov *et al.* (2013) foi proposto o *Continuous bag-of-words* (CBOW) e o *Skip-Gram*. O CBOW computa a probabilidade condicional de uma palavra alvo ocorrer dado as palavras que estão a sua volta, ou seja, a partir do contexto em que a palavra ocorre. Já o *Continuous Skip-Gram*, ou só *Skip-Gram*, dado uma palavra alvo, tenta prever as palavras que estão a volta dessa palavra, ou seja, tenta prever o contexto de ocorrência da palavra alvo. A principal contribuição desse trabalho segundo os autores é a possibilidade de treinar *word embeddings* usando grandes conjuntos de dados contendo bilhões de palavras com uma baixa complexidade computacional.

- **GloVe:** *Global Vectors for Word Representation* (GloVe) foi proposto por Pennington, Socher e Manning (2014), com o objetivo de deixar explícito quais propriedades o modelo precisa para que ocorram as regularidades sintáticas, diferentes de outros métodos como *Word2vec* o qual não ficava claro como ocorriam. Para isso, o treinamento consiste em usar estatísticas globais agregadas em uma matriz de co-ocorrência de palavra-palavra em um *corpus*. A principal contribuição, segundo os autores, é a combinação das vantagens das duas grandes famílias para geração de modelos de linguagem e obter bons resultados nas seguintes tarefas: analogia de palavras; similaridade de palavras e tarefas de reconhecimento de entidades nomeadas.

- **FastText:** É uma abordagem baseada no *Skip-Gram*, porém ao invés de uma usar uma BoW, o *FastText* usa uma *bag-of-characters n-grams*. Foi proposto por Pennington, Socher e Manning (2014) com objetivo de suprir uma limitação de modelos comuns os quais ignoram a morfologia da palavra. Uma das principais contribuições do modelo, segundos os autores, é a rápida etapa de treinamento, mesmo em um grande conjunto de dados. Além disso, *FastText* permite computar *word embeddings* para palavras que não estavam presentes no treinamento por meio de uma soma dos characteres *n-grams* que forma a palavra.

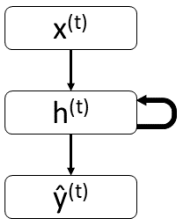
Apesar dos avanços alcançados por essas propostas ainda há limitações. Segundo Lucy e Gauthier (2017), algumas classes de características acabam sendo mal representadas pelos métodos avaliados, faltando elementos fundamentais de semântica e alguns domínios semânticos são particularmente afetados por esses problemas. Outro problema é que as *word embeddings*

geradas por esses modelos são representações vetoriais livres de contexto Peters *et al.* (2017), McCann *et al.* (2017), ou seja, após o treinamento uma palavra terá sempre o mesmo *word vector* independente do contexto que está inserido. Por exemplo, nesses modelos, o *word vector* da palavra "banco" será o mesmo, independente se ocorre no contexto de uma praça ou no contexto de serviços financeiros. Para lidar com esse problema diversos métodos foram propostos, sendo os principais descritos nas próximas subseções.

2.4.1 Redes neurais recorrentes e LSTM

Uma Rede Neural Recorrente (RNR) é qualquer rede neural que contém um ciclo em sua rede de conexões, ou seja, qualquer rede em que o valor de um nó é diretamente ou indiretamente dependente de saídas precedentes como uma entrada (YU *et al.*, 2019). Como é possível observal na representação dana Figura 10, a RNR recebe como entrada um vetor x, que é multiplicado por uma matriz de pesos W, seguido de uma função de ativação para computar h na camada escondida. Nesse tipo de rede, os elementos em uma sequência são processados um a um, introduzindo a noção de tempo (t). Assim, a principal característica da RNR está no nó de recorrência, representado na Figura 10 pelo nó intermediário e uma seta direcional, que adiciona como entrada na computação da camada escondida no passo t, o valor da camada escondida do passo anterior $(t-1)$. Desse modo, dado um vetor $[x^{(1)}, x^{(2)}, \ldots, x^{(T)}]$, a cada passo t, os nós com recorrência recebem uma entrada $x^{(t)}$ e o valor da camada escondida precedente $h^{(t-1)}$. Isso significa que o estado escondido produzido no passo $(t-1)$ influencia a saída $\hat{y}^{(t)}$, pois o valor da camada escondida em t é afetado pelos estados precedentes, semelhante ao que ocorre em um modelo de linguagem.

Figura 10 – Representação simplificada de uma Rede Neural Recorrente (RNR).



Fonte: Adaptada de Yu *et al.* (2019).

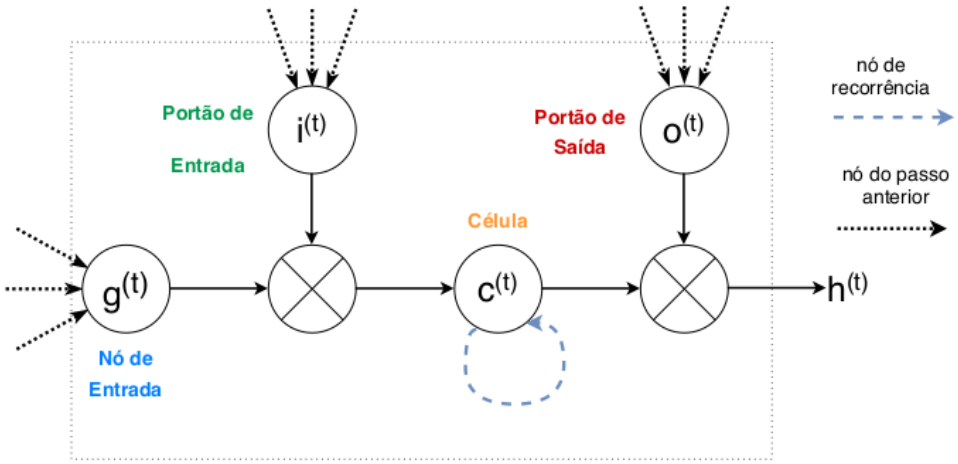
Por sua capacidade de processamento temporal, as RNRs são muito utilizadas na classificação de dados sequenciais, como processamento de texto, fala e sequenciamento de DNA. Um dos problemas enfrentados por elas é o chamado *vanishing gradients*, caracterizado pela perda de informações relevantes que estão muitos passos atrás no tempo t. Isso faz com que as RNRs tenham dificuldades de lidar com análises de dependência de longa distância.

Long Short-Term Memory (LSTM) é um tipo de rede neural recorrente cujo principal objetivo é superar o problema de *vanishing gradients* das RNRs (HOCHREITER; SCHMIDHUBER, 1997). Ela busca esquecer informações consideradas irrelevantes, mantendo somente

aquelas que podem ajudar na predição final. Devido a sua capacidade de filtrar informações contextuais relevantes, as LSTMs foram por muito tempo um dos modelos mais adotados em tarefas de PLN e MT.

A arquitetura das LSTM proposta por Hochreiter e Schmidhuber (1997) possui uma estrutura em cadeia, que contém blocos chamados de células de memória. O componente principal da célula de memória é um nó, chamado de *estado da célula*, que armazena as informações. As informações são reguladas por unidades neurais, chamadas de *portão*. Como é possível observar na Figura 11, as LSTMs possuem dois portões, o *portão de entrada* e o *portão de saída*. O primeiro adiciona as informações úteis no estado da célula e o segundo seleciona quais informações da célula serão enviadas como saída.

Figura 11 – Célula de memória de uma LSTM.



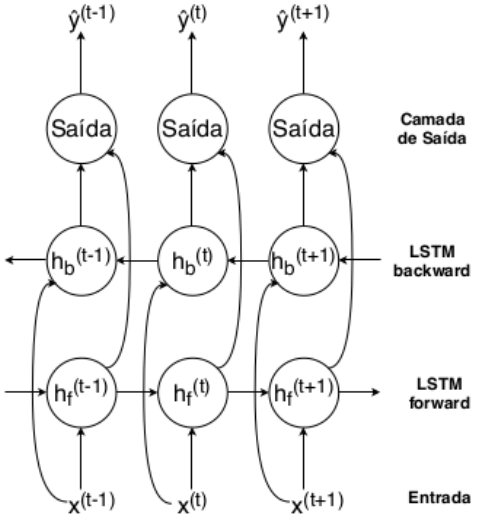
Fonte: Adaptada de Hochreiter e Schmidhuber (1997).

A célula de memória recebe uma entrada $x^{(t)}$ e o estado escondido do passo anterior $h^{(t-1)}$ e computa o valor do nó em $g^{(t)}$. Em seguida, o valor do portão de entrada $i^{(t)}$ é multiplicado a $g^{(t)}$ para selecionar as informações, que são atualizadas no estado da célula $c^{(t)}$. O último passo é multiplicar o estado interno $c^{(t)}$ pelo valor do portão de saída o t para obter o estado da camada escondida $h^{(t)}$. Os estados escondidos $h^{(t)}$ representam as informações contextuais fornecidas até o momento precedente a t, portanto o contexto à esquerda da sequência é usado para calcular a predição. Esse modelo é chamado de *forward*.

O contexto à direita da palavra também pode conter informações relevantes para a classificação da palavra. Esse contexto pode ser capturado por meio de um modelo *backward*, que percorre a sequência da direita para a esquerda. A combinação do modelo *forward* com o modelo *backward* forma uma Rede Neural LSTM Bidirecional (BiLSTM), capaz de combinar as saídas das duas redes em uma única representação que captura informações à esquerda e à direita (YU *et al.*, 2019). A Figura 12 representa uma Rede Neural LSTM Bidirecional (BiLSTM). Nela é possível verificar que, a cada passo t, os estados escondidos da LSTM *forward* e *backward* são combinados através de uma operação como concatenação, multiplicação ou soma. Essas

saídas são usadas como entrada para uma camada de pós-processamento, por exemplo, uma Rede Neural *Feedforward* com ativação *Softmax*, para obter as predições do modelo.

Figura 12 – Rede Neural BiLSTM.

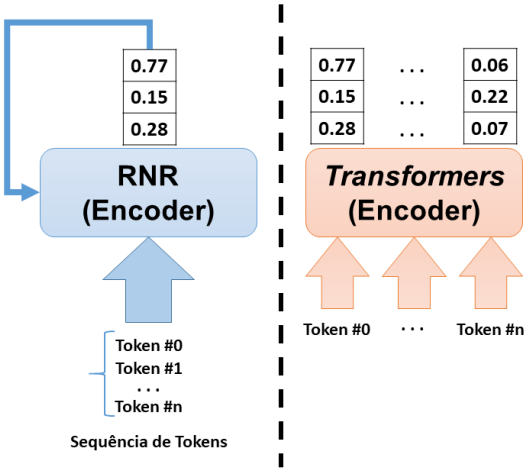


Fonte: Adaptada de Yu *et al.* (2019).

2.4.2 Transformer

Transformer é uma rede neural com arquitetura *encoder-decoder* baseada em mecanismos de atenção. Diferente das RNNs, cujas palavras são apresentadas uma por vez, em sequência, e o estado atual depende do resultado do estado anterior, no *Transformer* todas as palavras da sentença são passadas simultaneamente na forma de *tokens*, de forma que sues respectivos vetores (*embeddings*) sejam obtidos de forma simultânea (VASWANI *et al.*, 2017). Na Figura 13 é posível observar a diferença entre RNRs e *Transformers*.

Figura 13 – Diferença entre RNR e *Transformer*.

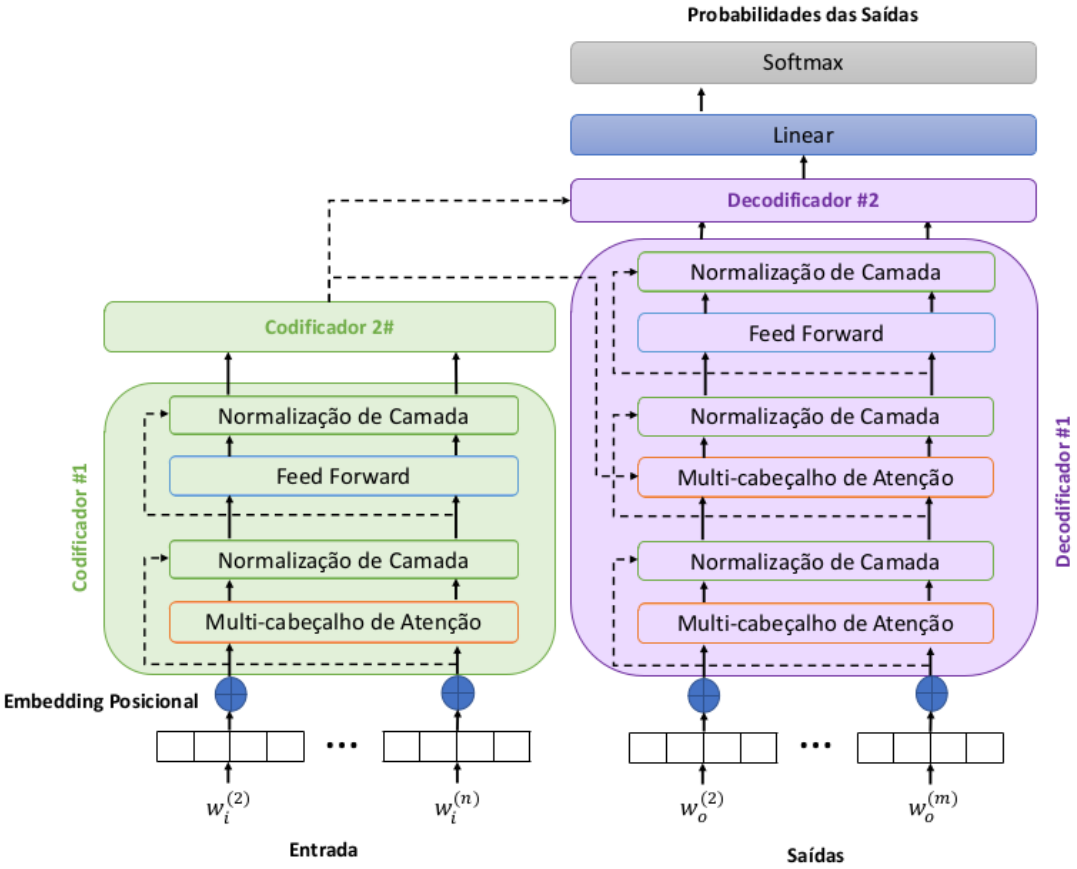


Fonte: Adaptada de Vaswani *et al.* (2017).

A rede recebe como entrada uma sequência de palavras, codifica-as em representações nas camadas de atenção e as decodifica em palavras novamente. À medida que o modelo processa cada palavra (cada posição na sequência de entrada), um mecanismo de atenção computa a importância de palavras em outras posições para codificar a palavra atual. Inúmeros modelos surgiram a partir da rede Transformer, entre eles é possível citar BERT (DEVLIN *et al.*, 2018) e GTP (RADFORD *et al.*, 2018).

Na Figura 14 é possível observar a estrutura da arquitetura do *Transformer*. No módulo *encoder*, cada camada é dividida em duas subcamadas, sendo a primeira um mecanismo chamado "Multi-cabelalho de Atenção" (em inglês, "*Multi-Head Self-Attention*") e a segunda uma rede neural *Feedforward*. Já nas camadas *decoder*, além das duas subcamadas citadas, há uma terceira subcamada de atenção que processa os dados vindos da camada *encoder*.

Figura 14 – Arquitetura de uma Rede Neural do tipo *Transformer*.



Fonte: Adaptada de Vaswani *et al.* (2017).

Nota-se que o modelo é composto por uma sequência de codificadores empilhados. O módulo *encoder* recebe como entrada uma sequência de palavras em determinada língua, por exemplo português, codifica as palavras em representações contextuais e as transfere para o módulo *decoder*. Nele, as representações são processadas passo-a-passo e suas saídas são passadas por uma camada de saída em que as predições do modelo são emitidas, ou seja, é gerada a tradução da sentença para um idioma alvo.

O mecanismo de atenção consiste no mapeamento de uma consulta, ou seja, um conjunto de pares chave-valor para uma saída, cujas consultas, chaves, valores e saídas são todos vetores. A saída é calculada por meio de uma soma ponderada de valores, os pesos associados a cada valores são gerados por uma função de compatibilidade da consulta com sua respectiva chave. De forma intuitiva, quanto maior a compatibilidade de uma chave com uma consulta, maior será o peso do valor associado aquela chave. A Equação 2.9 caracteriza o processo de auto-atenção, e pode ser descrito nos passos abaixo.

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{2.9}$$

O 1° passo do processo do mecanismo de atenção é calcular, para cada *token* de uma cadeia w, os vetores de consulta q, de chave k e de valor v. A criação dos vetores q, k e v se dá, respectivamente, pela multiplicação do vetor de *embedding* do *token* pelas matrizes de pesos W^Q, W^K e W^V, treinadas durante o processo de treinamento do modelo.

O 2° passo é calcular, para cada *token* da cadeia, uma representação com base em todos os *tokens* da cadeia. Essa representação se dá da seguinte forma: Seja w_t o *token* para o qual deseja-se calcular o *score*, com base no seu vetor se consulta associado, multiplica-se seu vetor de consulta associado pelos vetores de chave de todos os *tokens* da cadeia. A intuição por trás deste cálculo de compatibilidade entre consultas e chaves está em identificar o quanto o modelo deve prestar atenção em um conjunto de palavras para descrever uma palavra específica.

O 3° passo é realizar a divisão dos valores resultantes das multiplicações realizadas anteriormente, pela raíz da dimensionalidade do modelo, e então normalizados pela função *Softmax*.

O 4° passo ocorre pela multiplicação dos *scores Softmax* resultantes por seus respectivos vetores de valor associados.

O 5° passo se dá pela soma de todos estes vetores para resultar na representação do *token* w_t.

O *score* de um *token* tende a ser dominado pela chave e valor do mesmo, por isso, na arquitetura deste modelo é feito uso de múltiplos cabeçalhos de auto-atenção, permitindo que o modelo preste atenção em outros *tokens* para representar um *token* específico, aprendendo outras formas de representação que podem ser úteis, como é possível notar na Equação 2.10

$$MultHead(Q,K,V) = Concat(head_1, head_2, head_3, \ldots, head_h) * W^0 \tag{2.10}$$

tal que,

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Em vez de um único núcleo de atenção, a camada de atenção do *Transformer* (“Multi-cabelalho de Atenção”), é composta de oito núcleos de atenção. Assim, são projetadas paralelamente oito matrizes W^Q, W^K e W^V distintas, inicializadas aleatoriamente, então calcula-se a atenção de cada núcleo, como se viu anteriormente. Os vetores gerados são concatenados e multiplicados por uma matriz W^0 para obter o resultado final da camada. A ideia de utilizar inúmeros núcleos de atenção é que cada um dos núcleos esteja focado em diferentes palavras da sentença, o que melhora a codificação da palavra.

2.4.3 Bidirectional encoder representations from transformers (BERT)

Bidirectional Encoder Representation from Transformers (BERT) é uma rede neural que gera representações bidirecionais de palavras, ou seja, baseado no contexto à esquerda e à direita. Treinado usando grandes conjuntos de textos e que permite posteriormente ser refinado para uma tarefa específica (DEVLIN *et al.*, 2018). Para aprender representações de textos, o BERT faz o uso somente do encoder do Transformer com a implementação quase idêntica a do original (VASWANI *et al.*, 2017).

Existem duas arquiteturas do modelo, distintas pelo número de camadas *Transformer*, número de núcleos de atenção e tamanho da camada escondida. O $BERT_{LARGE}$ conta com 24 camadas, 16 núcleos de atenção e tamanho 1024 na camada escondida (DEVLIN *et al.*, 2018). Já o $BERT_{BASE}$ é composto de 12 camadas, 12 núcleos de atenção e tamanho 768 na camada escondida (XU *et al.*, 2019; SUN; HUANG; QIU, 2019).

O BERT pode receber como entrada uma sentença[11] ou um par de sentenças concatenados em uma sequência. As palavras da entrada são convertidas em *word embeddings* com base no vocabulário *WordPiece embeddings* (WU *et al.*, 2016), com mais de 30.000 *tokens*. O processo de conversão das palavras em *word embeddings* quebra as palavras que não existem no vocabulário, inserindo o símbolo nas partes não iniciais. Para exemplificar, se não existir a palavra “*playing*” no vocabulário então tal palavra poderia ser dividida em “*play*” e “*ing*”, considerando que existe um token denominado “*ing*” no vocabulário. Dos 30.000 *tokens* do vocabulário, há 5 *tokens* de uso especial, detalhados a seguir. A numeração indica o *id* de cada *token*:

102 - [CLS]: *Token* inserido no início da sequência. Usado na etapa de classificação para concatenar as representações individuais de palavras em uma sentença.

103 - [SEP]: *Token* usado para separar entradas compostas por um par de sentenças. Delimita a fronteira das sentenças.

1 - [PAD]: *Token* usado para preenchimento para lidar com sequências de tamanhos diferentes. Se é esperado que todas as sentenças tenham quatro tokens, a frase “dia ensolarado” ficaria

[11] Considera-se “sentença” uma sequência de palavras ou *tokens*, sem associação com a definição linguística.

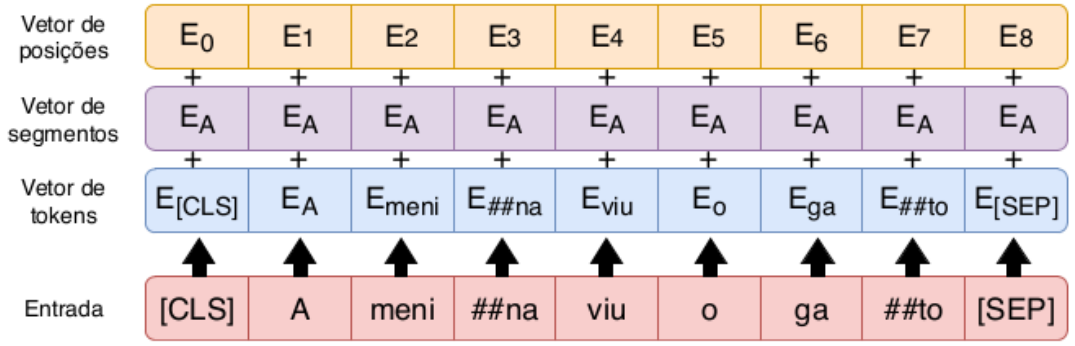
"dia ensolarado [PAD] [PAD]".

101 - [UNK]: Usado para indicar um *token* que não está no vocabulário.

104 - [MASK]: *Token* usado para mascarar valores. Este é o token usado no lugar da palavra que o modelo tentará prever em determinadas situações.

Após a preparação do conjunto de entrada, treina-se um vetor de segmentos indicando se a palavra pertence à sentença A ou à B. Por fim, assim como no *Transformer*, o BERT não captura a ordem de palavras, portanto necessita de um vetor de posições, nesse caso computando a posição absoluta da palavra na sentença. A entrada final do BERT é dada pela somatória dos vetores de *word embedding*, de posição e de segmento, como pode-se observar na Figura 15 cuja entrada é a sentença em português "*A menina viu o gato*".

Figura 15 – Exemplo de entrada no BERT.

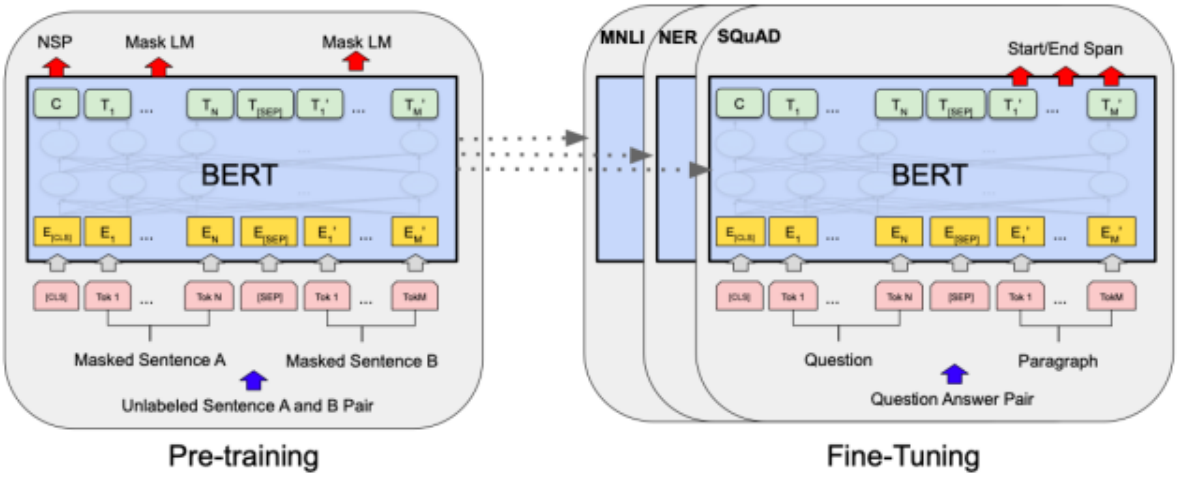


Fonte: Adaptada de Devlin *et al.* (2018).

Duas etapas são necessárias para a fase de treinamento do BERT: o pré-treinamento e o refinamento (do inglês, *fine-tuning*). O pré-treinamento é feito em corpora não-anotados em duas tarefas distintas: *Masked Language Model* (MLM) e *Next Sentence Prediction* (NSP). Por sua vez, o refinamento é treinado em corpora anotados para uma tarefa específica, como Reconhecimento de Entidades Nomeadas (REN), tradução automática, etc. A Figura 16 traz uma representação das etapas de pré-treinamento e de refinamento do modelo.

Um dos diferenciais do BERT está em sua etapa de pré-treinamento. Nela, o modelo é otimizado ao mesmo tempo em duas tarefas, uma delas um modelo de linguagem (ML) bidirecional. Em MLs bidirecionais anteriores, eram necessários dois modelos, um *forward* e outro *backward*, para lidar com o contexto à esquerda e à direita em uma sequência. Diferentemente, o BERT faz isso utilizando somente um modelo de linguagem, chamado de *Masked Language Model* (traduzindo, modelo de linguagem mascarado) com capacidade de observar todas as palavras em uma sequência para predizer qual está faltando. As subseções seguintes fornecem uma descrição das duas tarefas envolvidas no pré-treinamento do BERT.

Figura 16 – Representação da etapa de treinamento do BERT.



Fonte: Devlin *et al.* (2018).

Considerando a tarefa *Masked Language Model* (MLM), Devlin *et al.* (2018) utilizam um modelo de linguagem com máscaras, em que a tarefa é mascarar alguns *tokens* da entrada, substituindo-os pelo *token* especial [MASK] e predizer quais são eles com base nos demais *tokens* da sentença. Em seus experimentos, os autores mascaram aleatoriamente 15% dos *tokens* da entrada. Se o *token* da posição i é escolhido, ele é substituído em 80% dos casos por [MASK], em 10% por um *token* aleatório e nos outros 10% ele permanece o mesmo. Essas condições são estabelecidas como forma de diminuir a disparidade entre o pré-treinamento e o refinamento, já que o *token* [MASK] só aparece na primeira etapa.

A tarefa de predição da próxima sentença, *Next Sentence Prediction* (NSP), é voltada para a relação entre sentenças, calculando a probabilidade de uma sentença B estar imediatamente à direita de A. Para isso, ao selecionar um par de sentenças para treinamento, a chance de B seguir A é de 50%, recebendo a etiqueta "IsNext", enquanto nos outros 50%, B é selecionado aleatoriamente no *corpus*, sendo etiquetado como "NotNext".

Como já mencionado, o BERT pré-treinado pode ser refinado em uma tarefa específica. Nessa abordagem, treina-se o modelo em um *corpus* anotado, otimizando os pesos aprendidos. As predições são obtidas por uma camada de saída para classificação na tarefa em questão, que recebem como entrada as saídas do BERT. O refinamento é bem mais barato em termos computacionais do que o pré-treinamento, já que os pesos do BERT já passaram por treinamento, necessitando somente de ajustes para se adaptar na tarefa.

Além da opção acima, é possível utilizar o BERT como *word embeddings* em outros modelos, extraindo as representações de palavras geradas na etapa de pré-treinamento. Essa abordagem é conhecida como *feature-based*, uma técnica de transferência de aprendizado. Em vez de refinar o BERT em uma tarefa específica, retreinando os seus pesos, os traços do modelo pré-treinado são extraídos e podem alimentar outro modelo de classificação, mantendo os pesos

congelados, sem refinamento.

Devlin *et al.* (2018) apresentam os resultados do BERT na abordagem *feature-based* para o REN, extraindo os traços gerados no pré-treinamento e passando por uma rede BiLSTM. Os autores fazem os experimentos com o $BERT_{BASE}$, extraindo traços de diferentes camadas, como, por exemplo, somente a última ou as quatro finais, entre outras. Os resultados reportados por eles se comparam àqueles obtidos pelo modelo refinado, mostrando que as representações geradas pelo BERT são acuradas mesmo sem refinamento.

2.5 Considerações finais

Neste capítulo foi apresentada a fundamentação teórica dos principais temas que envolvem esta tese de doutorado. Inicialmente foram apresentados os principais conceitos relacionados à análise de textos, como Mineração de Textos e seu processo de desenvolvimento, que são essenciais para a contextualização deste trabalho. Foi dada uma explicação mais aprofundada para as etapas de Pré-processamento e Extração de Padrões, com as técnicas de representação de dados textuais e classificação automática de textos, que são os pilares do desenvolvimento deste trabalho.

Neste capítulo também foram definidos os aspectos semânticos em textos, bem como os níveis de complexidade semântica existentes em tarefas de classificação textual. Nesta linha também são apresentados as principais representações de textos utilizadas e técnicas de enriquecimento semântico, como por exemplo as expressões do domínio.

Expressões do domínio é uma das principais técnicas utilizadas como forma de enriquecimento semântico de representações de textos. Neste capítulo houve um detalhamento sobre o assunto que envolve expressões do domínio, bem como os principais trabalhos da literatura que abordam e desenvolvem o tema.

Por último, foram apresentados os conceitos que extração de termos, que também é base para uma parte do desenvolvimento desta tese, bem como as principais abordagens utilizadas na literatura. Em seguida, é feita uma apresentação dos conceitos de Modelos de Linguagem, Redes neurais recorrentes, LSTM, Transformer e BERT, ou seja, a base conceitural de redes neurais que engloba a extração de termos. Estes temas são base para a extração de termos e construção de listas de termos de forma mais automatizada.

CAPÍTULO

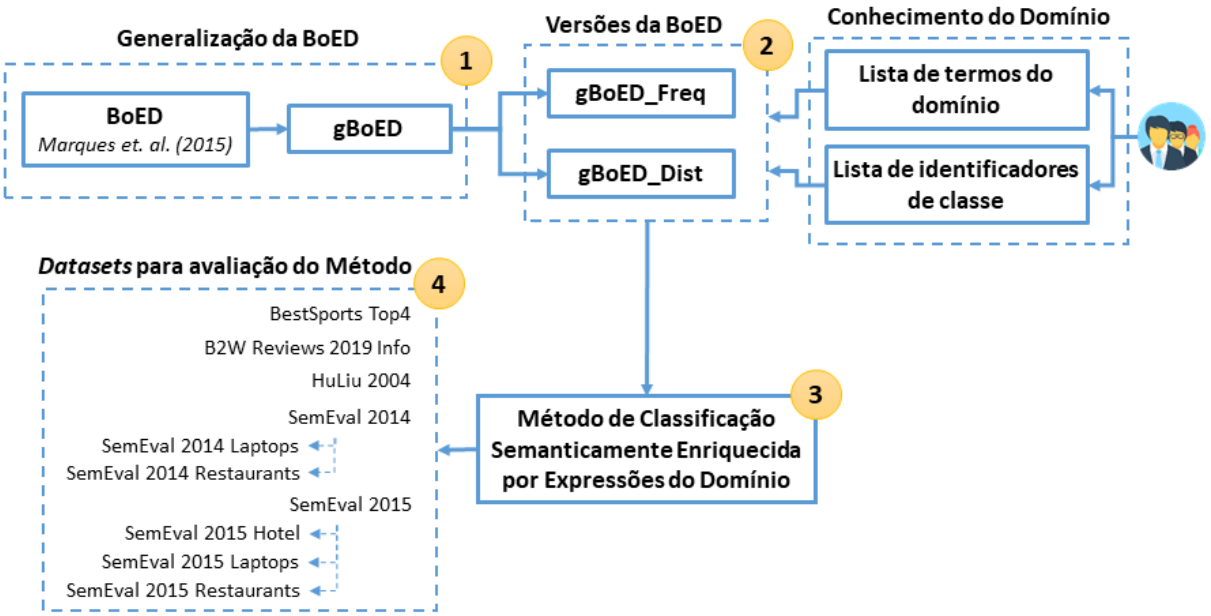
3

MÉTODO DE CLASSIFICAÇÃO SEMANTICAMENTE ENRIQUECIDA POR EXPRESSÕES DO DOMÍNIO

3.1 Considerações iniciais

Nesse capítulo são apresentados os primeiros trabalhos desenvolvidos nesta pesquisa de doutorado, ilustrados nas 4 etapas do diagrama da Figura 17.

Figura 17 – Diagrama da sequência de trabalhos desenvolvidos no capítulo 3.



Fonte: Elaborada pelo autor.

Na etapa 1, com base na literatura e no trabalho de Marques *et al.* (2015), realizou-se

o desenvolvimento da generalização da representação *Bag of Expressions of Domain* (BoED), transformando-a e *generalized Bag of Expressions of Domain* (gBoED) para que possa ser aplicada em diferentes domínios do conhecimento. Nesta etapa também foi executada uma avaliação experimental para testar a viabilidade de se utilizar a representação gBoED em tarefas de classificação de nível semântico, com o objetivo de melhorar os desempenho em relação ao modelo tradicional de treinamento com BoW. Essa etapa é apresentada com maiores detalhes na Seção 3.3.

Em seguida, na etapa 2, é desenvolvida uma nova versão da representação *gBoED* com métrica baseada em distância de termos, que visa melhorar a qualidade da representação enriquecida por expressões do domínio com outros tipos de métricas. Nestas versões, tanto gBoED_Freq quanto gBoED_Dist são construídas de forma manual por especialistas de cada domínio. Esta etapa é apresentada na Subseção 3.3.2. Na etapa 3, apresentada em maiores detalhes na Subseção 3.4.1, é proposto o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio que faz uso das informações enriquecidas para melhorar resultados de classificação de nível semântico.

Por último, ilustrada pela etapa 4 e apresentação com maiores detalhes na Subseção 3.4.2, é realizada uma avaliação experimental do método de classificação da etapa 3, com o objetivo de verificar sua utilidade, aplicabilidade e limitações. Para isso, a avaliação é realizada em 10 diferentes coleções de documentos (*datasets)*, nos idiomas português e inglês. Para cada coleção de documentos são construídas as duas versões do representação enriquecida, gBoED_Freq e gBoED_Dist, a construídas em 3 diferentes cenários. Na configuração experimental são utilizados 4 algoritmos de aprendizado de máquina diferentes voltados para a tarefa de classificação. Em cada um dele são aplicadas uma variação de parâmentros. Toda essa variação de coleções documentos, idiomas, representações, cenários, algoritmos e parâmetros, permite a realização de uma avaliação bastante completa do comportamento do método e das representações nos diferentes cenários. Na Seção 3.2 são apresentados alguns trabalhos relacionados a este trabalho de doutorado.

3.2 Trabalhos relacionados

A busca pelas melhores representações e métodos para obter melhores resultados de classificação de nível semânticos faz com que sujam diversos trabalhos com novas soluções. Na literatura é possível identificar diversos trabalhos que buscam incrementar representações tradicionais, como a BoW, enriquecendo-as com informações adicionais, de modo a permitir melhores resultados em tarefas de classificação de nível semântico. Nessa seção, primeiramente serão relacionados alguns desses trabalhos que possuem o objetivo de enriquecer semanticamente representações tradicionais com o objetivo de melhorar resultados de classificação. Em seguida, serão apresentados alguns trabalhos que utilizam outros tipos de representações semanticamente

enriquecidas.

Rossi e Rezende (2011) propõe uma representação enriquecida do tipo modelo espaço-vetorial, em que cada documento é representado por um vetor e cada palavra da coleção de documentos representa uma dimensão (*features*), de forma análoga a uma representação tradicional BoW. Os termos (*features*) da *Bag of Related Words* são identificados por meio de palavras que se correlacionam dentro de um documento, podem formar expressões como por exemplo "*text mining*", "*document enginnering*". As expressões são extraídas utilizando a técnica de Regras de Associação. Uma métrica baseada na métrica de Suporte é utilizada para podar as regras extraídas e selecionar os os termos e expressões. No trabalho de Rossi e Rezende (2011), a avaliação é realizada em tarefas de classificação por tópicos, usando duas coleções de documentos em inglês. Os resultados obtidos são mais interessantes do que aqueles utilizando BoW.

Seguindo uma linha de pesquisa semelhante, Georgieva-Trifonova (2017) realiza o enriquecimento de representações do tipo modelo espaço vetorial por meio da coocorrência entre termos da representação. A coocorrência é extraída usando a técnica de Regras de Associação. Nesse caso a métrica *Lift* é utilizada para designar a proximidade entre termos. Após a definição de uma matriz de associação entre os termos, ela é utilizada para modificar os pesos da representação BoW definida originalmente. Os experimentos foram realizados usando algoritmo *Support Vector Machine (SVM)* e a coleção de documentos Reuters-21578 composta por documentos de 135 temas diferentes, em inglês. Os resultados apresentam uma melhora na classificação de documentos em nível de tópicos.

Wang *et al.* (2013) propõe uma representação, do tipo modelo espaço-vetorial, que utiliza conhecimentos gerais e conhecimentos de mundo baseada na extração de conceitos da Wikipédia. A principal motivação deste trabalho está diretamente relacionada às limitações do enriquecimento semântico de representações com o uso de ontologias. Segundo Wang *et al.* (2013), nas ontologias a forma mais comum é combinar os conceitos da própria ontologia com termos do documento. Os conceitos das ontologias são usados de modo a substituir informações, ou como recursos adicionais à representação do documento original. A abordagem que substitui o conteúdo original por conceitos de ontologia pode alterar a semântica do documento original e o método que adiciona conceitos de ontologia ao texto original pode introduzir ruídos, especialmente quando a cobertura da ontologia é limitada. Portanto, para que nenhum desses problemas aconteçam é necessário garantir que a ontologia seja grande o suficiente para cobrir o domínio do conjunto de dados da maneira mais completa possível e o método de substituição não deve alterar os significados originais.

A Wikipédia é hoje uma das maiores enciclopédias do mundo e uma grande quantidade de pesquisas utiliza o conhecimento básico contido nela para melhorar a representação de documentos. Ela contém milhões de artigos com cada um explicando um conceito. No trabalho de Wang *et al.* (2013), o conteúdo textual e a estrutura de *links* na Wikipédia foi utilizada

para aprimorar a categorização do texto. Na BoW, os termos do documento foram ligados aos conceitos da Wikipédia adicionando mais significado à representação. A classificação é realizada diretamente no vetor de conceitos construídos a partir da Wikipédia. Um índice invertido ligam os termos aos conceitos vindos do conteúdo textual da Wikipedia. Também utiliza-se *links* entre conceitos para construir uma matriz semântica para entender melhor a relação semântica entre os conceitos. A validação foi realizada com o treinamento de modelos de classificação em 5 coleções de documentos, em inglês, categorizados por nível semântico, sendo 2 relacionados com o domínio de notícias em geral, 1 com o domínio médico, 1 com o domínio de opiniões de filmes e 1 relacionado com o domínio de *Snipets* de busca do Google (textos curtos). Os resultados apresentam melhorias em relação à classificação baseada em BoW.

Em Sinoara, Rossi e Rezende (2016) são propostos 2 tipos de representação de textos, do tipo modelo espaço-vetorial, enriquecidos semanticamente por meio da anotação de papéis semânticos. No primeiro tipo de representação, o conjunto de *features* é composto por palavras combinadas com seus respectivos papéis semânticos em cada sentença. No segundo tipo de representação, cada *feature* corresponde a uma estrutura verbo-argumento encontrada nos documentos, incluindo as palavras e seus rótulos de função semântica. Nesse trabalho, os autores buscam identificar se representações enriquecidas por meio de papéis semânticos melhoram resultados de classificação. As principais coleções utilizadas nos experimentos foram BestSpots Top 4 e SemEval 2015, idiomas português e inglês, também utilizadas na avaliação deste trabalho de doutorado, considerando classificação por nível semântico em ambos. O trabalho mostrou que o uso de informações semanticamente enriquecidas por meio de papéis semânticos melhora a acurácia de modelos de classificação em alguns cenários.

Albitar, Espinasse e Fournier (2014) proprõem duas estratégias para enriquecimento de representações semânticas do tipo modelo espaço-vetorial, aplicadas à tarefas de classificação no domínio médico. Como base de conceitos semânticos, esse trabalho utilizou dois recursos principais. A *Unified Medical Language System (UMLS)*, é uma base de palavras no formato de um dicionário, composta por termos e expressões ligadas à linguagem da saúde e biomedicina, com seus respectivos conceitos. Em seguida o *hsumed corpus*, um *corpus* composto por resumos de artigos biomédicos, desde o ano de 1991, da PubMed/MedLine, é utilizado como coleção de documentos. Foram selecionados os primeiros 20.000 documentos e rotulados em 23 sub-conceitos de doenças.

A primeira estratégia de enriquecimento semântico, chamada de *Bag of Concepts* (BOC) é baseada no método de Kernel Semântico. Nessa estratégia, cada *feature* da BoW é relacionada a um conceito por meio de uma matrix de similaridade. A segunda estratégia de enriquecimento semântico, denominada *Enriching Vectors*, atua entre a etapa de treinamento e predição. Este método enriquece a representação de texto BOC do modelo de classificação e os documentos de teste antes da predição criando uma matriz de proximidade entre os documentos e seus conceitos, com medidas clássicas, como distância de cosseno. Os resultados apresentaram uma melhora no

desempenho dos classificadores ao incrementar seus resultados com informações semânticas e de proximidade entre as classes.

Como é possível observar, diversos trabalhos na literatura visam a construção de métodos para enriquecimento de representações de forma semântica. Os trabalhos apresentados até o momento visam enriquecer representações do tipo modelo espaço-vetorial com diferentes tipos de informações privilegiadas refletindo em melhoria de classificação em diversos cenários.

Outros trabalhos podem ser citados usando diferentes técnicas de enriquecimento em representações do tipo modelo espaço-vetorial, como por exemplo em Ogada, Mwangi e Cheruiyot (2015) que utiliza *n-Grams* como informação enriquecida, Nikishina *et al.* (2022) que utiliza ontologias como enriquecimento por meio de estruturas taxonômicas, Nazir *et al.* (2018) e Moons, Tuytelaars e Moens (2018) que utilizam enriquecimento de representações textuais para melhoria na classificação de imagens. Outras abordagens que utilizam diferentes formatos de representações podem ser citadas, como por exemplo em Rossi *et al.* (2014), Rossi, Lopes e Rezende (2016) e Yan *et al.* (2020) que utilizam a redes complexas como representação ou como enriquecimento semântico.

Segundo Sinoara, Antunes e Rezende (2017) um dos fatores mais importantes para definir a semântica está na identificação do melhor tipo de informação privilegida e da melhor técnica para o domínio o qual o problema está inserido. Nesse trabalho de doutorado, o domínio está diretamente relacionado a tarefas de classificação de nível semântico. Um dos domínios que fazem parte de classificação de nível semântico é a análise de sentimentos. Sendo assim, diversas abordagens de análise de sentimentos podem ser consideradas relacionadas ao tema. No âmbito da análise de sentimentos, a classificação usando *Word Embeddings* e redes neurais são bastante comuns, como em Tang *et al.* (2014), Xiong (2016), Ju e Yu (2018).

El-Din (2016) propõe uma abordagem para enriquecimento da representação BoW para classificação de sentimentos de artigos online. O autor apresenta um método composto por 3 fases principais, sendo a primeira uma fase de *Web Scraping* que obtém os artigos e gera uma BoW a partir deles. Na segunda fase são identificados, por meio de anotações morfossintáticas, algumas palavras-chave contidas nos documentos, como nomes dos autores, nome e sigla da conferência. Na terceira fase é realizada a classificação dos artigos. Uma métrica é associada aos artigos que possuem as palavras-chave. A validação foi realizada usando um total de 1000 artigos extraídos do portal *CiteULike*, em inglês. Foi medida a performance dos classificadores em comparação com a BoW, apresentando uma melhora significativa nos resultados.

Todos os trabalhos apresentados nesta seção mostram a relevância do tema de enriquecimento de representações para melhorar resultados de classificação de nível semântico. Na Seção 3.3 é aprensentada a generalização da representação baseada em Expressões do Domínio.

3.3 *gBoED*: A generalização da *BoED*

O primeiro trabalho desenvolvido nesta pesquisa de doutorado tem como base o trabalho de Marques *et al.* (2015). Nesse trabalho é desenvolvida uma nova representação baseada em Expressões do Domínio (BoED) e aplicada em um processo de Classificação de Textos no domínio de Desenvolvimento de Produtos e Serviços, na Engenharia de Produção para classificar artigos científicos que relatam o desenvolvimento teórico de um método ou a aplicação de métodos já existentes.

Como apresentada na Subseção 2.2.3, a representação baseada em Expressões do Domínio visa carregar consigo um nível semântico maior do que a tradicional *Bag of Words*, devido a união de termos importantes dentro de um determinado domínio e termos que identificam cada uma das classes em um processo de classificação de nível semântico.

Em Marques *et al.* (2015) representação proposta é construída para um domínio específico. Neste trabalho de doutorado foi realizada a generalização da representação BoED. O principal objetivo ao generalizar a representação BoED é permitir melhor qualidade na representação em diferentes domínios do conhecimento, e ainda, busca melhores resultados de classificação de nível semântico. Baseando-se nesses objetivos é desenvolvida proposta de generalização da representação BoED para qualquer domínio do conhecimento, denominada *generalized Bag of Expressions of Domain* (gBoED).

O primeiro passo na generalização da representação BoED para qualquer domínio e problema é a alteração das listas M, A e T, denominadas *Lista de métodos e ferramentas*, *Lista de palavras de aplicação*, *Lista de palavras de desenvolvimento teórico*, repectivamente, apresentadas na Subseção 2.2.3. Tais listas necessitam ser generalizadas para que possam ser utilizadas em qualquer domínio de aplicação e problema.

As listas A e T são compostas por termos e expressões relacionados ao tipo de classificação que se desejava realizar. Em Marques *et al.* (2015) a classificação aplicada possuía o objetivo de identificar trabalhos que apresentavam aplicação prática ou o desenvolvimento teórico de um método na Engenharia de Produção. Nos diferentes domínios do conhecimento tais listas irão conter termos e expressões relacionadas às classes do problema. Por exemplo, no domínio da Análise de sentimentos, é possível classificar como “Positivo” ou “Negativo”, mas também pode-se classificar como “Feliz”, “Triste” ou “Raiva”. Ambas as formas trabalham com classificação de nível semântico e a composição dos termos e expressões podem mudar em cada uma delas. Nesse caso, as listas A e T são generalizadas e recebem o nome de **Listas de Identificadores de Classe**. Elas irão receber os termos importantes para cada classe do problema, podem ser compostas de uma ou mais listas para cada classe.

A lista M corresponde à *Lista de métodos e ferramentas*. Como o próprio nome diz, ela é composta por termos relacionados com métodos ou ferramentas. Por termos relacionados pode-se considerar, nomes de métodos e ferramentas, características, funções, etc. Na Enge-

nharia de Produção, mais especificamente na área de Desenvolvimento de Produtos e Serviços, existe grande interesse em identificar trabalhos que descrevem a aplicação de métodos ou o desenvolvimento de ferramentas que contribuem com o avanço do estado-da-arte nessa área. Portanto, essa lista está diretamente relacionada a um domínio de problema da área, e assim, para generalizar a lista *M* existe a necessidade de uma lista única denominada **Listas de Termos do Domínio**.

Portanto, o primeiro passo para a generalização da construção da representação *Bag of Expressions of Domain* e geração da representação *generalized Bag of Expressions of Domain* é a definição de uma lista de termos do domínio e um conjunto de listas de identificadores de classe. Tais elementos são descritos e formalizados seguir.

- *Lista de Termos do Domínio (Domain Keywords):* formada por termos ou expressões que são importantes para aquela coleção de documentos e para a organização ou classificação esperada como resultado do processo de Mineração de Textos.

  $$Domain\_Keywords = \{k_1, k_2, \ldots, k_i\}$$

  Cada elemento da lista *Domain_Keywords* é formado por um termo do domínio t e seus sinônimos s, isto é, $k_i = \{t_i\} \cup \{s_1, \ldots, s_j\}$.

- *Conjunto de Listas de Identificadores de Classe (Class Keywords):* formado por uma ou mais listas de palavras ou expressões que estão particularmente ligadas a uma determinada classe e, assim, são consideradas como termos ou palavras-chaves daquela classe. O número de listas de identificadores de classe pode variar de acordo com a coleção de documentos e com o objetivo do processo de Mineração de Textos.

  $$Class\_Keywords\_Set = \{\{ck_{11}, ck_{12}, \ldots, ck_{1j}\}, \ldots, \{ck_{m1}, ck_{m2}, \ldots, ck_{ml}\}\}$$

  Cada elemento da m-ésima lista do conjunto *Class_Keywords_Set*, ck_{mj}, é formado por um termo identificador de classe t e seus sinônimos s, isto é, $ck_{ml} = \{t_l\} \cup \{s_1, \ldots, s_p\}$).

Como passo seguinte para a construção da representação gBoED está a maneira como ela é construída e como os atributos são formados. Os atributos são representados pelas colunas da matriz documento-termo e formados por expressões do domínio criadas a partir da combinação dos elementos da lista *Domain_Keywords* com os elementos das listas *Class_Keyword_Set*. Uma métrica é associada às expressão do domínio presentes em cada documento. Nesta primeira versão da representação, denominada gBoED_Freq a métrica corresponde à frequência em que os dois termos coocorrem em uma mesma sentença do documento. Um esquema da representação gBoED para uma coleção de n documentos é apresentado na Figura 18. Vale ressaltar que a gBoED também pode ser vista como uma BoED independente de domínio. Sendo assim, para o

problema do domínio de Sistemas Produto-Serviço tratado por Marques *et al.* (2015), as duas representações são equivalentes.

Figura 18 – Esquema da representação de coleção de documentos gBoED.

	$k_1\_ck_{11}$	...	$k_1\_ck_{1j}$	...	$k_i\_ck_{11}$	...	$k_i\_ck_{1j}$	...	$k_1\_ck_{m1}$	...	$k_1\_ck_{ml}$	...	$k_i\_ck_{m1}$	...	$k_i\_ck_{ml}$
d_1															
d_2															
⋮															
d_n															

Fonte: Elaborada pelo autor.

Experimentos que visam verificar a utilização da representação *gBoed* em duas diferentes versões, comparando-as com resultados de modelos gerados pelo representação BoW em tarefas de classificação de segundo nível semântico, foram realizados e alguns dos resultados que serão apresentados a seguir foram publicados nos eventos DocEng 2019 - ACM Symposium on Document Engineering 2019 (SCHEICHER *et al.*, 2019) e ERAMIA 2020 - I Escola Regional de Aprendizado de Máquina e Inteligência Artificial de São Paulo (SCHEICHER; SINORARA; REZENDE, 2020).

3.3.1 Avaliação experimental de viabilidade da representação gBoED

Os experimentos são planejados e executados com o objetivo de verificar o impacto da representação gBoED em diferentes cenários de classificação de documentos. Assim, nessa avaliação experimental foram executadas as três etapas centrais do processo de Mineração de Textos.

Na etapa de Pré-processamento são geradas as representações BoW e gBoED_Freq para as coleções de documentos. Para a construção das representações BoW aplica-se a remoção de *stopwords* e radicalização dos termos. Para gerar a gBoED_Freq são utilizadas as listas de termos do domínio e de identificadores de classes. As expressões do domínio foram geradas com os termos das listas radicalizados. As coleções de documentos e as listas de termos são apresentadas na Subsubseção 3.3.1.1.

Na etapa de Extração de Padrões são construídos classificadores utilizando-se sete algoritmos diferentes de Aprendizado de Máquina, apresentados na Subsubseção 3.3.1.2, incluindo dois algoritmos de classificação indutiva baseados em redes bipartidas (ROSSI *et al.*, 2014; ROSSI; LOPES; REZENDE, 2016).

Os classificadores gerados foram avaliados considerando os melhores valores de acurácia (porcentagem de documentos corretamente classificados) obtidas por um processo de validação cruzada (*10-fold cross validation*). Os resultados obtidos são apresentados na Subsubseção 3.3.1.3.

A seguir, são apresentadas as coleções de documentos utilizadas, juntamente com as listas de termos consideradas na geração da gBoED.

3.3.1.1 Coleções de documentos

Com o objetivo de avaliar o uso da representação gBoED_Freq em tarefas de classificação de nível semântico, alguns experimentos foram realizados utilizando bases de dados de *benchmarking*.

Com relação às bases de dados a serem utilizadas nesse processo, os experimentos foram conduzidos inicialmente utilizando-se duas coleções de textos: *Best Sports* (SINOARA; REZENDE, 2018) e *SemEval-2015* (PONTIKI *et al.*, 2015). Para cada uma das coleções, foram geradas três configurações diferentes de *datasets* que reproduzem distintos níveis de complexidade semântica em cenários de classificação (SINOARA; REZENDE, 2018).

A primeira coleção de textos, denominada *BEST Sports - Top 4* (BS-Top4), é um conjunto de 283 notícias de esportes escritas em língua portuguesa extraídas do *website BEST sports*[1] e preparados para execução em diferentes níveis semânticos (SINOARA; REZENDE, 2018). Cada documento possui a classificação correspondente a um esporte, podendo ser Fórmula 1, Motovelocidade, Futebol ou Tênis. A partir dessa coleção de documentos, foram geradas três configurações com finalidades diferentes para realização dos experimentos:

1. ***BS-tópico*:** a primeira configuração, denominada *BS-tópico*, corresponde à classificação padrão por esporte: Fórmula 1, Motovelocidade, Futebol e Tênis. Segundo Sinoara (2018) esta configuração é considerada do primeiro nível de complexidade semântica, correspondendo a uma classificação por tópico.

2. ***BS-semântico*:** a segunda configuração, denominada *BS-semântico*, está relacionada ao desempenho dos atletas brasileiros. Nesse caso, existem quatro possíveis rótulos de classificação: "Brasileiro venceu", "Brasileiro não venceu", "Não foi citado brasileiro" ou "Não definido". Com base nessas classes, espera-se que o classificador consiga assimilar do conjunto de textos certo nível semântico de modo a obter desempenho satisfatório. Essa característica é necessária na identificação de atletas brasileiros e, também, de vitórias ou derrotas

3. ***BS-tópico-semântico*:** a terceira configuração, denominada *BS-tópico-semântico* une as configurações de classe das duas versões anteriores. Essa versão relaciona o tipo de esporte praticado com o desempenho dos atletas brasileiros, e.g., considera-se a união das classes apresentadas nas configurações *BS-tópico* com *BS-semântico*.

A segunda coleção de textos é denominada *SemEval-2015*. Escrita em língua inglesa, ela é composta por 801 opiniões de satisfação de usuários sobre hotéis, restaurantes e *laptops*. Essa

[1] *BEST Sports* - Arquivo de notícias: <http://bestsports.com.br/db/notarqhome.php>

coleção foi disponibilizada para a *SemEval-2015 Aspect Based Sentiment Analysis Task* (PONTIKI *et al.*, 2015). Todos os textos são anotados com a polaridade (positivo, negativo ou neutro) de cada aspecto dos produtos avaliados pelo autor.

Assim como para a coleção *BEST sports*, foram construídas para a *SemEval-2015* três configurações distintas que foram utilizadas nos experimentos:

1. ***SE-produto*:** a primeira configuração, denominada *SE-produto*, corresponde à classificação padrão por produto: hotéis, restaurantes e *laptops*. Semelhante à configuração *BS-topic*, de acordo com Sinoara (2018), esta configuração é considerada do primeiro nível de complexidade semântica, correspondendo a uma classificação por tópico.

2. ***SE-polaridade*:** a segunda configuração, denominada *SE-polaridade*, está relacionada à polaridade das opiniões. Originalmente a coleção possui três possíveis rótulos de classificação: Positivo, Negativo e Neutro. Porém, nos experimentos executados nesta Seção, considerou-se apenas as classes Positivo e Negativo. Ao considerar essas classes espera-se que o classificador considere certo nível semântico para obter desempenho satisfatório, pois essa característica é necessária para identificação da polaridade da opinião.

3. ***SE-produto-polaridade*:** a terceira configuração, denominada *SE-produto-polaridade* une as configurações de classe das duas versões anteriores. Essa versão relaciona o tipo de produto avaliado com a polaridade da opinião, e.g., considera-se a união das classes apresentadas nas configurações *SE-produto* com *SE-polaridade*.

Para a construção das representações gBoED da coleção *BS-Top4* foram utilizadas listas que contêm nomes de atletas brasileiros (lista de termos do domínio) e verbos usados para expressar vitórias e derrotas (listas de identificadores de classe). Para a coleção *SemEval-2015* foram usadas listas que contêm aspectos dos produtos (lista de termos do domínio) e palavras positivas e negativas (listas de identificadores de classe). Em seguida, são apresentadas a configuração dos experimentos.

3.3.1.2 Configuração dos experimentos de viabilidade da representação gBoED

Para avaliar a proposta, executou-se três conjuntos de experimentos de classificação: (i) classificação indutiva supervisionada utilizando a representação BoW, que foi utilizada como *baseline*; (ii) classificação indutiva supervisionada utilizando a representação gBoED_Freq; e (iii) combinação de classificadores gerados utilizando ambas as representações BoW e gBoED_Freq. Os algoritmos utilizados são apresentados a seguir, seguidos dos parâmetros de configuração utilizados em cada caso.

- **J48**, algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor*.

- ***K-nearest neighbor* (KNN)**, algoritmo IBk. Foram utilizadas as opções de voto com peso e voto sem peso. Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: distância euclideana e cosseno.
- ***Naive Bayes* (NB)**, algoritmo baseado em problabilidade condicional;
- ***Multinomial Naive Bayes* (MNB)**, algoritmo baseado em problabilidade condicional aplicado a valores discretos;
- ***Suport Vector Machine* (SVM)**, algoritmo *Sequential Minimal Optimization* (SMO). Nesse algoritmo foram considerados três tipos de kernel: linear, polinomial (expoente=2) e RBF (Radial Basis Function). Os valores considerados para cada tipo de kernel foram 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10, 10^2, 10^3, 10^4, 10^5.
- **IMBHN$^{\mathbf{C}}$** (ROSSI *et al.*, 2014) e **IMBHN$^{\mathbf{R}}$** (ROSSI; LOPES; REZENDE, 2016), algoritmos de classificação indutiva baseados em redes heterogêneas bipartidas. Nesses algoritmos utilizou-se taxa de correção de erros de 0,01, 0,05, 0,1, 0,5. O número máximo de iterações foi ajustado para 1000 e utilizou-se o erro dos mínimos quadrados com critério de parada de 0,01.

Para realizar a combinação de classificadores gerados com cada representação (BoW e gBoED) foram utilizados três abordagens: (i) uso da resposta do classificador com maior confiança (*Most Confident* (MC)); (ii) uso da resposta com maior soma de confianças (*Sum of Confidences* (SC)); e (iii) uso da resposta com maior soma de confianças ponderadas pela acurácia dos classificadores no conjunto de treinamento (*Weighted Sum of Confidences* (WSC)). Além disso, para as três abordagens também foram utilizados diferentes pesos para as representações. Considere w_1 o peso do classificador gerado com a representação BoW e w_2 o peso do classificador gerado com a gBoED. Nesses experimentos, foram utilizados os seguintes valores para esses pesos: $w_1 = \{0,1; 0,3; 0,5; 0,7; 0,9\}$ e $w_2 = 1 - w_1$.

Em seguida, são apresentados os resultados dos experimentos realizados, bem como as discussões que os envolvem.

3.3.1.3 Resultados - viabilidade da representação gBoED

A Tabela 3 apresenta os resultados obtidos neste experimento. Inicialmente, verifica-se que a primeira coluna da tabela corresponde aos algoritmos de classificação aplicados nos documentos que utilizaram métodos de representação de documentos. Na segunda coluna, observa-se os resultados de acurácia obtidos por cada algoritmo utilizando a representação BoW, que serve de *Baseline* para comparação dos resultados. Como citado anteriormente o método BoW é tradicionalmente conhecido por gerar resultados significativos na representação de documentos, quando submetido a tarefas de classificação. Na terceira coluna da tabela,

estão presentes os resultados das acurácias nas tarefas de classificação, após ser submetido à representação de documentos gBoED_Freq. As colunas seguintes da tabela são correspondentes às acurácias quando aplicadas a combinação dos modelos de representação (BoW e gBoED_-Freq). Acurácias maiores que as obtidas com o *baseline* BoW, quando não é o resultado obtido com a própria representação BoW, são apresentadas em negrito. A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia de cada representação para o respectivo *dataset*.

Ao observar as acurácias obtidas pelos classificadores em cada uma das representações usando os diferentes *datasets*, verifica-se que a representação gBoED_Freq quando utilizada individualmente não obtém bons resultados em nenhum dos cenários apresentados. A representação *Bag of Words* obtém melhores valores de acurácias apenas nos cenários cuja classificação está diretamente relacionada ao assuntos dos documentos. Outra observação importante é que apesar da gBoED_Freq ser formada por atributos mais expressivos do que simples palavras ela não apresenta informação suficiente para possibilitar a classificação dos documentos. No caso da BoW, mesmo obtendo resultados mais significativos para os *datasets BS-tópico* e *SE-produto*, a ausência de informações semânticas associadas à representação não garante o mesmo nível de resultados em todas as situações que possuem determinados níveis de complexidade semântica, como em *BS-semântico* e *SE-polaridade*. Nos resultados obtidos com a combinação de representações, pode-se verificar principalmente nos *datasets* das versões semânticas, resultados aproximados ou um pouco melhores do que aqueles obtidos com a representação BoW.

A partir deste trabalho, pode-se verificar que nos cenários de baixa complexidade semântica, aqueles que envolvem as coleções *BS-tópico* e *SE-produto*, a representação BoW aparenta ser suficiente para a construção de classificadores cuja acurácia é próxima de 100%. Para as coleções *BS-semântico*, *BS-tópico-semântico*, *SE-polaridade* e *SE-produto-polaridade*, a BoW não é suficiente. Nesses cenários, verifica-se que a combinação das representações BoW e gBoED_Freq atingiu melhores valores de acurácia na maioria das configurações testadas. Vale notar a melhora de acurácia obtida com a combinação de classificadores no caso *SE-polaridade* utilizando o algoritmo KNN. Para esse caso, o melhor modelo gerado utilizando a BoW obteve 77,4179% de acurácia, enquanto a combinação de classificadores gerados com as duas representações obteve 93,0051%. Este trabalho foi publicado no XIII Encontro Nacional de Inteligência Artificial e Computacional 2016 (ENIAC 2016) (SCHEICHER *et al.*, 2016).

A partir dos resultados apresentados nesse primeiro conjunto de experimentos é possível validar a viabilidade no uso de representações semanticamente enriquecidas por meio de expressões do domínio em processos de classificação de textos de nível semântico. Portanto, um dos caminhos de pesquisa realizados nesse trabalho de doutorado diz respeito à construção de outras versões de gBoED que utilizam diferentes métricas e formas de construção, bem como construção de um método de classificação que combine abordagens tradicionais com informações enriquecidas semanticamente. Essa abordagem de pesquisa permite avançar o estado da arte na

Tabela 3 – Melhores acurácias para as coleções *BS-Top4* e *SemEval-2015*

			BOW + *gBoED_Freq*		
	BOW	***gBoED_Freq***	**SC**	**WSC**	**MC**
BS-tópico	100,0000	78,4360	100,0000	100,0000	100,0000
IMBHN^{C}	98,9286	67,2167	98,9286	98,9286	98,9286
IMBHN^{R}	99,6429	68,9532	99,6429	99,6429	99,6429
J48	96,8227	61,8719	96,8227	96,8227	96,8227
k-NN	99,6552	73,1158	99,6552	99,6552	99,6552
MNB	100,0000	78,4360	100,0000	100,0000	100,0000
NB	99,6429	71,4163	99,6429	99,6429	99,6429
SVM	100,0000	71,7118	100,0000	100,0000	100,0000
BS-semântico	68,9532	53,7315	**69,2857**	**69,2980**	68,9532
IMBHN^{C}	64,6552	44,2118	**65,3818**	**65,3571**	64,6552
IMBHN^{R}	68,9532	43,1034	**69,2857**	**69,2980**	68,9532
J48	59,0517	46,3054	**61,4655**	**63,2266**	**61,5025**
k-NN	65,3818	53,7315	**68,5345**	**68,1650**	**66,4778**
MNB	59,7414	52,9926	**63,9778**	**63,6330**	**60,4557**
NB	57,6108	46,6379	**59,0271**	**57,9557**	**59,0025**
SVM	63,6576	48,8300	**64,7167**	**64,7167**	63,6576
BS-tópico-semântico	66,8596	44,5074	**68,9901**	**68,9901**	66,8596
IMBHN^{C}	62,5739	42,0443	**67,0813**	**68,1281**	**62,9310**
IMBHN^{R}	57,9926	38,5099	56,5764	56,5764	57,9926
J48	55,1478	31,8596	**56,2069**	**56,5640**	**57,2537**
k-NN	65,7512	44,5074	**66,7734**	**67,5000**	65,7512
MNB	62,5985	44,1749	**65,0369**	**63,9655**	**65,0369**
NB	57,2537	43,8054	**57,6108**	**57,6108**	57,2537
SVM	66,8596	38,5345	**68,9901**	**68,9901**	66,8596
SE-produto	99,5077	89,9353	99,5077	99,5077	99,5077
IMBHN^{C}	98,1587	86,0193	**98,5276**	**98,5276**	98,1587
IMBHN^{R}	99,1388	89,6914	99,0169	99,0169	99,1388
J48	92,2704	80,6158	92,2704	92,2704	92,2704
k-NN	98,5245	89,9353	**98,7699**	**98,7699**	**98,6480**
MNB	99,5077	86,2572	99,5077	99,5077	99,5077
NB	92,7612	84,5348	**93,9898**	**94,4791**	**94,2307**
SVM	96,4408	89,4550	96,3189	96,3189	96,4408
SE-polaridade	84,5438	69,0786	**84,5453**	**84,5453**	84,5438
IMBHN^{C}	80,4908	65,0316	**80,6143**	**80,6128**	80,4908
IMBHN^{R}	82,8214	64,2939	81,9587	82,2057	82,8214
J48	71,5266	66,5056	**72,6257**	**72,3893**	**72,7492**
k-NN	77,4179	68,5953	**91,0388**	**92,8696**	**93,0051**
MNB	84,5438	66,7525	**84,5453**	**84,5453**	84,5438
NB	70,3071	63,6886	70,1867	**70,3086**	70,3071
SVM	81,6110	69,0786	81,2436	81,3640	81,6110
SE-produto-polaridade	83,6811	62,2162	**83,8046**	**83,8046**	83,6811
IMBHN^{C}	77,4345	60,1235	**77,7989**	**78,0367**	**77,5565**
IMBHN^{R}	73,9762	60,7347	72,5083	72,6302	73,9762
J48	71,0479	56,1999	**71,6652**	**71,4213**	**71,6667**
k-NN	75,8220	61,2195	**75,9455**	**75,9455**	**75,9455**
MNB	83,6811	62,2162	**83,8046**	**83,8046**	83,6811
NB	68,9521	57,0671	**69,5604**	**69,5649**	**69,0696**
SVM	77,8034	60,8627	**78,0473**	**78,0473**	77,8034

Nota – *Most Confident* (MC): resposta do classificador com maior confiança; *Sum of Confidences* (SC): resposta com maior soma de confianças; *Weighted Sum of Confidences* (WSC): resposta com maior soma de confianças ponderadas pela acurácia dos classificadores no conjunto de treinamento.

Fonte: Scheicher *et al.* (2016).

literatura relacionada à semântica de textos.

Outra importante conclusão que pode ser feita a partir dos experimentos realizados

nesse trabalho e que envolvem as tarefas desta proposta de doutorado diz respeito ao fato da representação gBoED_Freq abordar a semântica dos textos por meio de conhecimento do domínio. Tal conhecimento é obtido por meio das listas de termos do domínio e identificadores de classe que podem ser incorporadas às representações dos textos enriquecendo seu conteúdo e consequentemente trazendo um maior nível de informações para os modelos de classificação. Portanto, outro caminho de pesquisa que busca avançar o estado da arte na literatura relacionada à semântica de textos está relacionado à geração de listas de termos e identificadores de classe de forma semiautomática, com o objetivo de obter listas tão interessantes ou próximas àquelas formadas pelos especialistas do domínio. Essa abordagem permite maior automatização dos métodos de classificação e da geração das representações semanticamente enriquecidas por expressões do domínio.

3.3.2 gBoED_Dist: gBoED com métrica de distância

Na Seção 3.3 é apresentada a generalização da representação semanticamente enriquecida baseada em expressões do domínio com métrica de frequência (gBoED_Freq), bem como o experimento realizado para validar a representação e suas limitações. A partir dos experimentos realizados é possível perceber que, em cenários de baixa complexidade semântica a representação BoW é suficiente para a construção de classificadores com bons resultados. Em cenários de mais alta complexidade semântica a BoW não é suficiente e apresenta a necessidade de combinação com representações que agreguem informações enriquecidas, como é o caso da gBoED.

Baseando-se nas observações extraídas a partir do experimento apresentado na Subseção 3.4.2, foi desenvolvida uma nova versão da gBoED cuja formação das expressões e a métrica associada baseia-se na distância entre um Termo do Domínio e um Identificador de Classe. Assim como na gBoED_Freq, as expressões do domínio são formadas pela combinação de todos Termos do Domínio com todos os Identificadores de Classe existentes em cada sentença. A métrica associada corresponde ao inverso da distância em número de palavras entre um Termo do Domínio e um Identificador de Classe, dada pela equação:

$$dist(Td, Ic) = \frac{1}{n(Td, Ic)}, \tag{3.1}$$

no qual, *Td* corresponde ao Termo do Domínio e *Ic* ao Identificador de Classe. *n(Td,Ic)* corresponde a quantidade de palavras entre o Termo do Domínio e o Identificador de Classe. Assim, termos que possuem maior distância entre si em uma sentença tendem a estar menos relacionados do que termos que estão mais próximos. Como nesta métrica é considerado o inverso, os termos mais distantes recebem um peso menor em relação aos termos mais próximos. O exemplo do Quadro 4 mostra como a representação é construída a partir de um documento de opinião de produto. Nele é possível observar que a opinião forma tanto expressões positivas quanto negativas. Sua classificação original é negativa. O símbolo <*TD*> indicam Termos do Domínio,

<*TP*> Termo Positivo e <*TN*> Termo Negativo. Na representação gBoED_Dist, os números 0 e 1 contidos entre os termos das expressões são referentes às classes positiva (0) e negativa (1).

Quadro 4 – gBoED baseada em distância de termos.

Documento original:

Produto entregue no prazo. O produto veio com defeito. O som distorce quando aumenta todo o volume, em seguida, a caixa desconecta do modo Bluetooth.

Marcação de termos:

Sentença 1: <Produto><*TD*> <entregue no prazo><*TP*>.
Sentença 2: O <produto><*TD*> veio com <defeito><*TN*>.
Sentença 3: O <som><*TD*> <distorce><*TN*> quando aumenta todo o volume, em seguida, a <caixa><*TD*> <desconecta><*TN*> do modo Bluetooth.

---

gBoED_Dist:

produto_0_prazo	produto_1_defeito	som_1_distorce	som_1_desconecta	caixa_1_distorce	caixa_1_desconecta
1	0,33	1	0,09	0,11	1

Fonte: Elaborada pelo autor.

O experimento da Seção 3.3 também mostra que a combinação da representação tradicional BoW com a representação gBoED resultou melhores acurácias em cenários de maior complexidade semântica. Na Seção 3.4 é proposto um Método de Classificação Semanticamente Enriquecida por Expressões do Domínio que combina resultados de classificadores gerados pelo modelo tradicional BoW com as predições feitas a partir das representações semanticamente enriquecidas gBoED_Freq e gBoED_Dist.

3.4 Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

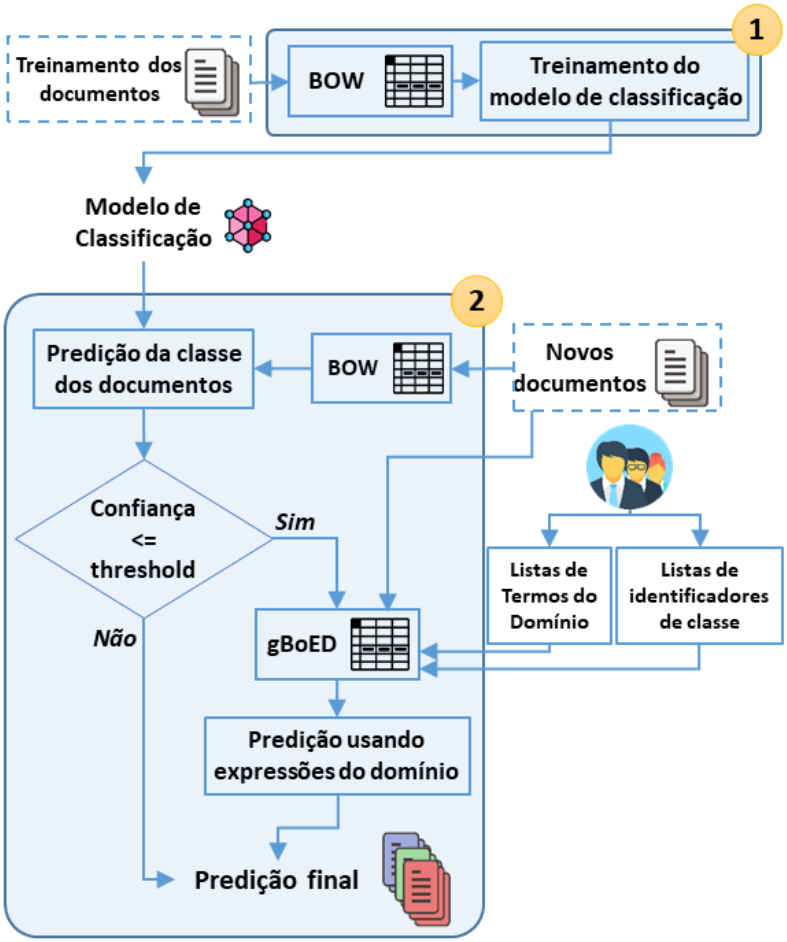
Um dos objetivos desta pesquisa de doutorado é a melhoria de resultados de classificação, em cenários de mais alta complexidade semântica, utilizando informações semanticamente enriquecidas. Com base no objetivo apresentado, nas diferentes versões de representações baseadas em expressões do domínio construídas e nas percepções extraídas a partir dos experimentos descritos na Seção 3.3, verificou-se que a combinação entre modelos gerados por métodos

tradicionais de representação (BoW) e informações enriquecidas semanticamente podem trazer melhores resultados nesses cenários.

3.4.1 Método proposto para classificação enriquecida por expressões do domínio

Visando atingir tais objetivos e considerando a combinação de diferentes modelos, foi desenvolvido um Método de Classificação Semanticamente Enriquecida por Expressões do Domínio que visa melhorar resultados de classificação de nível semântico utilizando conhecimentos oriundos do especialista do domínio. A Figura 19 apresenta um diagrama do método proposto que consiste em duas etapas principais bem definidas.

Figura 19 – Diagrama do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio.



Fonte: Elaborada pelo autor.

A etapa 1 corresponde ao treinamento de um modelo de classificação baseado em uma representação da coleção de textos do tipo *Bag of Words* (BoW). Nesta etapa, qualquer algoritmo de classificação pode ser aplicado para obter um modelo de classificação. A BoW é usada, pois

esse modelo tradicional alcança bons resultados em diversos cenários simples de classificação de texto.

Na etapa 2, as informações semanticamente enriquecidas da gBoED são aplicadas para melhorar os resultados da classificação para aqueles documentos cuja confiança na previsão é menor ou igual a um limite global definido. Primeiro, os novos documentos são preparados para classificação, ou seja, uma representação BoW é construída. Em seguida, a representação BoW dos novos documentos é apresentada ao modelo de classificação treinado na etapa 1 para predizer a classe dos documentos. A gBoED construída a partir das listas de termos do domínio e identificadores de classes definidas pelo especialista do domínio.

A saída do modelo é uma classe prevista de cada documento pela gBoED, a partir de um valor de confiança. Se a confiança for maior que um limite definido, a classe prevista é considerada como a previsão final do documento ou, caso contrário, a melhoria da classificação é realizada. A previsão da classe em cada uma das representações (gBoED_Freq e gBoED_Freq) é dada pela soma dos pesos associados às expressões do domínio que representam cada classe, existentes em cada representação. Na representação gBoED_Freq a predição é dada pela soma das frequências das expressões do domínio de cada classe, assim, a classe que contém a maior quantidade de expressões do domínio é escolhida para determinado documento. Na representação gBoED_Freq a predição também é dada pela soma de pesos das expressões do domínio de que representa cada classe. É importante lembrar, como apresentado na Subseção 3.3.2, que os pesos associados a uma expressão do domínio são atribuídos como o inverso da distância entre um termo do domínio e um identificador de classe.

3.4.2 Avaliação experimental do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

O método desenvolvido tem por objetivo melhorar resultados de classificação em cenários de maior complexidade semântica, combinando resultados de um modelo gerado a partir de uma representação tradicional BoW com dados oriundos de uma representação semanticamente enriquecida. Nesta subseção é apresentada a avaliação experimental realizada para validar o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio.

3.4.2.1 Coleção de documentos

As coleções de documentos que serão apresentadas nessa seção, serão utilizadas nessa e em outras avaliações experimentais realizadas ao longo deste trabalho de doutorado. Como um dos objetivos do método proposto é ser aplicável em diferentes idiomas e domínios, serão utilizadas coleções de documentos que sigam essa premissa. Para cada coleção de documentos também foram geradas as listas de termos do domínio e identificadores de classe. A seguir são apresentadas as coleções de dados utilizadas, bem como as configurações das listas de termos do

domínio e identificadores de classe.

- **B2W Reviews 2019 Info:** coleção de documentos composta por 132.374 opiniões de produtos em português separadas em dezenas de categorias (REAL; OSHIRO; MAFRA, 2019). A *B2W Digital* é uma das maiores plataformas de e-commerce da América Latina. Para os experimentos deste trabalho foi selecionada uma categoria relacionada a um único domínio: Informática. Essa categoria é formada por 4262 opiniões de produtos como computadores, notebooks e tablets. A base original possui rotulação por pontos que vão de 1 (ruim) a 5 (excelente). Neste trabalho, por questões de capacidade de processamento, foram selecionadas aleatoriamente 1000 opiniões, 500 opiniões para cada classe. Os rótulos foram modificados para "Positivo" as opiniões pertencentes à pontuação 5 e "Negativo" as opiniões pertencentes às pontuações 1 e 2, de acordo com instruções dos autores.

  As listas de termos domínio e identificadores de classe foram montadas de forma manual. Os gráficos da Figura 20 apresentam uma relação entre a quantidade de termos únicos e com sinônimos para as listas formadas a partir de 10% e 100% da coleção de documentos. No gráfico da Figura 20a pode-se observar as quantidades de termos do domínio, na Figura 20b observa-se as quantidades para identificadores da classe "*Positivo*" e na Figura 20c observa-se as quantidades para identificadores da classe "*Negativo*".

  A Tabela 4 apresenta exemplos de termos do domínio e identificadores de classe para as listas da coleção de documentos *B2W Reviews 2019 Info*.

Tabela 4 – Exemplos de termos - Coleção *B2W Reviews 2019 Info*.

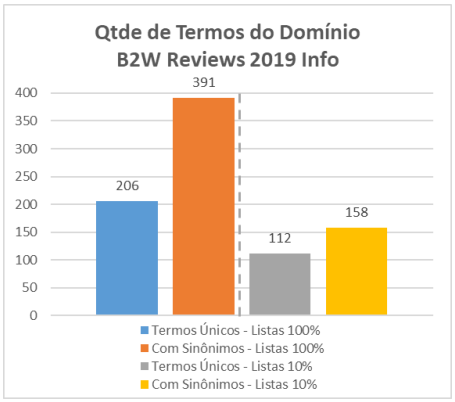
Termos de Domínio	**Identificadores da Classe *Positivo***	**Identificadores da Classe *Negativo***
relação custo benefício; custo beneficio; custo-beneficio; custo; custos	perfeitamente; perfeitas; perfeito; prefeito; perfeita	insatisfatório; insatisfeito; insatisfação
placa do roteador; roteadores; roteador	rapidíssimo; rapidinha; rapidês; rápido	baixa qualidade; baixa
funcionamento; funcionando; funcionou	bem; bom; boa	decepcionante; decepcionado; decepcionada; decepcionei; decepciona; decepção

Fonte: Elaborada pelo autor.

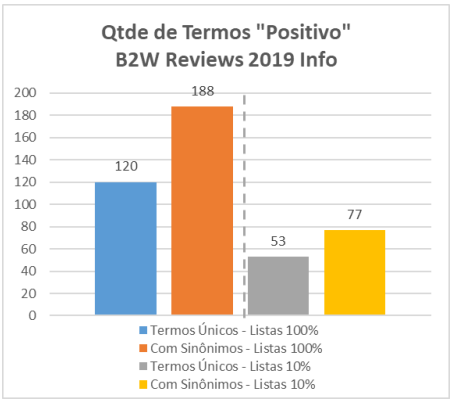
- **BestSports Top4:** é um conjunto de notícias de esportes escritas em língua portuguesa extraídas do *website BEST Sports*[2] e preparados para execução em diferentes níveis semânticos (SINOARA; REZENDE, 2018). A configuração utilizada nessa avaliação experimental é referente ao segundo nível de complexidade semântica, com um total de 181 notícias, sendo 93 notícias da classe "Brasileiro venceu" e 88 da classe "Brasileiro

[2] *BEST sports* - Arquivo de notícias: <http://bestsports.com.br/db/notarqhome.php>

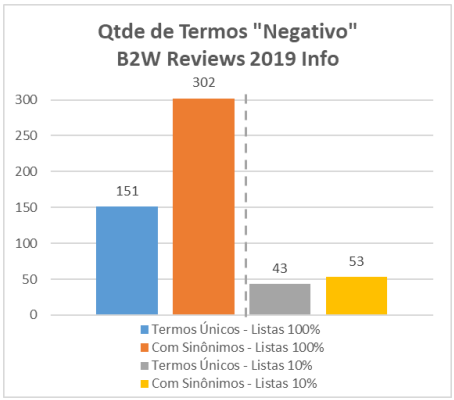
Figura 20 – Gráficos de quantidade de termos por tipo de lista - Coleção *B2W Reviews 2019 Info*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positivo*"



(c) Identificadores de Classe "*Negativo*"

Fonte: Elaborada pelo autor.

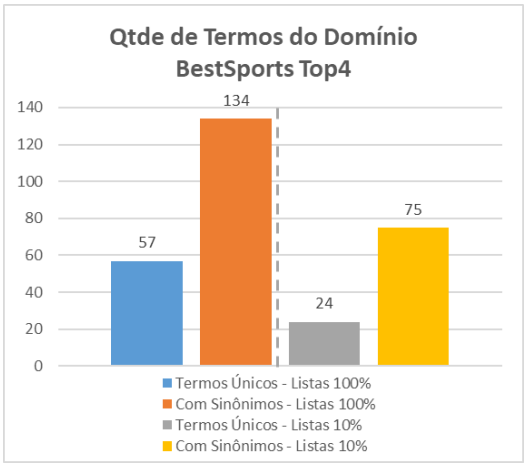
não venceu". A lista de termos do domínio é formada pelos nomes dos atletas brasileiros presentes no conjunto de notícias e as listas de identificadores de classe são formadas por palavras que indicam vitória ou derrota e seus respectivos sinônimos.

Os gráficos da Figura 21 apresentam uma relação entre a quantidade de termos únicos e com sinônimos para as listas formadas a partir de 10% e 100% da coleção de documentos. No gráfico da Figura 21a pode-se observar as quantidades de termos do domínio, na Figura 21b observa-se as quantidades para identificadores da classe "*Brasileiro Venceu*" e na Figura 21c observa-se as quantidades para identificadores da classe "*Brasileiro Não Venceu*".

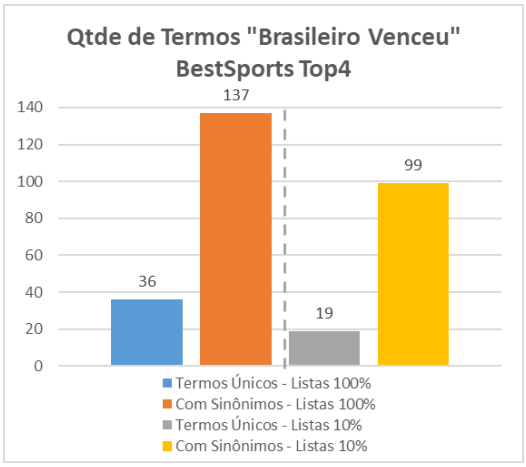
A Tabela 5 apresenta exemplos de termos do domínio e identificadores de classe para as listas da coleção de documentos *BestSports Top4*.

- **HuLiu 2004:** coleção composta por opiniões, em inglês, de 5 produtos (duas câmeras fotográficas digitais, um telefone celular, um *MP3 Player* e um *DVD Player*) (HU; LIU, 2004). A coleção original possui uma lista de aspectos separadas por polaridade (*positive*, *negative* e *neutral*) associadas a cada opinião. A rotulação das opiniões foi realizada

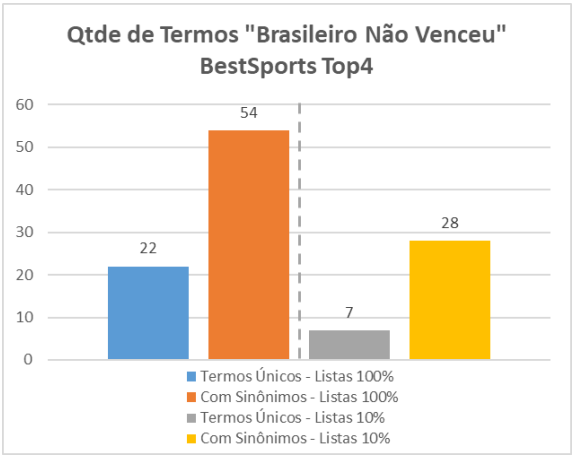
Figura 21 – Gráficos de quantidade de termos por tipo de lista - Coleção *BestSports Top4*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Brasileiro Venceu*"



(c) Identificadores de Classe "*Brasileiro Não Venceu*"

Fonte: Elaborada pelo autor.

Tabela 5 – Exemplos de termos - Coleção *BestSports Top4*.

Termos de Domínio	**Identificadores da Classe *Brasileiro Venceu***	**Identificadores da Classe *Brasileiro Não Venceu***
wanderley luxemburgo; luxemburgo; wanderley	ganhando; ganhar; ganhou	desclassificadas; desclassificado; desclassificados
emerson fittipaldi; emerson; fittipaldi	dispara; disparando; disparava; disparavam; disparou	abandona; abandonado; abandonando; abandonar; abandonaram; abandonou
brasileiro; brasileira; brasileiros; brasileiras; brasil	arrasa; arrasar; arrasou	perde; perdendo; perder; perderão; perdeu; perdia; perdido

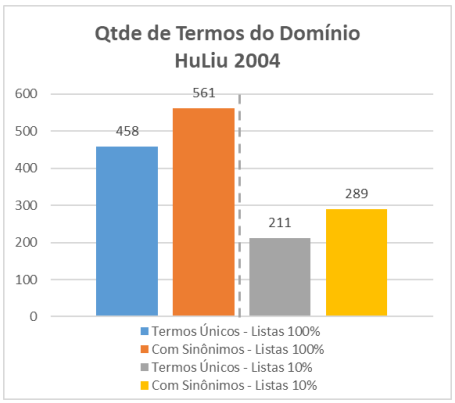
Fonte: Elaborada pelo autor.

considerando a polaridade mais frequente entre os aspectos. As listas de termos também

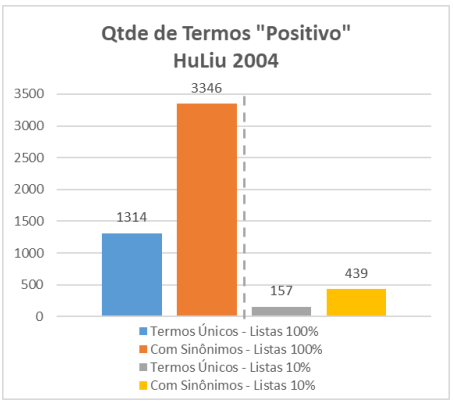
foram geradas a partir dos aspectos. Nesse cenário, o conjunto possui 186 avaliações positivas e 110 negativas, totalizando 296 avaliações. Não foram consideradas as 18 avaliações neutras restantes.

Os gráficos da Figura 22 apresentam uma relação entre a quantidade de termos únicos e com sinônimos para as listas formadas a partir de 10% e 100% da coleção de documentos. No gráfico da Figura 22a pode-se observar as quantidades de termos do domínio, na Figura 22b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 22c observa-se as quantidades para identificadores da classe "*Negative*".

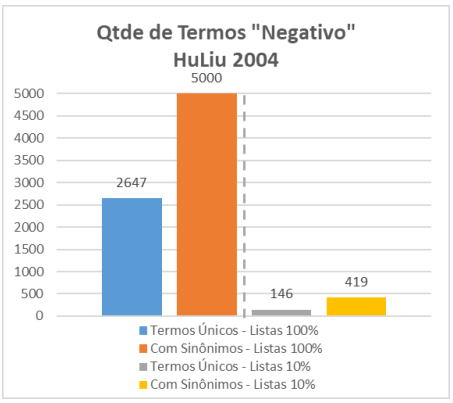
Figura 22 – Gráficos de quantidade de termos por tipo de lista - Coleção *HuLiu 2004*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

A Tabela 6 apresenta exemplos de termos do domínio e identificadores de classe para as listas da coleção de documentos *HuLiu 2004*.

- **SemEval 2014:** criada para o desafio de mineração de textos *SemEval-2014 Aspect Based Sentiment Analysis task 4* (PONTIKI *et al.*, 2014), é uma coleção para análise de sentimentos em inglês, composta por opiniões de *laptops* e restaurantes.

  A coleção original está anotada de acordo com a polaridade de aspectos, como "*positive*", "*negative*" e "*neutral*". A polaridade dos documentos foi determinada da mesma forma

Tabela 6 – Exemplos de termos - Coleção *HuLiu 2004*.

Termos de Domínio	**Identificadores da Classe *Positive***	**Identificadores da Classe *Negative***
rechargable battery; rechargable batteries	large-capacity; large capacity	incompatability; incompatibility; incompatible
mobile phone; mobile phones; phone; phones	sustainability; sustainability; sustainable;sustainable	scare; scared; scarier; scariest; scarily; scarred; scars; scary
digital camera; digital cameras	congratulate; congratulate; congratulation; congratulation	worry; worrying; worryingly; worried; worriedly; worrier; worries; worrisome

Fonte: Elaborada pelo autor.

que em *HuLiu 2004*, compondo um total de 1.836 opiniões positivas e 1.073 opiniões negativas, ao todo 2.909 opiniões. Sub-conjuntos das listas de termos foram criados para melhor identificação das expressões do domínio em cada sub-coleção de documentos. São eles:

- **SemEval 2014 Laptop:** composta por 619 opiniões positivas e 622 negativas.
- **SemEval 2014 Restaurant:** composta por 1.217 opiniões positivas e 451 negativas.

Os gráficos da Figura 23 apresentam uma relação entre a quantidade de termos únicos e com sinônimos para as listas formadas a partir de 10% e 100% da coleção de documentos. No gráfico da Figura 23a pode-se observar as quantidades de termos do domínio, na Figura 23b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 23c observa-se as quantidades para identificadores da classe "*Negative*".

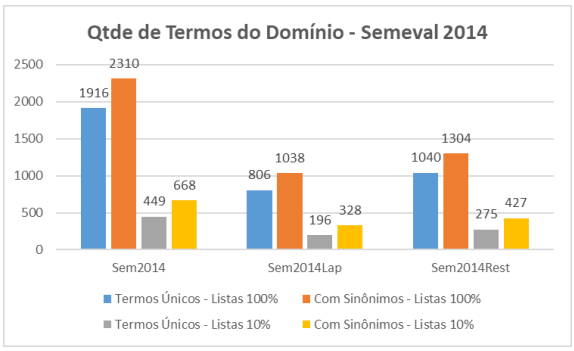
A Tabela 7 apresenta exemplos de termos do domínio e identificadores de classe para as listas da coleção de documentos *SemEval 2014*.

- **SemEval 2015:** coleção criada para o desafio de mineração de textos *SemEval-2015 Aspect Based Sentiment Analysis task 12* (PONTIKI *et al.*, 2015), também é uma coleção para análise de sentimentos em inglês, composta por opiniões de *laptops*, hotéis e restaurantes. Assim como em *HuLiu 2004* e *SemEval 2014*, a polaridade dos documentos foi determinada pela polaridade da maior quantidade de aspectos, resultando em 555 opiniões positivas, 246 opiniões negativas, em um total de 801 opiniões.

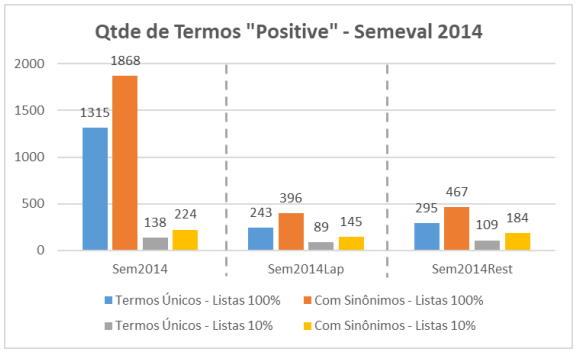
  Nessa coleção também foram consideradas as sub-coleções:

  - **SemEval 2015 Laptop:** composta por 277 opiniões positivas e 151 negativas.
  - **SemEval 2015 Hotel:** composta por 21 opiniões positivas e 8 negativas.
  - **SemEval 2015 Restaurant:** composta por 257 opiniões positivas e 87 negativas.

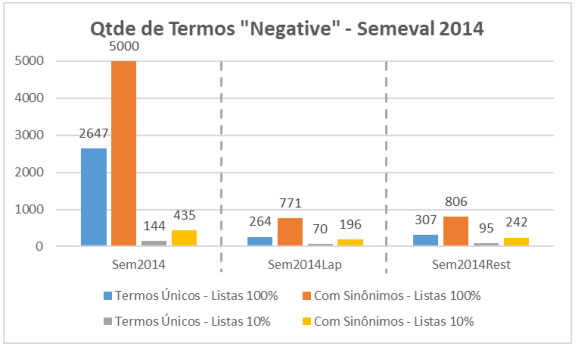
Figura 23 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014* e sub-coleções.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Tabela 7 – Exemplos de termos - Coleção *SemEval 2014* e sub-coleções.

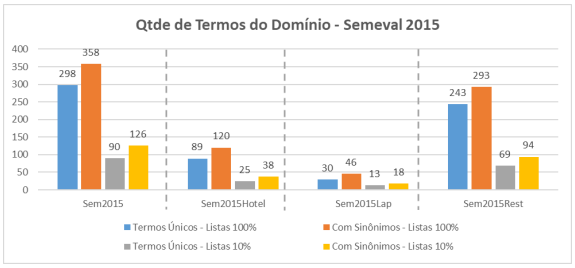
Termos de Domínio	**Identificadores da Classe *Positive***	**Identificadores da Classe *Negative***
operating system; operating systems	large-capacity; large capacity	incompatability; incompatibility; incompatible
bottle of wine; bottles of wine; botle of wine	sustainability; sustainability; sustainable; sustainable	scare; scared;scarier; scariest; scarily; scarred; scars; scary
bathroom; bathrooms	congratulate; congratulate; congratulation; congratulation	worry; worrying; worryingly; worried; worriedly; worrier; worries; worrisome

Fonte: Elaborada pelo autor.

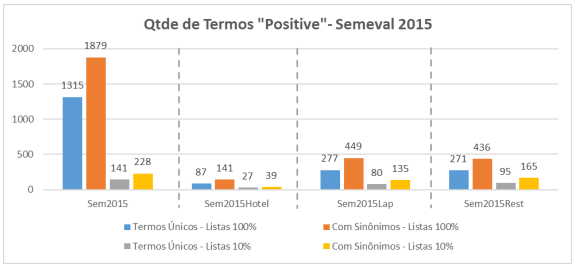
Os gráficos da Figura 24 apresentam uma relação entre a quantidade de termos únicos e com sinônimos para as listas formadas a partir de 10% e 100% da coleção de documentos. No gráfico da Figura 24a pode-se observar as quantidades de termos do domínio, na Figura 24b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 24c observa-se as quantidades para identificadores da classe "*Negative*".

A Tabela 8 apresenta exemplos de termos do domínio e identificadores de classe para as

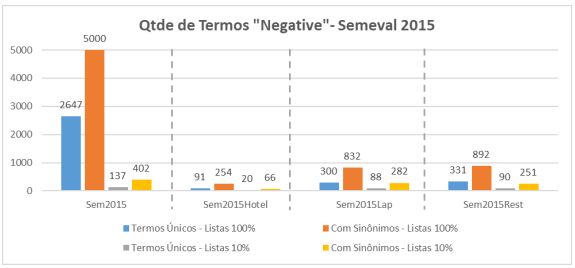
Figura 24 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015* e sub-coleções.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

listas da coleção de documentos *SemEval 2015*.

Tabela 8 – Exemplos de termos - Coleção *SemEval 2015* e sub-coleções.

Termos de Domínio	**Identificadores da Classe *Positive***	**Identificadores da Classe *Negative***
indo chinese food; indo-chinese food	romantically; romantic	anti-social; anti social
customer service	terrifically; terrific	allergies; allergy
elevators; elevator	elegance; elegant	noisier; noises; noise; noisy

Fonte: Elaborada pelo autor.

3.4.2.2 Configuração dos experimentos para validação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

Como apresentado anteriormente, o método proposto possui duas etapas principais. A primeira etapa visa um processo de classificação por meio de um método tradicional e a segunda etapa visa melhorar os resultados cujo grau de confiança da resposta do primeiro classificador foi menor do que um determinado limiar.

O pré-processamento aplicado nas etapas 1 e 2 são referentes à construção das representações *Bag of Words* (BoW), responsável pela construção dos modelos da etapa 1, e das representações gBoED_Freq e gBoED_Dist, responsáveis pela melhoria dos resultados na etapa

2. Na etapa 1, a BoW é construída pela seguinte sequência de técnicas já apresentadas na Subseção 2.1.1 e aplicadas na mesma ordem que aparecem: 1) Padronização de caixa, 2) Tokenização 3) Remoção de pontuação 4) Remoção de caracteres especiais 5) Remoção de acentos (para as coleções de documentos em português 6) Remoção de números 7) Remoção de *Stopwords* 8) Radicalização.

Na etapa 2, ambas as representações, gBoED_Freq e gBoED_Dist, são construídas a partir da seguinte sequência de técnicas de pre-procesamento: 1) Padronização de caixa, 2) Tokenização 3) Remoção de pontuação 4) Remoção de caracteres especiais 5) Remoção de acentos (para as coleções de documentos em português 6) Identificação dos termos de domínio e identificadores de classe

Na Remoção de *Stopwords*, técnica aplicada na etapa 1, foram removidas as *Stopwords* padrão para o idioma inglês e português, de acordo com cada coleção de documentos. Porém, levando em consideração o tipo de classificação semântica que será aplicado nas coleções de documentos utilizadas, necessita-se considerar as negações como fator importante. Em coleções de documentos de opiniões de produtos ou serviços cuja classificação está no nível da análise de sentimentos, a negação de aspectos positivos ocorre com bastante frequência. Portanto, na aplicação da técnica de Remoção de *Stopwords* foi desconsiderada a remoção de palavras que trazem tal sentido, como "não" e "nunca" em português e "*not*", "*do not*", "*don't*", "*won't*", "*can't*" , "*can not*", entre outras, em inglês.

Considerando o pré-processamento aplicado aos conjuntos de documentos e com o objetivo de avaliar o desempenho e as limitações do método proposto, a avaliação experimental é realizada em três diferentes cenários:

- **Oráculo:** esse cenário testa o limite superior do método, ou seja, é o cenário que permite verificar o desempenho do método em uma situação ideal. Como o método prosposto possui duas etapas dependentes uma da outra, admite-se o nome ***Oráculo*** para simular o cenário cuja etapa 2 responde de forma assertiva a classe do documento consultado, quando não for atingido o limiar de confiança. Assim é possível medir quais os valores máximos o método é capaz de atingir.

- **Listas 100%:** o cenário denominado ***Listas 100%*** corresponde àquele cujas listas de termos foram formadas a partir de 100% dos documentos da coleção. É um cenário cujos resultados podem ser considerados como o mais próximo da realidade. A partir dele considera-se o principal resultado dos experimentos.

- **Listas 10%:** esse cenário possui importância semelhante ao Listas 100%. É aquele cujas listas de termos foram formadas a partir de apenas 10% dos documentos da coleção. Ele simula o pior cenário para o método proposto, trazendo seus resultados ao limite inferior.

Em cada cenário apresentado anteriormente, os modelos foram gerados por meio de aprendizado supervisionado. Os algoritmos utilizados são apresentados a seguir, seguidos dos parâmetros de configuração utilizados em cada caso. Nos experimentos executados foi utilizado método de amostragem *10-fold cross validation*. Tais algoritmos e variações são utilizadas na etapa 1 do método.

- **C4.5**, algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor* e critérios para escolha do atributo: Entropia e Gini.
- ***K-nearest neighbor* (KNN)**, algoritmo IBk. Foram utilizadas as opções de voto com peso e voto sem peso. Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: Distância Euclideana e Distância de Cosseno.
- ***Multinomial Naive Bayes* (MNB)**, parâmentro alpha(α) considerando os valores: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} e 1.
- ***Suport Vector Machine* (SVM)**, algoritmo *Sequential Minimal Optimization* (SMO). Nesse algoritmo foram considerados três tipos de kernel: linear, polinomial (expoente=2) e RBF (*Radial Basis Function*). Os valores considerados para cada tipo de kernel foram 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}.

Além da variação dos algoritmos e seus parâmentros, na avaliação experimental também foram utilizados diferentes valores para o grau de confiança que o método de classificação considera para selecionar os documentos para reclassificação. Nos experimentos realizados o grau de confiança variou entre os valores 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% e 100%. Nas tabelas de resultados, apresentadas e detalhadas na Subsubseção 3.4.2.3, é considerado o valor de 75% como um máximo de confiança para comparação do custo x benefício de reclassificar documentos.

Nesse experimento foi utilizada a implementação em linguagem Python na versão 3.7, biblioteca Scikit-Learn na versão 0.22.1, NLTK na versão 3.4.5 e Numpy na versão 1.18.1. Na sequência são apresentados os resultados da avaliação experimental para validação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio.

3.4.2.3 Resultados - validação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

Nessa subsubseção são apresentados os principais resultados obtidos a partir da avaliação experimental executada, para validação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio em diferentes cenários. Os resultados são apresentados individualmente para cada coleção de textos. Iniciando com as coleções em idioma português

e, em seguida, as coleções em inglês, os resultados estão no formato de tabelas contendo as melhores acurácias obtidas pelos modelos gerados por cada algoritmo utilizado.

As Tabelas 9, 11, 13, 15, 17, 19, 21, 23, 25 e 27 demonstram a representatividade da gBoED nos conjuntos de dados. Nelas são apresentadas a quantidade de documentos que nos quais as representações gBoED_Freq e gBoED_Dist puderam representar e que não puderam representar pois as listas de termos não permitiram tal abrangência. Além disso, nelas são apresentadas as quantidades de documentos cujas predições a partir de cada uma das representações semanticamente enriquecidas obtiveram acertos e erros. Os documentos que não obtveram representação por parte das gBoEDs ou que obtiveram empate na predição, nestas tabelas estão sendo considerados como neutros e ao serem consultados dentro do método mantém a predição inicial do classificador.

Já nas Tabelas 10, 12, 14, 16, 18, 20, 22, 24, 26 e 28 são apresentados o resultados com as melhores acurácias obtidas em cada algoritmo. Elas apresentam os algoritmos de acordo com um nível decrescente de explicabilidade. Cada tabela de resultados está organizada de maneira que a linha em cinza corresponde ao melhor resultado obtido para toda a coleção. Na segunda coluna estão as melhores acurácias obtidas a partir dos modelos gerados via BoW. Em seguida as melhores acurácias obtidas a partir dos modelos gerados pelo método de classificação semanticamente enriquecida incrementado pela representação gBoED_Freq e, por último, as melhores acurácias obtidos dos modelos gerados pelo mesmo método e incrementado pela representação gBoED_Dist.

Por último, nos gráficos das Figuras de 25 a 34 é apresentado o custo x benefício de se utilizar a representação semanticamente enriquecida em cada um dos conjuntos de documentos, ou seja, como a predição usando as gBoEDs se comportam em cada cenário e em cada conjuntos de documentos.

Nas tabelas que apresentam as melhores acurácias, os algoritmos estão organizados de acordo com um nível decrescente de explicabilidade. Os resultados obtidos pelo método de classificação enriquecida pela gBoED_Freq e gBoED_Dist estão divididos em três partes. A primeira parte contém os resultados do experimento aplicado no cenário Oráculo (aquele cuja segunda etapa do método reclassifica os documentos sempre de forma correta). A primeira linha é o melhor resultado do oráculo com o limite de confiança do melhor resultado de Lista 100% e a segunda linha é o melhor resultado do valor máximo de confiança de 75%. A segunda parte contém os resultados do experimento aplicado no cenário Listas 100% (aquele cuja gBoED é construída a partir de listas de termos formadas por 100% dos documentos da coleção). A terceira parte contém os resultados do experimento aplicado no cenário Listas 10% (aquele cuja gBoED é construída a partir de listas de termos formadas por apenas 10% dos documentos da coleção).

A forma de leitura das tabelas que apresentam as melhores acurácias deve ser realizada da seguinte maneira. Primeiramente deve-se observar o resultado da melhor acurácia para o

cenário Listas 100%, já que ele corresponde ao cenário mais próximo do real. Em **Acc** é descrita a melhor acurácia e na coluna **Conf** é descrita o grau de confiança em que esse valor de acurácia foi obtido.

No cenário Oráculo, observa-se duas linhas em **Acc**. A primeira contém a melhor acurácia deste cenário com o mesmo limite de confiança atingido pela melhor acurácia do cenário Listas 100%. A primeira linha corresponde à acurácia obtida no cenário Oráculo com grau, por determinado algoritmo, com o mesmo grau de confiança de Listas 100%. A segunda linha contém a melhor acurácia para um limite de confiança de 75%. O valor 75% está sendo considerado como um valor de referência para comparação de acordo com o custo computacional para se realizar reclassificações em documentos.

No cenário Listas 10% é apresentada a melhor acurácia (**Acc**) obtida e o grau de confiança. Tanto em Listas 100% quanto em Listas 10% estão presentes uma sequência de valores, logo abaixo do valor de acurácia, que correspondem à quantidade de documentos enviados para reclassificação devido à confiança ser menor do que o valor descrito, seguido da quantidade desses documentos que foram realmente reclassificados, ou seja, sua predição sofreu alteração de classe. Portando, ao observar tal sequência em Listas 100%, representação gBoED_Freq, algoritmo C4.5-Entropia, grau de confiança 65%, na Tabela 10 por exemplo, obtém-se 33-14. Isso significa que para a confiança de 65%, 33 documentos foram enviados para reclassificação e, desses, 14 sofreram alteração em sua predição para a coleção *B2W Reviews 2019 Info*.

Iniciando a apresentação dos resultados com a coleção de documentos *B2W Reviews 2019 Info*, na Tabela 9 é apresentada a representatividade para o conjunto de documentos. Como esperado, nela é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Porém, também é possível observar que apesar das representações semanticamente enriquecidas atingirem um nível alto de representatividade em relação ao conjunto de documentos, o nível de acerto máximo para Listas 100% ficou em 65.30% para gBoED_Freq e 68.10% para gBoED_Dist.

Na Tabela 10 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *B2W Reviews 2019 Info*. O melhor resultado, destacado pela linha cinza, ocorreu utilizando o algoritmo ***Support Vector Machine* – SVM, kernel RBF, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 92,291%, Medida-F1 de 92,236%**, utilizando a representação **gBoED_Dist como enriquecimento semântico**, $\gamma = 10^{-1}$, grau de confiança de 50%, como 3 documentos sendo consultados e 2 documentos reclassificados. É importante observar que tanto o resultado obtido pelo modelo gerado pela BoW e Listas 10% atingem a acurácia máxima de 92,200%. O limite superior definido pelo Oráculo é de 92,300%, com grau de confiança de 50% e 96,400% com grau de confiança de 75%. Ou seja, em um grau de confiança de 50% gBoED_Dist atingiu um resultado próximo ao limite superior.

Tabela 9 – Representatividade da gBoED no conjunto de dados para *B2W Reviews 2019 Info*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	920	92%	865	86,50%	920	92%	865	86,50%
Qtde de documentos sem representação	80	8%	135	13,5%	80	8%	135	13,5%
Número de ACERTOS na predição	653	65,30%	569	56,89%	681	68,10%	600	60%
Número de ERROS na predição	191	19,10%	216	21,60%	163	16,30%	185	18,50%
Número de NEUTROS na predição	156	15,60%	215	21,51%	156	15,60%	215	21,50%

Fonte: Elaborada pelo autor.

Tabela 10 – Melhores acurácias dos classificadores gerados para *B2W Reviews 2019 Info*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,88600	0,90000 0,92200	65% 75%	0,89000 33-14	65%	0,88700 12-6	60%	0,93500 0,92200	85% 75%	0,89100 157-55	85%	0,88700 12-5	60%
C4.5-Gini	0,87500	0,92200 0,88600	90% 75%	0,87900 175-57	90%	0,87600 1-1	60%	0,93200 0,88600	95% 75%	0,88100 225-69	95%	0,87600 1-1	60%
KNN-Cosseno	0,90100	0,92700 0,98500	55% 75%	0,90200 66-40 $n = 35$	55%	0,89300 46-28 $n = 35$	55%	0,95800 0,98600	60% 75%	0,90800 128-67 $n = 45$	60%	0,89700 46-25 $n = 45$	55%
KNN-Euclideana	0,89700	0,92700 0,98500	55% 75%	0,89900 67-40 $n = 35$	55%	0,89000 47-29 $n = 35$	55%	0,95900 0,98500	60% 75%	0,90500 70-45 $n = 35$	60%	0,89400 47-26 $n = 35$	55%
MNB	0,90400	0,92400 0,98900	55% 75%	0,89700 50-36 $\alpha = 1$	55%	0,89200 50-40 $\alpha = 1$	55%	0,95500 0,98900	60% 75%	0,90700 123-64 $\alpha = 1$	60%	0,89700 50-35 $\alpha = 1$	55%
SVM-Linear	0,90900	0,93500 0,95800	60% 75%	0,91300 51-33 $\gamma = 10^{-4}$ a 10^4	60%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,94400 0,95800	65% 75%	0,91700 73-40 $\gamma = 10^{-4}$ a 10^4	65%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,91500	0,92400 0,96400	55% 75%	0,91500 18-14 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%	0,94900 0,96400	65% 75%	0,91900 79-44 $\gamma = 10$	65%	0,91400 1-1 $\gamma = 10$	50%
SVM-RBF	0,92200	0,92300 0,96400	50% 75%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92300 0,96400	50% 75%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

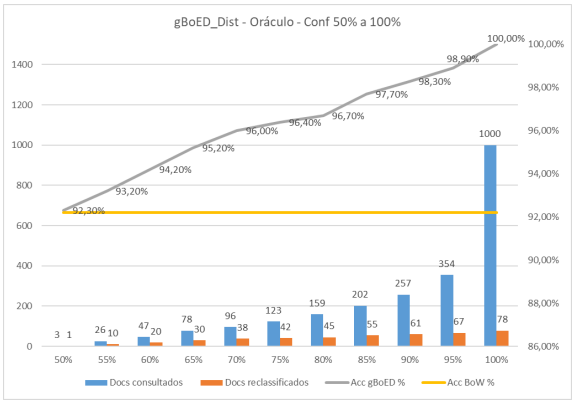
Sobre os resultados apresentados na Tabela 10, vale destacar que para os algoritmos C4.5-Entropia, C4.5-Gini, KNN-Cosseno, KNN-Euclidiana, SVM-Linear e SVM-Polinomial, apesar de não obterem os melhores resultados, em Listas 100%, tanto gBoED_Freq quanto gBoED_Dist conseguiram melhorar o valores de acurácia em relação aos modelos BoW. Comparativamente, em 548 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq. Assim, para esse conjunto de documentos o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Na Figura 25, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-RBF), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-RBF, $\gamma = 10^{-1}$. No cenário ideal, apresentado na Figura 25a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 25b Listas 100% e Figura 25c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 65% de assertividade com relação à classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção de documentos *B2W Reviews 2019 Info*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

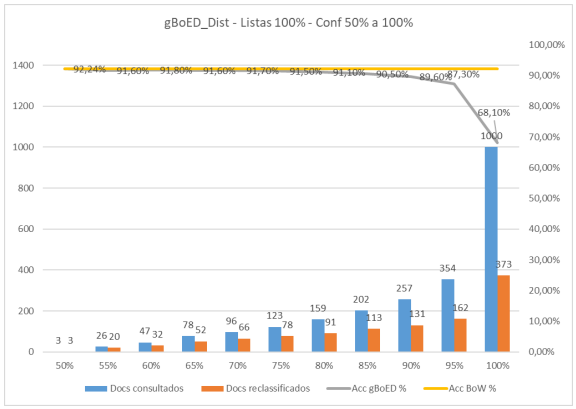
Em seguida são analisados os resultados para a coleção de documentos *BestSports Top 4*. Na Tabela 11 é apresentada a representatividade para este conjunto de documentos. Como esperado, nela é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Para este conjunto, na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação ao consultar a representação gBoED_Dist é maior do que gBoED_Freq, 56,35% e 52,48%, respectivamente.

Na Tabela 12 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *BestSports Top 4*. Nela, o melhor resultado, destacado pela linha cinza, ocorreu utilizando o algoritmo ***Support Vector Machine* – SVM, kernel Linear, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 71,823%, Medida-F1 de 69,736%**, utilizando as **representações gBoED_Freq e gBoED_Dist como enriquecimento semântico**, $\gamma = 10^{-4} a 10^{-4}$, grau de confiança de 50%, com 4 documentos sendo consultados e 1 documento reclassificado. É importante observar que tanto o resultado obtido pelo modelo gerado pela BoW e Listas 10% atingem a acurácia máxima de 71,315%. O limite superior definido pelo Oráculo é de 71,871%, com grau de confiança de 50% e 93,333% com grau de confiança de 75%. Ou seja, em um grau de confiança de 50% gBoED_Freq e gBoED_Dist atingiram um resultado próximo ao limite superior. Comparativamente, em 489 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq. Assim, para esse conjunto de documentos o esquema de ponderação baseado na

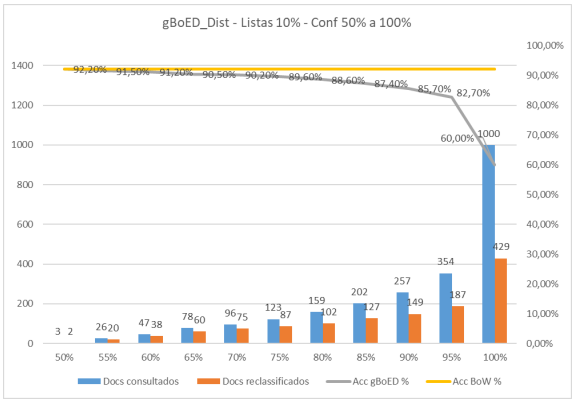
Figura 25 – Uso da gBoED_Dist nos cenários de *B2W Reviews 2019 Info* para SVM-RBF, $\gamma = 10^{-1}$.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Tabela 11 – Representatividade da gBoED no conjunto de dados para *BestSports Top 4*.

	gBoED_Freq				**gBoED_Dist**			
	Listas_100%		**Listas_10%**		**Listas_100%**		**Listas_10%**	
Qtde de documentos representados	162	89,50%	161	88,95%	162	89,50%	161	88,95%
Qtde de documentos sem representação	19	10,49%	20	11,04%	19	10,49%	20	11,04%
Número de ACERTOS na predição	95	52,48%	91	50,27%	102	56,35%	97	53,59%
Número de ERROS na predição	54	29,83%	54	29,83%	47	25,96%	48	26,51%
Número de NEUTROS na predição	32	15,60%	36	19,88%	32	17,67%	36	19,88%

Fonte: Elaborada pelo autor.

distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Sobre os resultados apresentados na Tabela 12, vale destacar que para os algoritmos

C4.5-Entropia e C4.5-Gini, apesar de não obterem os melhores resultados, em Listas 100%, tanto gBoED_Freq quanto gBoED_Dist conseguiram melhorar o valores de acurácia em relação aos modelos BoW. C4.5-Entropia, em Listas 100% e gBoED_Dist obteve um resultado superior em quase 5%.

Tabela 12 – Melhores acurácias dos classificadores gerados para *BestSports Top 4*.

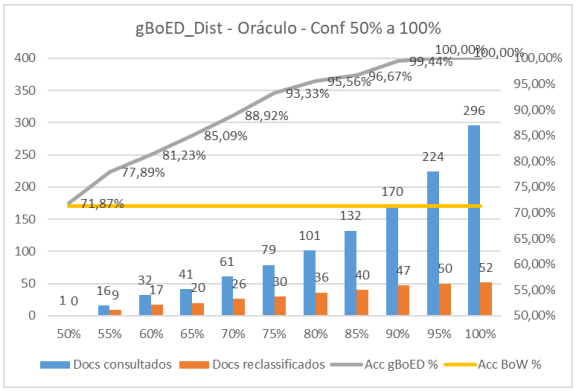
Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,51871	0,58479 0,59035	60% 75%	0,52953 19-4	60%	0,53508 46-7	90%	1,00000 0,59035	100% 75%	0,56432 181-85	100%	0,53654 181-82	100%
C4.5-Gini	0,58596	0,71754 0,59152	95% 75%	0,59707 52-12	95%	0,60263 52-11	95%	0,71754 0,59152	95% 75%	0,60818 52-11	95%	0,61374 52-10	95%
KNN-Cosseno	0,70643	0,69035 0,93888	55% 75%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,69035 0,93888	55% 75%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%
KNN-Euclideana	0,70643	0,69035 0,93888	55% 75%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,69035 0,93888	55% 75%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%
MNB	0,69064	0,66783 0,75029	55% 75%	0,67368 22-10 $\alpha = 1$	55%	0,67368 16-12 $\alpha = 1$	60%	0,66783 0,75029	55% 75%	0,67368 22-11 $\alpha = 1$	55%	0,67368 22-11 $\alpha = 1$	55%
SVM-Linear	0,71315	0,71871 0,93333	50% 75%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,71871 0,93333	50% 75%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,71286	0,71286 0,93333	50% 75%	0,70730 2-1 $\gamma = 10^{-1}$	50%	0,71286 2-0 $\gamma = 10^{-1}$	50%	0,71286 0,93333	50% 75%	0,70730 2-1 $\gamma = 10^{-1}$	50%	0,71286 2-0 $\gamma = 10^{-1}$	50%
SVM-RBF	0,70175	0,74590 0,93362	55% 75%	0,69619 18-5 $\gamma = 1$	55%	0,69619 18-5 $\gamma = 1$	55%	0,74590 0,93362	55% 75%	0,69619 18-4 $\gamma = 1$	55%	0,69619 18-4 $\gamma = 1$	55%

Fonte: Elaborada pelo autor.

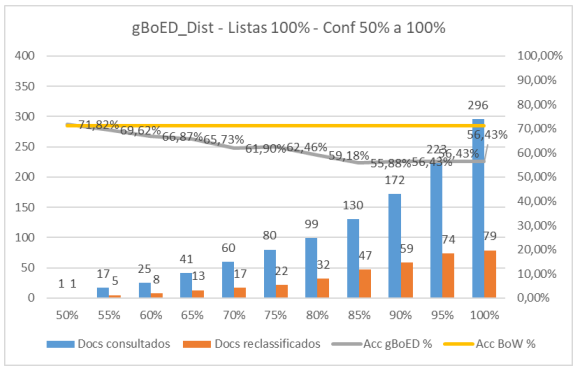
Na Figura 26, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-Linear), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-Linear, $\gamma = 10^{-4}$ a 10^4. No cenário ideal, apresentado na Figura 26a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 26b Listas 100% e Figura 26c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 56% de assertividade com relação à classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção de documentos *BestSports Top 4*, o melhor custo x benefício

ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

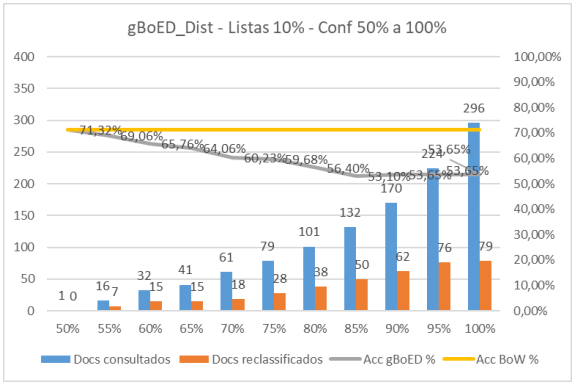
Figura 26 – Uso da gBoED_Dist nos cenários de *BestSports Top 4*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Iniciando a análise dos resultados das coleções de documentos em inglês, os próximos resultados estão relacionados à coleção de documentos *HuLiu 2004*. Na Tabela 13 é apresentada a representatividade para este conjunto de documentos. Como esperado, nela é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Para este conjunto, na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação ao consultar a representação gBoED_Dist é maior do que gBoED_Freq, 74,32% e 73,64%, respectivamente.

Na Tabela 14 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *HuLiu 2004*. O melhor resultado está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel Polinomial, com Listas 100%, enriquecimento pela representação gBoED_Freq**. Nesse cenário, a **acurácia obtida foi de 85,816%, Medida-F1 de 85,816%**, utilizando a representação gBoED_Dist como enriquecimento semântico, $\gamma = 10^{-1}$, grau de confiança de 65%, com 48 documentos sendo consultados

Tabela 13 – Representatividade da gBoED no conjunto de dados para *HuLiu 2004*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	285	96.28%	280	94.59%	285	96.28%	280	94.59%
Qtde de documentos sem representação	11	3.71%	16	5.40%	11	3.71%	16	5.40%
Número de ACERTOS na predição	218	73.64%	215	72.63%	220	74.32%	220	74.32%
Número de ERROS na predição	57	19.25%	54	18.24%	55	18.58%	49	16.55%
Número de NEUTROS na predição	21	7.09%	27	9.12%	21	7.09%	27	9.12%

Fonte: Elaborada pelo autor.

e 21 documentos reclassificados. Para gBoED_Dist, Listas 100%, o valor da acurácia foi de 85,483%, Medida-F1 de 85,483%, $\gamma = 10^{-1}$, grau de confiança de 70%, com 59 documentos sendo consultados e 21 documentos reclassificados. Vale destacar que em todos os algoritmos apresentados, apesar de não obterem os melhores resultados, em Listas 100%, tanto gBoED_Freq quanto gBoED_Dist conseguiram melhorar o valores de acurácia em relação aos modelos BoW. Comparativamente, em Listas 100% ocorreram 322 casos que gBoED_Freq obteve maior acurácia do que o gBoED_Dist. Assim, para esse conjunto de documentos o esquema de ponderação baseado na frequência entre os termos ainda apresenta um impacto positivo na efetividade da gBoED.

Na Figura 27, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Freq (no cenário com melhor acurácia, SVM-Polinomial), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM, kernel Polinomial, $\gamma = 10^{-1}$. No cenário ideal, apresentado na Figura 27a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 27b (Listas 100%) e Figura 27c (Listas 10%), a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. A coleção Huliu 2004 apresenta uma característica de aumento da acurácia relacionada ao aumento de documentos consultados e reclassificados entre os níveis de confiança de 50% a 70% aproximadamente, porém a partir dessa faixa de confiança a acurácia

Tabela 14 – Melhores acurácias dos classificadores gerados para *HuLiu 2004*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,68988	0,82735 0,80375	90% 75%	0,74057 86-34	90%	0,72678 296-118	100%	0,82735 0,80379	90% 75%	0,74390 86-32	90%	0,74345 296-109	100%
C4.5-Gini	0,67609	1,00000 0,67551	100% 75%	0,73701 296-114	100%	0,72678 296-111	100%	1,00000 0,67551	100% 75%	0,74379 296-118	100%	0,74345 296-106	100%
KNN-Cosseno	0,78712	0,93563 0,93563	70% 75%	0,80781 121-57 $n = 9$	70%	0,79402 63-35	55%	0,93563 0,93563	70% 75%	0,79436 121-55 $n = 9$	70%	0,79391 63-31	55%
KNN-Euclideana	0,78712	0,93563 0,93563	70% 75%	0,80781 121-57 $n = 9$	70%	0,79402 63-35	55%	0,93563 0,93563	70% 75%	0,79436 121-55 $n = 9$	70%	0,79391 63-31	55%
MNB	0,81402	0,88551 0,89908	70% 75%	0,83793 48-29 $\alpha = 10^{-2}$	70%	0,82115 14-5	55%	0,88551 0,89908	70% 75%	0,83126 48-31 $\alpha = 10^{-2}$	70%	0,81770 14-8	55%
SVM-Linear	0,82793	0,92620 0,92954	70% 75%	0,85161 66-30 $\gamma = 10^{-4}$ a 10^4	70%	0,84149 62-27 $\gamma = 10^{-4}$ a 10^4	70%	0,92620 0,92954	65% 75%	0,85149 47-21 $\gamma = 10^{-4}$ a 10^4	65%	0,83804 62-26 $\gamma = 10^{-4}$ a 10^4	70%
SVM-Polinomial	0,83436	0,91942 0,92954	70% 75%	0,85816 59-21 $\gamma = 10^{-1}$	70%	0,84827 60-17 $\gamma = 10^{-1}$	70%	0,90241 0,92954	65% 75%	0,85483 48-21 $\gamma = 10^{-1}$	65%	0,84149 59-20 $\gamma = 10^{-1}$	70%
SVM-RBF	0,81425	0,91597 0,92597	70% 75%	0,84149 72-35 $\gamma = 1$	70%	0,83839 69-33 $\gamma = 1$	70%	0,91597 0,92597	70% 75%	0,83805 72-29 $\gamma = 1$	70%	0,82816 69-29 $\gamma = 1$	70%

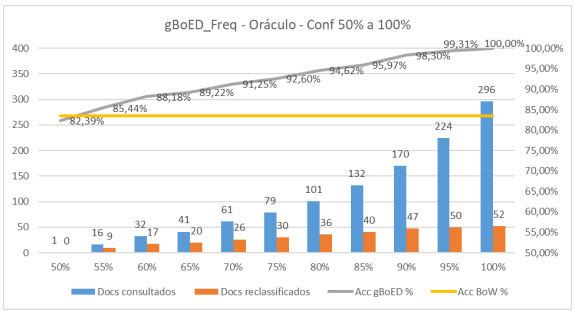
Fonte: Elaborada pelo autor.

volta a diminuir. Essa característica mostra que o conjunto de termos e, consequentemente, as representações enriquecidas conseguiram representar a coleção de forma mais eficiente. Para a coleção de documentos *HuLiu 2004*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

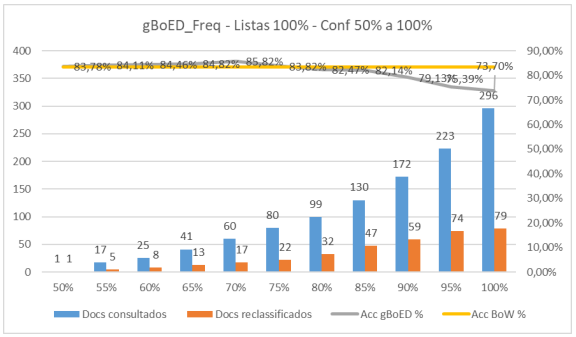
O próximo conjunto de resultados a serem analisados são referentes à coleção de documentos *SemEval 2014*. Como apresentado em Subsubseção 3.4.2.1, essa coleção possui a característica de conter documentos de análise de sentimentos de dois domínios diferentes: *Laptops* e *Restaurants*. Na Tabela 15 é apresentada a representatividade para este conjunto de documentos. Nela é possível observar que pelo conjunto de termos identificados como termos de domínio e identificadores de classe e pelo fato de existir mais de um domínio associado à mesma coleção, tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos. Na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação fica em torno de 60% para ambas. A taxa de erro é baixa, principalmente para gBoED_Dist. Porém, a taxa de Neutros (aqueles documentos que não puderam ser reclassificados pelas gBoED), que em coleções anteriores estavam entre 10% e 20%, nessa coleção aumentou para mais de 30%.

Na Tabela 16 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014*. O melhor resultado está destacado pela linha cinza e se dá pelo algoritmo *Support Vector Machine* – SVM, kernel RBF, com Listas 100%. Nesse

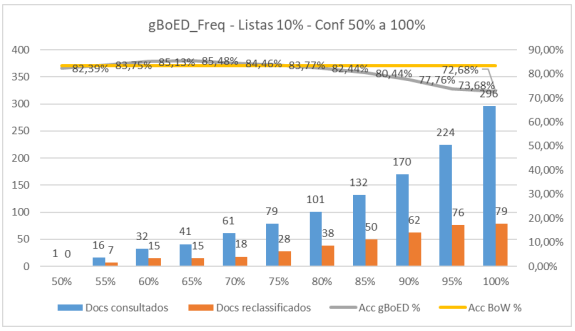
Figura 27 – Uso da gBoED_Freq nos cenários de *HuLiu 2004*.



(a) gBoED_Freq Oráculo



(b) gBoED_Freq Listas 100%



(c) gBoED_Freq Listas 10%

Fonte: Elaborada pelo autor.

Tabela 15 – Representatividade da gBoED no conjunto de dados para *SemEval 2014*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	2186	75,14%	1820	62,56%	2186	75,14%	1820	62,56%
Qtde de documentos sem representação	723	24,85%	1089	37,43%	723	24,85%	1089	37,43%
Número de ACERTOS na predição	1719	59,09%	1436	49,36%	1830	62,90%	1513	52,01%
Número de ERROS na predição	275	9,45%	247	8,49%	164	5,63%	170	5,84%
Número de NEUTROS na predição	915	31,45%	1226	42,14%	915	31,45%	1226	42,14%

Fonte: Elaborada pelo autor.

cenário, a acurácia obtida foi de 80,407%, Medida-F1 de 80,407%, utilizando a representação gBoED_Dist como enriquecimento semântico, $\gamma = 1$, grau de confiança de 55%, com 185 documentos sendo consultados e 140 documentos reclassificados. Para gBoED_Freq, Listas 100%, o valor da acurácia foi de 80,235%, Medida-F1 de 80,235%, $\gamma = 1$, grau de confiança de 50%, com 8 documentos sendo consultados e 6 documentos reclassificados. Ambos contra uma

acurácia de 79,925% do modelo formado pela BoW. Vale destacar que os modelos formados pelos algoritmos C4.5-Entropia, C4.5-Gini, KNN-Euclidiana e SVM-Polinomial, apesar de não obterem os melhores resultados, em Listas 100%, tanto gBoED_Freq quanto gBoED_Dist conseguiram melhorar o valores de acurácia em relação aos modelos BoW. Comparativamente, em Listas 100%, todos os 601 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq. Assim, para esse conjunto de documentos o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Tabela 16 – Melhores acurácias dos classificadores gerados para *SemEval 2014*.

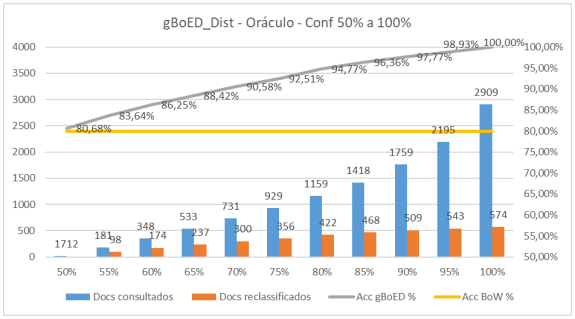
Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,63700	0,67044 0,67113	55% 75%	0,63871 2082-1239	55%	0,54348 2088-1390	55%	0,67044 0,67113	60% 75%	0,66759 2082-1153	55%	0,56032 2088-1331	55%
C4.5-Gini	0,64525	0,66322 0,67387	60% 75%	0,65005 1235-707	55%	0,57818 27-12	55%	0,66322 0,67387	60% 75%	0,66587 1235-654	55%	0,59090 27-12	55%
KNN-Cosseno	0,77760	0,79924 0,95015	55% 75%	0,75973 317-214 $n = 17$	55%	0,74355 40-29 $n = 17$	55%	0,79924 0,95015	55% 75%	0,76419 421-254 $n = 13$	55%	0,74767 40-29 $n = 13$	55%
KNN-Euclideana	0,77691	0,80990 0,95840	55% 75%	0,76076 326-217 $n = 17$	55%	0,74871 39-28 $n = 17$	55%	0,80990 0,95840	55% 75%	0,76523 326-207 $n = 17$	55%	0,75180 39-28 $n = 17$	55%
MNB	0,80648	0,80681 0,87831	55% 75%	0,79823 141-100 $\alpha = 10^{-1}$	55%	0,78584 24-19 $\alpha = 10^{-1}$	55%	0,82365 0,91784	55% 75%	0,80063 141-94 $\alpha = 10^{-1}$	55%	0,78687 35-25 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79718	0,79649 0,93056	50% 75%	0,79581 25-18 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 17-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79649 0,93056	60% 75%	0,79615 25-17 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 16-8 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,78824	0,79614 0,91715	50% 75%	0,79615 19-13 $\gamma = 10^{-1}$	50%	0,78053 17-14 $\gamma = 10^{-1}$	50%	0,79614 0,91715	50% 75%	0,79650 19-12 $\gamma = 10^{-1}$	50%	0,78145 17-14 $\gamma = 10^{-1}$	50%
SVM-RBF	0,79925	0,80680 0,92505	50% 75%	0,80235 8-6 $\gamma = 1$	50%	0,78045 17-14 $\gamma = 1$	50%	0,80680 0,92505	50% 75%	0,80407 185-140 $\gamma = 1$	55%	0,78045 17-14 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

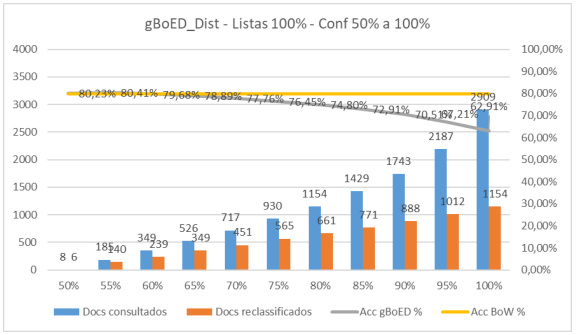
Na Figura 28, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-RBF), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-RBF, $\gamma = 1$. No cenário ideal, apresentado na Figura 28a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 28b Listas 100% e Figura 28c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 65% de assertividade com relação à

classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção de documentos *SemEval 2014*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está entre 50% e 55%, em seguida a acurácia começa a cair.

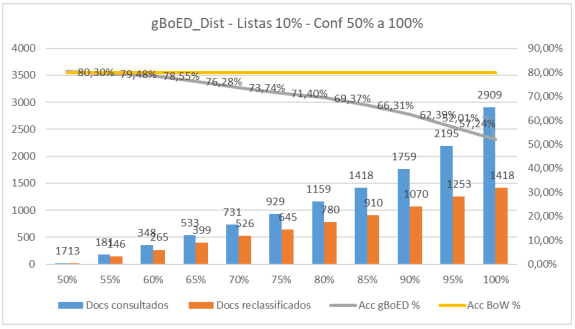
Figura 28 – Uso da gBoED_Dist nos cenários de *SemEval 2014*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Especificando os domínios da coleção *SemEval 2014*, o próximo conjunto de resultados a serem analisados são referentes à coleção de documentos *SemEval 2014 Laptop*. Na Tabela 17 são apresentados os valores de completude e representatividade para este conjunto de documentos. Como esperado, nela é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação está entre 57.93% para gBoED_Freq e 61.64% para gBoED_Dist. A taxa de erro é baixa, principalmente para gBoED_Dist e a taxa de Neutros fica em 32,39% para ambas as representações.

Na Tabela 18 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Laptop*. O melhor resultado está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 80,254%, Medida-F1 de 80,356%**, utilizando a **representação gBoED_Dist como enriquecimento semântico**, $\gamma = 1$, grau de confiança de 50%,

Tabela 17 – Representatividade da gBoED no conjunto de dados para *SemEval 2014 Laptop*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	916	73,81%	627	50,52%	916	73,81%	627	50,52%
Qtde de documentos sem representação	325	26,18%	614	49,47%	325	26,18%	614	49,47%
Número de ACERTOS na predição	719	57,93%	485	39,08%	765	61,64%	503	40,53%
Número de ERROS na predição	120	9,67%	108	8,70%	74	5,96%	90	7,25%
Número de NEUTROS na predição	402	32,39%	648	52,21%	402	32,39%	648	52,21%

Fonte: Elaborada pelo autor.

com 8 documentos sendo consultados e 8 documentos reclassificados. Para gBoED_Freq, Listas 100%, o valor da acurácia se manteve igual à acurácia apresentada pelo modelo gerado pela BoW que foi de 80,173%. Outro destaque é o modelo formado pelo algoritmo SVM-Polinomial, que apesar de não obter o melhor resultado, em Listas 100%, gBoED_Dist atingiu melhor acurácia em relação ao modelo BoW. Comparativamente, em Listas 100%, todos os 622 casos, gBoED_-Dist obteve maior acurácia do que o gBoED_Freq. Assim, para esse conjunto de documentos o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Na Figura 29, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-RBF), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-RBF, $\gamma = 1$. No cenário ideal, apresentado na Figura 29a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 29b Listas 100% e Figura 29c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 65% de assertividade com relação à classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção

Tabela 18 – Melhores acurácias dos classificadores gerados para *SemEval 2014 Laptop*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,67436	0,72436 0,96774	60% 75%	0,65823 174-122	60%	0,62194 174-152	60%	0,72436 0,96774	60% 75%	0,66388 174-116	60%	0,62194 174-152	60%
C4.5-Gini	0,68000	0,70661 0,95969	60% 75%	0,67355 88-62	60%	0,65500 88-77	60%	0,70661 0,95969	60% 75%	0,67677 88-59	60%	0,65500 88-77	60%
KNN-Cosseno	0,77352	0,81864 0,96211	55% 75%	0,76869 126-82 $n = 25$	55%	0,74615 126-100 $n = 25$	55%	0,81864 0,96211	55% 75%	0,77352 126-79 $n = 25$	55%	0,75097 126-98 $n = 25$	55%
KNN-Euclideana	0,77757	0,82429 0,97662	55% 75%	0,76790 141-94 $n = 25$	55%	0,74051 141-115 $n = 25$	55%	0,82429 0,97662	55% 75%	0,77353 141-90 $n = 25$	55%	0,74534 141-113 $n = 25$	55%
MNB	0,79535	0,84609 0,92345	60% 75%	0,78727 137-99 $\alpha = 10^{-1}$	60%	0,77519 74-61 $\alpha = 10^{-1}$	55%	0,81792 0,92345	55% 75%	0,79291 137-95 $\alpha = 10^{-1}$	60%	0,77519 74-60 $\alpha = 10^{-1}$	55%
SVM-Linear	0,80093	0,80174 0,91777	50% 75%	0,79771 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79852 5-5 $\gamma = 10^{-4}$ a 10^4	50%	0,80174 0,91777	50% 75%	0,79851 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79932 5-4 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79610	0,79932 0,91777	50% 75%	0,79530 75-59 $\gamma = 10^{-1}$	55%	0,79449 9-9 $\gamma = 10^{-1}$	50%	0,82673 0,91777	55% 75%	0,79772 75-54 $\gamma = 10^{-1}$	55%	0,79449 9-9 $\gamma = 10^{-1}$	50%
SVM-RBF	0,80173	0,80657 0,92099	50% 75%	0,80173 8-8 $\gamma = 1$	50%	0,80173 8-6 $\gamma = 1$	50%	0,80657 0,92099	50% 75%	0,80254 8-8 $\gamma = 1$	50%	0,80173 8-6 $\gamma = 1$	50%

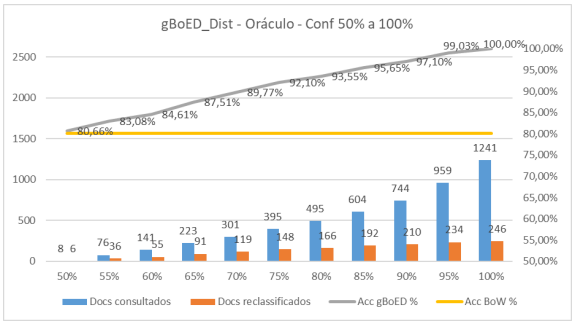
Fonte: Elaborada pelo autor.

de documentos *SemEval 2014 Laptop*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

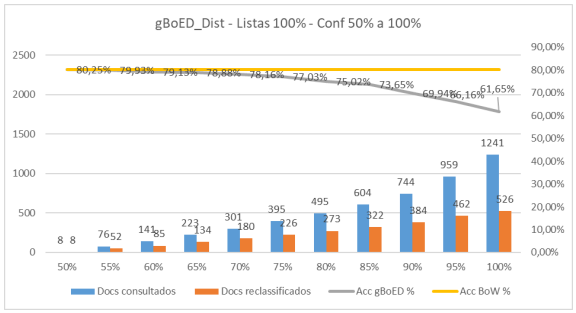
Ainda especificando os domínios da coleção *SemEval 2014*, o próximo conjunto de resultados a serem analisados são referentes à coleção de documentos *SemEval 2014 Restaurant*. Na Tabela 19 são apresentados os valores de completude e representatividade para este conjunto de documentos. Como esperado, semelhante à coleção de documentos *SemEval 2014 Laptop*, é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação está entre 61.03% para gBoED_Freq e 64.80% para gBoED_Dist. A taxa de erro é baixa, principalmente para gBoED_Dist e a taxa de Neutros fica em 29.55% para ambas as representações.

Na Tabela 20 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Restaurant*. O melhor resultado está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 81,658%, Medida-F1 de 80,890%**, utilizando a **representação gBoED_Dist como enriquecimento semântico**, $\gamma = 1$, grau de confiança de 50%, com 6 documentos sendo consultados e 2 documentos reclassificados. Para gBoED_Freq, Listas 100%, o valor da acurácia se manteve igual à acurácia apresentada pelo modelo gerado

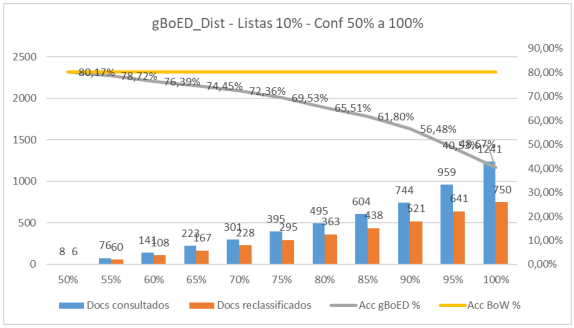
Figura 29 – Uso da gBoED_Dist nos cenários de *SemEval 2014 Laptop*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Tabela 19 – Representatividade da gBoED no conjunto de dados para *SemEval 2014 Restaurant*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	1286	77,09%	1025	61,45%	1286	77,09%	1025	61,45%
Qtde de documentos sem representação	382	22,90%	643	38,54%	382	22,90%	643	38,54%
Número de ACERTOS na predição	1018	61,03%	823	49,34%	1081	64,80%	853	51,13%
Número de ERROS na predição	157	9,41%	140	8,39%	94	5,63%	110	6,59%
Número de NEUTROS na predição	493	29,55%	705	42,26%	493	29,55%	705	42,26%

Fonte: Elaborada pelo autor.

pela BoW que foi de 81,598%. Outros destaques são os modelos formados pelos algoritmos C4.5-Gini, MNB e SVM-Polinomial, que apesar de não obter o melhor resultado, em Listas 100%, atingiram melhor acurácia em relação ao modelo BoW. Comparativamente, em Listas 100%, todos os 561 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq. Assim, para esse conjunto de documentos o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Tabela 20 – Melhores acurácias dos classificadores gerados para *SemEval 2014 Restaurant*.

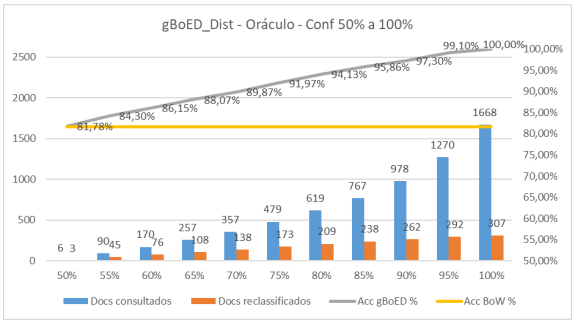
Algoritmos	BoW	gBoED_Freq						gBoED_Dist					
		Oráculo		Listas 100%		Listas 10%		Oráculo		Listas 100%		Listas 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75003	0,75001 0,95444	50% 75%	0,74940 1-0	50%	0,74940 1-1	50%	0,75001 0,95444	50% 75%	0,74940 1-0	50%	0,74940 1-1	50%
C4.5-Gini	0,75543	0,76199 0,90355	60% 75%	0,76139 3-3	60%	0,76139 3-3	60%	0,76199 0,90355	60% 75%	0,76139 3-3	60%	0,76139 3-3	60%
KNN-Cosseno	0,79196	0,83454 0,92327	55% 75%	0,78059 164-108 $n = 13$	55%	0,76081 164-126 $n = 13$	55%	0,83454 0,92327	55% 75%	0,78478 164-101 $n = 13$	55%	0,76200 164-123 $n = 13$	55%
KNN-Euclideana	0,78959	0,83574 0,92687	55% 75%	0,78058 166-110 $n = 13$	55%	0,76080 166-128 $n = 13$	55%	0,83574 0,92687	55% 75%	0,78478 166-103 $n = 13$	55%	0,76200 166-125 $n = 13$	55%
MNB	0,79856	0,85012 0,91006	60% 75%	0,79976 162-108 $\alpha = 10^{-1}$	60%	0,79315 78-65 $\alpha = 10^{-1}$	55%	0,85012 0,91006	60% 75%	0,80575 162-95 $\alpha = 10^{-1}$	60%	0,79495 78-62 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79438	0,79678 0,92387	50% 75%	0,79438 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79678 0,92387	50% 75%	0,79438 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79558	0,79798 0,91548	50% 75%	0,79558 7-7 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%	0,79798 0,91548	50% 75%	0,79558 7-7 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%
SVM-RBF	0,81598	0,81778 0,91969	50% 75%	0,81658 6-2 $\gamma = 1$	50%	0,81598 6-3 $\gamma = 1$	50%	0,81778 0,91969	50% 75%	0,81658 6-2 $\gamma = 1$	50%	0,81598 6-3 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

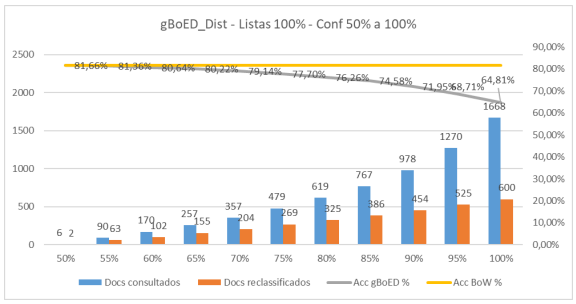
Na Figura 30, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-RBF), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-RBF, $\gamma = 1$. No cenário ideal, apresentado na Figura 30a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 30b Listas 100% e Figura 30c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 65% de assertividade com relação à classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção de documentos *SemEval 2014 Laptop*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

O próximo conjunto de resultados a serem analisados são referentes à coleção de do-

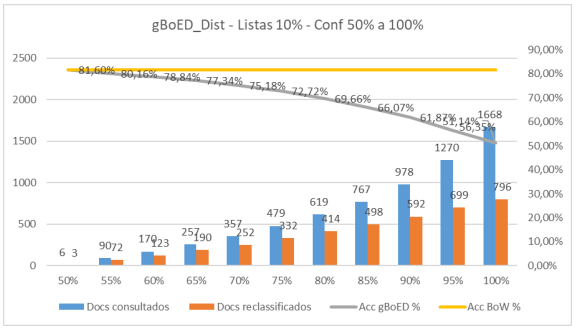
Figura 30 – Uso da gBoED_Dist nos cenários de *SemEval 2014 Restaurant*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

cumentos *SemEval 2015*. Como apresentado em Subsubseção 3.4.2.1, essa coleção possui a característica de conter documentos de análise de sentimentos de domínios diferentes: *Hotels*, *Laptops* e *Restaurants*. Na Tabela 21 são apresentados os valores de completude e representatividade para este conjunto de documentos. Isso significa que para essa coleção de documentos, as listas de termos não representam tão bem quanto em outras coleções.

Tabela 21 – Representatividade da gBoED no conjunto de dados para *SemEval 2015*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	447	55,80%	424	52,93%	447	55,80%	424	52,93%
Qtde de documentos sem representação	354	44,19%	377	47,06%	354	44,19%	377	47,06%
Número de ACERTOS na predição	351	43,82%	321	40,07%	357	44,56%	334	41,69%
Número de ERROS na predição	72	8,98%	79	9,86%	66	8,23%	66	8,23%
Número de NEUTROS na predição	378	47,19%	401	50,06%	378	47,19%	401	50,06%

Fonte: Elaborada pelo autor.

Na Tabela 22 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015*. Nessa coleção, devido à baixa representação das gBoEDs, os melhores resultados foram obtidos pelos modelos gerados pela BoW. O melhor resultado ocorreu por meio do algoritmo ***Support Vector Machine* – SVM, kernel RBF**. Nesse modelo, a **acurácia foi de 87,390% com Medida-F1 de 87,325%**. O enriquecimento semântico não foi eficaz e se manteve no nível mínimo das Listas 10% com **acurácia de 87,142%** em ambos.

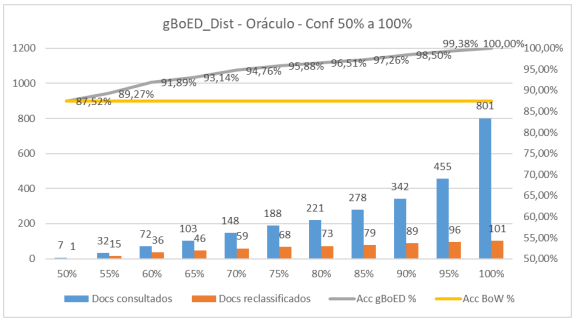
Tabela 22 – Melhores acurácias dos classificadores gerados para *SemEval 2015*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75404	0,77029 0,92386	60% 75%	0,74904 32-18	60%	0,74529 32-17	55%	0,77029 0,92386	60% 75%	0,74779 32-17	60%	0,74529 32-17	60%
C4.5-Gini	0,77528	0,78903 0,91261	60% 75%	0,76028 31-21	60%	0,76028 31-20	60%	0,78903 0,91261	60% 75%	0,76028 31-20	60%	0,76028 31-20	60%
KNN-Cosseno	0,81526	0,83148 0,94756	55% 75%	0,79778 40-30 $n = 25$	55%	0,79528 40-30 $n = 25$	55%	0,83148 0,94756	55% 75%	0,79653 40-29 $n = 25$	55%	0,79528 40-31 $n = 25$	55%
KNN-Euclideana	0,81651	0,83148 0,94756	55% 75%	0,79903 39-29 $n = 25$	55%	0,79653 39-29 $n = 25$	55%	0,83148 0,94756	55% 75%	0,79778 39-28 $n = 25$	55%	0,79653 39-30 $n = 25$	55%
MNB	0,86145	0,87018 0,92261	55% 75%	0,85145 24-19 $\alpha = 10^{-2}$	55%	0,85020 24-18 $\alpha = 10^{-2}$	55%	0,87018 0,92261	55% 75%	0,85148 35-25 $\alpha = 10^{-1}$	55%	0,85020 24-17 $\alpha = 10^{-2}$	55%
SVM-Linear	0,86148	0,86518 0,94640	50% 75%	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,90392 0,94640	60% 75%	0,86398 61-48 $\gamma = 10^{-4}$ a 10^4	60%	0,86148 61-49 $\gamma = 10^{-4}$ a 10^4	60%
SVM-Polinomial	0,87270	0,70659 0,84893	50% 75%	0,87020 5-3 $\gamma = 1$	50%	0,87020 5-2 $\gamma = 1$	50%	0,70659 0,84893	50% 75%	0,87270 5-4 $\gamma = 1$	50%	0,87147 30-16 $\gamma = 1$	55%
SVM-RBF	0,87390	0,87015 0,95011	50% 75%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87015 0,95011	50% 75%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87142 2-2 $\gamma = 10^{-1}$	50%

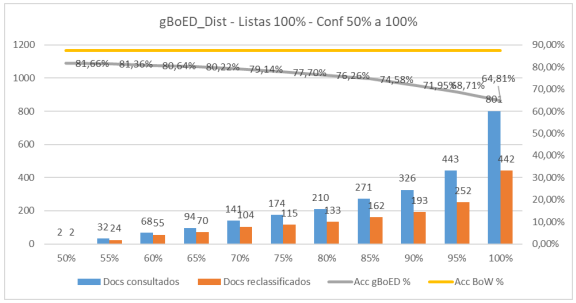
Fonte: Elaborada pelo autor.

Mesmo não obtendo o melhores resultados devido a baixa representatividade das representações semanticamente enriquecidas, na Figura 31 serão apresentados os gráficos comparativos que mostram o custo e relação à quantidade de consultas à gBoED. A figura contém os dados relacionados à quantidade de documentos que necessitaram consulta à gBoED_Dist algoritmo SVM-RBF, nível de confiança, $\gamma = 10^{-1}$. Foi escolhido esse cenário pois apresentou melhor resultado usando BoW. Verifica-se que no cenário ideal Figura 31a que conforme o grau de confiança aumenta, a quantidade de documentos consultados aumenta proporcionalmente. A quantidade de documentos que sofre reclassificação também aumenta, porém em uma proporção menor. A acurácia final do classificador também aumenta chegando ao nível máximo de 100%. Nos cenários reais, Figura 31b Listas 100% e Figura 31c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados.

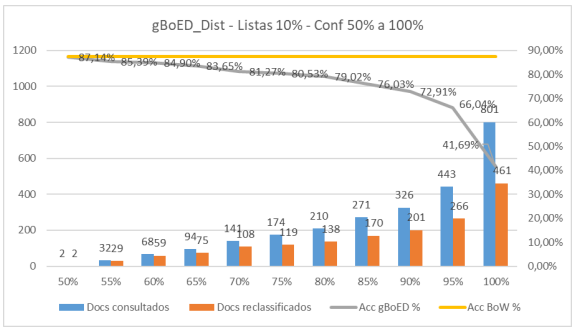
Figura 31 – Uso da gBoED_Dist nos cenários de *SemEval 2015*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Especificando os domínios da coleção *SemEval 2015*, o próximo conjunto de resultados a serem analisados são referentes à coleção de documentos *SemEval 2015 Hotel*. Na Tabela 23 são apresentados os valores de completude e representatividade para este conjunto de documentos. Como esperado, semelhante à coleção de documentos *SemEval 2014 Laptop*, é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação. Na representação formada pelas Listas 100%, a quantidade de acertos na reclassificação está entre 82,75% para gBoED_-Freq e gBoED_Dist. A taxa de erro está em 13,79% e a taxa de Neutros fica em 3,44% para ambas as representações. É importante observar que essa é a coleção de documentos cujas representações semanticamente enriquecidas melhor representou o conjunto, apesar do conjunto SemEval 2015 ter a pior representação de todas as coleções. Porém, a coleção é considerada muito pequena em comparação com as coleções utilizadas em experimentos da literatura, o que faz com que um conjunto pequeno de termos possa representar bem o conjunto de documentos.

Na Tabela 24 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Hotel*. O melhor resultado está destacado pela linha cinza e se dá pelo algoritmo ***Multinomial Naïve Bayes* – MNB, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 86,666%, Medida-F1 de 83,6666%, tanto para gBoED_Freq quanto para gBoED_Dist como enriquecimento semântico**, $\alpha = 1$, grau de confiança de 85%, com 14

Tabela 23 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Hotel*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	29	100,00%	28	96,55%	29	100,00%	28	96,55%
Qtde de documentos sem representação	0	00,00%	1	3,44%	0	0,00%	1	3,44%
Número de ACERTOS na predição	24	82,75%	22	75,86%	24	82,75%	23	79,31%
Número de ERROS na predição	4	13,79%	5	17,24%	4	13,79%	4	13,79%
Número de NEUTROS na predição	1	3,44%	2	6,89%	1	3,44%	2	6,89%

Fonte: Elaborada pelo autor.

documentos sendo consultados e 5 documentos reclassificados. Outros destaques são os modelos formados pelos algoritmos KNN-Cosseno, KNN-Euclidiana, SVM-Polinomial, SVM-Linear, SVM-Polinomial e SVM-RBF que apesar de não obter os melhores resultados, em Listas 100%, atingiram melhores acurácias em relação aos respectivos modelos BoW.

Tabela 24 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Hotel*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,83333	1,00000	100%	0,83333 29-8	100%	0,76667 29-10	100%	1,00000	100%	0,83333 29-8	100%	0,80000 29-9	100%
C4.5-Gini	0,83333	1,00000	100%	0,83333 29-8	100%	0,76667 29-10	100%	1,00000	100%	0,83333 29-8	100%	0,80000 29-9	100%
KNN-Cosseno	0,80000	0,96667 0,90000	90% 75%	0,86667 15-5 $n=7$	90%	0,80000 11-4 $n=3$	70%	0,96667 0,90000	90% 75%	0,86667 15-4 $n=7$	90%	0,80000 11-4 $n=3$	70%
KNN-Euclideana	0,80000	0,96667 0,90000	90% 75%	0,86667 15-5 $n=7$	90%	0,80000 11-4 $n=3$	70%	0,96667 0,90000	90% 75%	0,86667 15-4 $n=7$	90%	0,80000 11-4 $n=3$	70%
MNB	0,73333	1,00000 0,80000	85% 75%	0,86667 14-5 $\alpha=1$	85%	0,80000 8-3 $\alpha=1$	80%	1,00000 0,80000	85% 75%	0,86667 14-4 $\alpha=1$	85%	0,80000 8-3 $\alpha=1$	80%
SVM-Linear	0,73333	0,93333 0,93333	70% 75%	0,83333 8-4 $\gamma=10^{-4}$ a 10^4	70%	0,80000 8-2 $\gamma=10^{-4}$ a 10^4	70%	0,93333 0,93333	70% 75%	0,83333 8-4 $\gamma=10^{-4}$ a 10^4	70%	0,80000 8-2 $\gamma=10^{-4}$ a 10^4	70%
SVM-Polinomial	0,73333	0,90000 0,93333	70% 75%	0,83333 6-3 $\gamma=10^{-1}$	70%	0,80000 6-2 $\gamma=10^{-1}$	70%	0,90000 0,93333	70% 75%	0,83333 6-3 $\gamma=10^{-1}$	70%	0,80000 6-2 $\gamma=10^{-1}$	70%
SVM-RBF	0,73333	0,90000	75%	0,83333 21-3 $\gamma=1$	75%	0,76667 21-2 $\gamma=1$	60%	0,90000	75%	0,87142 2-2 $\gamma=10^{-1}$	75%	0,76667 21-2 $\gamma=1$	60%

Fonte: Elaborada pelo autor.

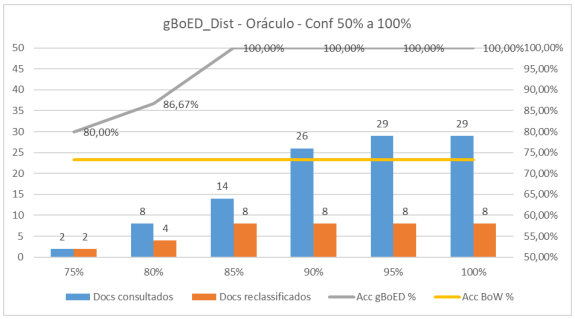
Na Figura 30, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, SVM-RBF), a quan-

tidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-RBF, $\gamma = 1$. No cenário ideal, apresentado na Figura 30a, é possível verificar no comportamento do método que, conforme o grau de confiança aumenta a quantidade de documentos consultados também aumenta. A quantidade de documentos que sofre reclassificação aumenta em uma proporção menor e a acurácia final do classificador cresce até o ao nível máximo de 100%. Nos cenários reais, Figura 30b Listas 100% e Figura 30c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Essa característica acontece pois, apesar do alto nível de representatividade por parte das representações semanticamente enriquecidas, ambas não atingiram mais do que 65% de assertividade com relação à classe do documento, pois elas acabam trocando respostas corretas do modelo BoW (aqueles que possui alto valor de confiança) por respostas incorretas geradas pela gBoed. Para a coleção de documentos *SemEval 2014 Laptop*, o melhor custo x benefício ao utilizar a representação gBoED ocorre quando a predição no modelo BoW está em 50%, em seguida a acurácia começa a cair.

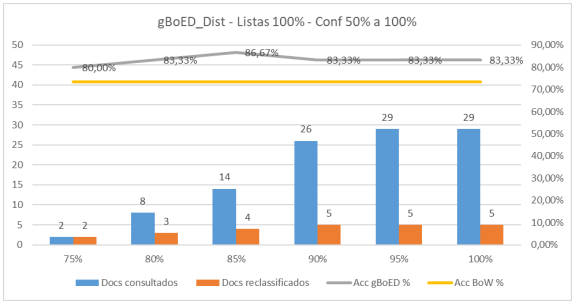
Na Figura 32, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Dist (no cenário com melhor acurácia, MNB), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método MNB, $\alpha = 1$. Verifica-se que no cenário ideal Figura 32a que conforme o grau de confiança aumenta, a quantidade de documentos consultados aumenta proporcionalmente. A quantidade de documentos que sofre reclassificação também aumenta, porém em uma proporção menor. A acurácia final do classificador também aumenta chegando ao nível máximo de 100%. Nos cenários reais, Figura 32b Listas 100% e Figura 32c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados. Para a coleção de documentos *SemEval 2015 Hotel*, o melhor custo x benefício ao utilizar a representação gBoED ocorre em todos os níveis de confiança.

Ainda especificando os domínios da coleção *SemEval 2015*, o próximo conjunto de resultados a serem analisados são referentes à coleção de documentos *SemEval 2015 Laptop*. Na Tabela 25 é apresentada a representatividade para este conjunto de documentos. Verifica-se que para essa coleção de documentos as listas de termos, e consequentemente as representações gBoED_Freq e gBoED_Dist não representaram bem a coleção. É possível observar nos dados da tabela que apenas 21,02% dos documentos foram representados e uma taxa de 13,55% de acertos para gBoED_Freq e 14,95% para gBoED_Dist.

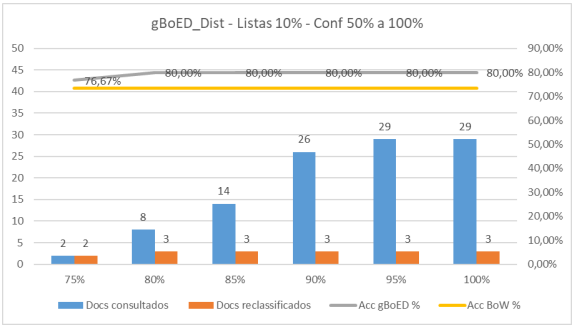
Figura 32 – Uso da gBoED_Dist nos cenários de *SemEval 2015 Hotel*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

Tabela 25 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Laptop*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	90	21,02%	67	15,65%	90	21,02%	67	15,65%
Qtde de documentos sem representação	338	78,97%	361	84,34%	338	78,97%	361	84,34%
Número de ACERTOS na predição	58	13,55%	44	10,28%	64	14,95%	47	10,98%
Número de ERROS na predição	23	5,37%	19	4,43%	17	3,97%	16	3,73%
Número de NEUTROS na predição	347	81,07%	365	85,28%	347	81,07%	365	85,28%

Fonte: Elaborada pelo autor.

Na Tabela 26 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Laptop*. Diferentemente da coleção *SemEval 2015 Hotel*, mas de forma semelhante à coleção de documentos *SemEval 2015*, tanto gBoED_Freq quanto gBoED_Dist não foram capazes de representar os documentos e melhorar os resultados de classificação. Os melhores resultados se deram pelos modelos formados pela BoW, com

destaque para ***Support Vector Machine* – SVM, kernel Polinomial**. Nesse cenário, a **acurácia obtida foi de 80,020%, Medida-F1 de 88,882%**.

Tabela 26 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Laptop*.

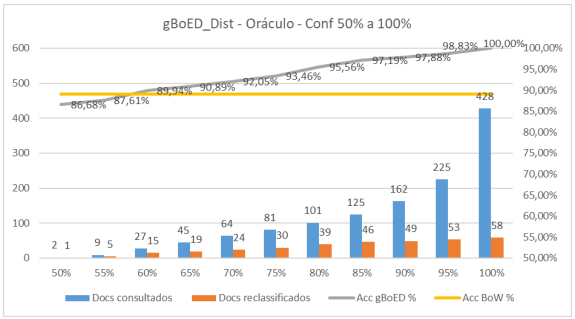
Algoritmos	BoW	gBoED_Freq						gBoED_Dist					
		Oráculo		Listas 100%		Listas 10%		Oráculo		Listas 100%		Listas 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75238	0,78959 0,88344	55% 75%	0,71285 41-37	55%	0,71052 41-38	55%	0,78959 0,88344	55% 75%	0,71285 41-37	55%	0,71052 41-38	55%
C4.5-Gini	0,75476	0,76871 0,84352	55% 75%	0,75011 12-11	55%	0,74779 12-11	55%	0,76871 0,84352	55% 75%	0,75011 12-11	55%	0,74779 12-11	55%
KNN-Cosseno	0,85753	0,87159 0,94142	55% 75%	0,81085 31-26 $n = 17$	55%	0,80853 31-26 $n = 17$	55%	0,87159 0,94142	55% 75%	0,81085 31-26 $n = 17$	55%	0,80853 31-26 $n = 17$	55%
KNN-Euclideana	0,85753	0,87159 0,94142	55% 75%	0,80847 32-27 $n = 17$	55%	0,80615 32-27 $n = 17$	55%	0,87159 0,94142	55% 75%	0,80847 32-27 $n = 17$	55%	0,80615 32-27 $n = 17$	55%
MNB	0,87625	0,89485 0,92287	55% 75%	0,85997 15-14 $\alpha = 10^{-1}$	55%	0,85997 15-14 $\alpha = 10^{-1}$	55%	0,89485 0,92287	55% 75%	0,86229 15-14 $\alpha = 10^{-1}$	55%	0,86229 15-14 $\alpha = 10^{-1}$	55%
SVM-Linear	0,86916	0,88079 0,93466	50% 75%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^4	50%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^4	50%	0,88079 0,93466	50% 75%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^4	50%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^4	70%
SVM-Polinomial	0,89020	0,91345 0,99053	50% 75%	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88555 2-2 $\gamma = 10^{-2}$	50%	0,91345 0,99053	50% 75%	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88555 3-2 $\gamma = 10^{-2}$	55%
SVM-RBF	0,88311	0,89944 0,95565	50% 75%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,89944 0,95565	50% 75%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

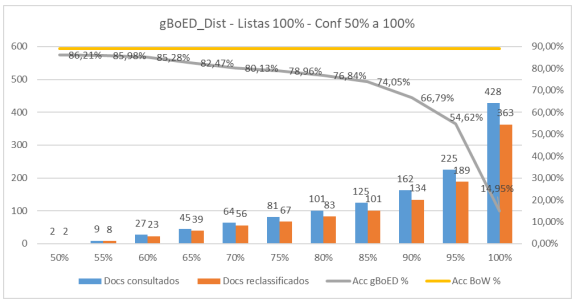
Mesmo não obtendo o melhores resultados, na Figura 33 serão apresentados os gráficos comparativos que mostram o custo e relação à quantidade de consultas à gBoED. A figura contém os dados relacionados à quantidade de documentos que necessitaram consulta à gBoED_Dist algoritmo SVM-Polinomial, nível de confiança 50%, $\gamma = 10^{-1}$. Foi escolhido esse cenário pois apresentou melhor resultado usando BoW. Verifica-se que no cenário ideal Figura 33a que conforme o grau de confiança aumenta, a quantidade de documentos consultados aumenta proporcionalmente. A quantidade de documentos que sofre reclassificação também aumenta, porém em uma proporção menor. A acurácia final do classificador também aumenta chegando ao nível máximo de 100%. Nos cenários reais, Figura 33b Listas 100% e Figura 33c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo, porém a acurácia do modelo decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados.

O último resultado a ser apresentado é referente à coleção *SemEval 2015 Restaurant*, parte da coleção *SemEval 2015*. Na Tabela 27 é apresentada a representatividade para este conjunto de documentos. Como esperado, semelhante à coleção de documentos *SemEval 2015 Hotel*, nela é possível observar que tanto gBoED_Freq quanto gBoED_Dist são capazes de representar a mesma quantidade de documentos e a representação formada pelas Listas 100% é maior do que aquela formada pelas Listas 10%, bem como a quantidade de acertos na reclassificação, um total

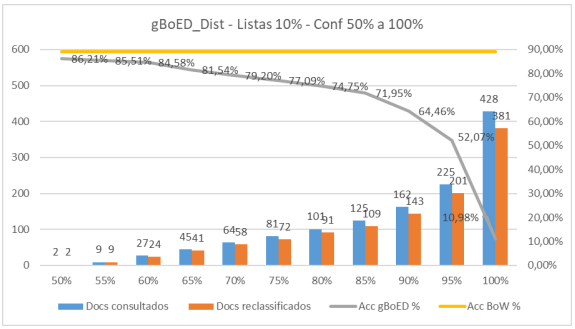
Figura 33 – Custo de uso da gBoED_Dist nos diferentes cenários de *SemEval 2015 Laptop*.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist Listas 100%



(c) gBoED_Dist Listas 10%

Fonte: Elaborada pelo autor.

de 95,34%. Na representação formada pelas Listas 100%, a taxa de acertos na reclassificação está entre 78,19% ambas. A taxa de erro é baixa, ficando em 13,08$, e a taxa de neutros fica em 8,72%.

Tabela 27 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Restaurant*.

	gBoED_Freq				**gBoED_Dist**			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	328	95,34%	312	90,69%	328	95,34%	312	9,30%
Qtde de documentos sem representação	16	4,65%	32	9,30%	16	4,65%	643	38,54%
Número de ACERTOS na predição	269	78,19%	247	71,80%	269	78,19%	255	74,12%
Número de ERROS na predição	45	13,08%	51	14,82%	45	13,08%	43	12,50%
Número de NEUTROS na predição	30	8,72%	46	13,37%	30	8,72%	46	13,37%

Fonte: Elaborada pelo autor.

Na Tabela 28 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Restaurant*. O melhor resultado está destacado pela

linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel Linear, com Listas 100%**. Nesse cenário, a **acurácia obtida foi de 89,513%, Medida-F1 de 89,418%**, utilizando as **representações gBoED_Freq e gBoED_Dist como enriquecimento semântico**, $\gamma = 10^{-4}$ a 10^4, grau de confiança de 70%, com 50 documentos sendo consultados e 15 documentos reclassificados. Todos os outros algoritmos, apesar de não obterem o melhor resultado, em Listas 100%, atingiram melhor acurácia em relação ao modelo BoW. Comparativamente, em Listas 100%, em 125 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq, que foi maior em 105 casos. Apesar do melhor resultado ter sido usando gBoED_Freq, para esse conjunto de documentos o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Tabela 28 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Restaurant*.

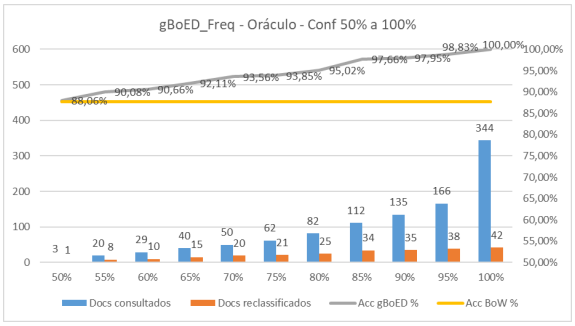
Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,81723	0,83731	85%	0,81403 30-8	85%	0,81118 30-7	55%	0,83731	85%	0,81983 30-7	85%	0,81697 30-6	60%
C4.5-Gini	0,77328	1,0000	100%	0,78168 344-96	100%	0,72672 272-46	95%	1,0000	100%	0,78176 344-97	100%	0,74126 272-46	100%
KNN-Cosseno	0,81378	0,92143 0,95639	65% 75%	0,84303 81-29 $n = 11$	65%	0,81378 28-16 $n = 13$	55%	0,92143 0,95639	65% 75%	0,85168 81-30 $n = 11$	65%	0,82563 96-38 $n = 5$	60%
KNN-Euclideana	0,81378	0,92143 0,95639	65% 75%	0,84303 81-29 $n = 11$	65%	0,81378 28-16 $n = 13$	55%	0,92143 0,95639	65% 75%	0,85168 81-30 $n = 11$	65%	0,82563 96-38 $n = 5$	55%
MNB	0,86908	0,90975 0,93294	65% 75%	0,88084 34-12 $\alpha = 10^{-2}$	65%	0,87773 11-5 $\alpha = 10^{-2}$	55%	0,90681 0,94160	65% 75%	0,88076 34-10 $\alpha = 10^{-1}$	65%	0,87773 11-5 $\alpha = 10^{-2}$	55%
SVM-Linear	0,87765	0,93563 0,93849	70% 75%	0,89513 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,88933 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,93563 0,93849	70% 75%	0,89513 50-14 $\gamma = 10^{-4}$ a 10^4	70%	0,86398 50-14 $\gamma = 10^{-4}$ a 10^4	70%
SVM-Polinomial	0,88059	0,93563 0,94143	70% 75%	0,89504 50-13 $\gamma = 10^{-1}$	70%	0,88639 50-13 $\gamma = 10^{-1}$	70%	0,93563 0,94143	70% 75%	0,89218 50-10 $\gamma = 10^{-1}$	70%	0,89210 50-10 $\gamma = 10^{-1}$	55%
SVM-RBF	0,82261	1,00000 0,97059	90% 75%	0,83983 139-29 $\gamma = 1$	90%	0,82563 27-3 $\gamma = 1$	60%	1,00000 0,97059	90% 75%	0,84588 166-30 $\gamma = 1$	95%	0,82857 27-4 $\gamma = 1$	60%

Fonte: Elaborada pelo autor.

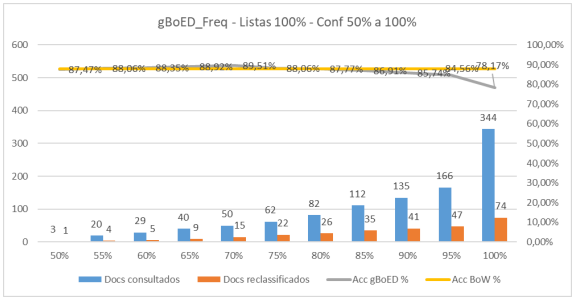
Na Figura 34, observam-se os gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Freq (no cenário com melhor acurácia, SVM-Linear), a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do método SVM-Linear, $\gamma = 10^{-4} a 10^4$. Verifica-se que no cenário ideal Figura 34a que conforme o grau de confiança aumenta, a quantidade de documentos consultados aumenta proporcionalmente. A quantidade de documentos que sofre reclassificação também aumenta, porém em uma proporção menor. A acurácia final do classificador também aumenta chegando ao nível máximo de 100%. Nos cenários reais, Figura 34b Listas 100% e Figura 34c Listas 10%, a relação entre documentos consultados e documentos reclassificados mantêm uma proporção semelhante ao cenário Oráculo. A acurácia sobre cerca de 2%, entre os

graus de confiança 50% e 75%, depois decresce à medida em que aumenta o grau de confiança e mais documentos são consultados e reclassificados.

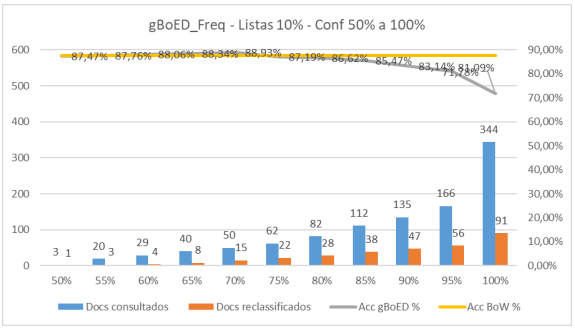
Figura 34 – Uso da gBoED_Freq nos cenários de *SemEval 2015 Restaurant*.



(a) gBoED_Freq Oráculo



(b) gBoED_Freq Listas 100%



(c) gBoED_Freq Listas 10%

Fonte: Elaborada pelo autor.

3.5 Considerações finais

Neste capítulo foi apresentado o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio. Como base do método de classificação proposto, foi realizado o desenvolvimento da generalização da representação semanticamente enriquecida *generalized Bag of Expressions of Domain (gBoED)* apresentada na Subseção 2.2.3.

Duas versões da *gBoED* foram criadas: inicialmente a gBoED_Freq, cuja métrica que associa uma expressão do domínio ao seu documento é a frequência daquela expressão naquele documento e a gBoED_Freq cuja métrica que associa uma expressão do domínio ao seu documento é a distância entre um termo do domínio e um identificador de classe. Inicialmente, como forma de avaliar a viabilidade do uso da representação semanticamente enriquecida em tarefas de classificação, foi realizada uma avaliação experimental utilizando coleções de documentos em diferentes configurações, bem como treinamento de modelos de classificação com diferentes tipos de algoritmos.

Em seguida, foi apresentado o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, que dá nome a este capítulo. Esse método foi desenvolvido com

o objetivo de melhorar resultados de classificação em cenários em que somente o conjunto de palavras não separa bem os documentos em cada classe. Nesses casos são necessários informações sobre o significado das palavras.

O método realiza a predição de documentos por meio um modelo treinado a partir da representação tradicional *Bag of Words (BoW)* e, a partir de um grau de confiança das predições, realiza consultas à representação gBoED e substitui as predições cujo resultado é diferente. Para avaliar esse método foram utilizados um total de 10 coleções de documentos, variando idioma, domínio e tipo de classificação. De todas as coleções utilizadas, apenas em 2 delas não foi possível obter resultados melhores que aqueles obtidos apenas usando BoW. O principal motivo deste fato se dá pelo conjunto de termos não ter tido boa representatividade com relação à coleção. Em todas as outras coleções foram os resultados obtidos pela BoW foram melhorados com a utilização do método de classificação semanticamente enriquecido e pelas representações gBoED_Freq e gBoED_Dist de cada coleção.

De maneira geral, o destaque vai para os modelos gerados utilizando o algoritmos SVM, que atingiu os melhores resultados em 7 das melhores acurácias obtidas. Dos modelos gerados pelas representações semanticamente enriquecidas, a acurácia na representação gBoED_Freq foi maior que a gBoED_Freq em 3941 casos. Isso significa que o esquema de ponderação baseado na distância entre os termos apresenta um impacto positivo na efetividade da gBoED.

Nesse capítulo, foi apresentado que usando Método de Classificação Semanticamente Enriquecida por Expressões do Domínio a partir de listas de termos que representem bem os dados que estão sendo classificados, é possível atingir melhores resultados do que modelos de classificação usando apenas o método tradicional BoW. Esse resultado está relacionado à questão de pesquisa **Q1** e aos objetivos específicos **O1** e **O2**.

CAPÍTULO

4

EXTRAÇÃO DE TERMOS E CLASSIFICAÇÃO SEMÂNTICA BASEADOS EM REGRAS MORFOSSINTÁTICAS

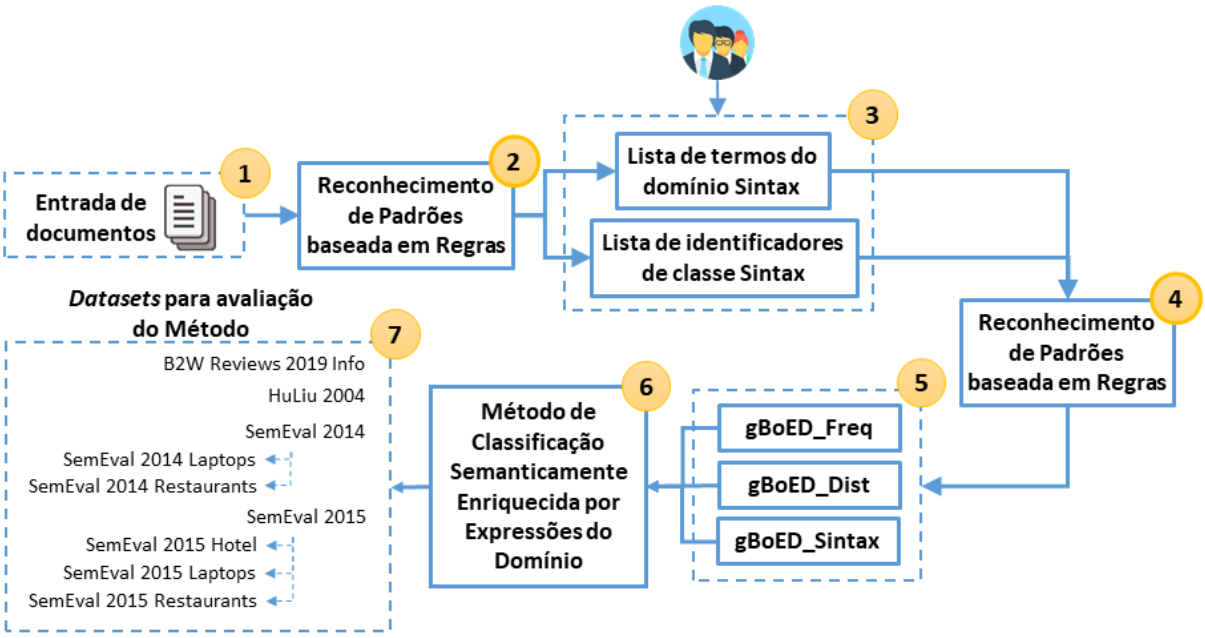
4.1 Considerações iniciais

Na Seção 2.3, "Extração de termos", foram descritos as três principais abordagens para extração de termos: (i) Abordagem Estatística, (ii) Abordagem Linguística e (iii) Abordagem Híbrida. A **Abordagem Estatística** caracteriza-se pela seleção e extração dos termos a partir de da aplicação de medidas estatísticas e de estruturas como N-Gramas. Na **Abordagem Linguística**, termos são identificados e extraídos de um *corpus* com base em suas características ou propriedades linguísticas, as quais podem ser de diferentes tipos ou níveis, morfológico, sintático, semântico e pragmático. Já a **Abordagem Híbrida** utiliza ambas as técnicas para indentificar e extrair candidatos a termos.

Nesta tese de doutorado, dentre os objetivos definidos na Seção 1.3 está o objetivo específico 3, que visa aplicar e analisar o impacto de soluções de extração de termos para a construção de listas de termos do domínio e identificadores de classe. Como apresentado no Capítulo 3, as listas são construídas manualmente pelos especialistas de modo a utilizar o conhecimento do especialista como enriquecimento das representações. Este é um processo que requer um grande esforço por parte do especialista. Portanto, a aplicação de técnicas que tornem o processo de construção das listas e das representações enriquecidas por expressões do domínio mais automatizado torna-se interessante para a diminuição do custo total de uso das informações enriquecidas e do método de classificação enriquecida. Nesse capítulo é apresentado o desenvolvimento, aplicação e validação do método de extração de termos e construção da representação semanticamente enriquecida (gBoED) que utilizada abordagem linguística baseada em regras que identificam padrões sintáticos.

O desenvolvimento e aplicação de regras que identificam padrões sintáticos é uma tarefa que possui muita dependência do domínio e do idioma de cada conjunto de dados. Portanto, neste capítulo para a construção foram escolhidas coleções de documentos da área de análise de sentimentos, nos idiomas português e inglês, para a aplicação das regras de extração de padrões sintáticos. O Capítulo 4 está organizado de acordo com o diagrama Figura 35 que ilustra as 7 principais etapas de desenvolvimento e validação.

Figura 35 – Diagrama das etapas de desenvolvimento do Capítulo 4.



Fonte: Elaborada pelo autor.

Na etapa 1, documentos são introduzidos ao processo e aplicados à etapa 2 de reconhecimento de padrões baseada em regras. As regras construídas para extração de termos são apresentadas na Seção 4.4 e o método de extração de termos baseado em regras é apresentado na Seção 4.5. A etapa 3 ilustra a fase de atuação dos especialistas de domínio que buscam selecionar os termos, realizar um processo de limpeza e organização das listas.

Na etapa 4, as regras para reconhecimento de padrões são aplicadas aos documentos, tendo como base as listas construídas para elaboração das representações gBoED. Na Seção 4.6, uma nova representação é proposta nesse capítulo e representada pela etapa 5, a gBoED_Syntax.

As etapas 6 e 7 fazem parte do processo de validação dos métodos de extração de termos e construção da representação gBoED_Syntax baseado em regras. Apresentado na Subseção 4.6.2, a validação ocorre por meio do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, apresentado no Capítulo 3. O método de classificação faz uso das diferentes representações gBoED construídas a partir das listas de termos extraídas pelo método baseado em regras e aplicadas em 9 coleções de documentos distintas da área de análise de sentimentos.

Os resultados da extração de termos, representatividade e classificação são apresentados com detalhes na Subsubseção 4.6.2.3.

4.2 Trabalhos relacionados

Extração de termos baseada em regras linguísticas e estatísticas são técnicas bastante utilizadas apoiadas pelas ferramentas de anotação morfossintática do tipo *Part-of-Speech (POS)* (JIANHUA *et al.*, 2008). Segundo Pazienza, Pennacchiotti e Zanzotto (2005), mesmo existindo diferentes abordagens para extração automática de termos, essa tarefa consiste em dois passos principais: 1) Identificação de termos candidatos, com suporte em algum tipo de conhecimento linguístico e 2) a seleção e filtro de termos, com base em medidas estatísticas que identificam a importância em relação ao domínio. A seguir, são apresentados alguns trabalhos que utilizam que realizam extração de termos baseada em regras linguísticas.

Em Marciniak e Mykowiecka (2014) é proposto um método para extração automática de termos e expressões da área biomédica, em textos de prontuários médicos que realizam tratamentos de pacientes em ambientes hospitalares. O método utiliza regras linguísticas baseadas em anotações morfossintáticas no idioma polonês, para a identificação das sentenças que possuem os termos candidatos e para a extração desses termos, e medidas estatísticas *C/NC-value*, propostas por Frantzi, Ananiadou e Tsujii (1998), para seleção dos termos. Os termos e expressões extraídos são comparados com os termos presentes na *Polish Medical Subject Headings (MeSH)*, uma base terminológica da área médica em idioma polonês que, segundo os autores, não possui boa cobertura para representar tarefas clínicas. Os resultados apresentaram uma cobertura total de 84% dos termos dos textos de domínio. Ao final do processo de extração, 20% dos termos e expressões extraídos não eram relacionados ao domínio e, dos 80% que eram relacionados ao domínio, 70% não estavam presentes no *Polish MeSH*.

De modo semelhante ao trabalho anterior, em Stanković *et al.* (2016) é desenvolvido um método automático para extração de termos complexos, ou "multi-palavras", baseado em regras gramaticais no idioma sérvio. Também faz uso dos dois passos principais para extração automática de termos: 1) Identificação de termos candidatos e 2) a seleção e filtro de termos. Neste trabalho o método de extração combina extração por regras, dicionários lexicais em um formato de máquina de estados finitos com uma abordagem estatística para filtragem e seleção de termos utilizando as métricas de *Frequência*, *C-Value*, *T-Score*, *Keyness* e uma combinação de medidas na forma $TK\_Value = T\_Score * Keyness$. O método foi aplicado a um *corpus* composto de 51 artigos de periódicos, estes com 32.633 sentenças e 625.105 palavras simples. A avaliação deste trabalho se dá pela qualidade das expressões multi-palavras geradas e se elas realmente eram expressões válidas. Os resultados apresentaram que das multi-palavras formadas, 94% eram válidas e 97% delas realmente existiam no *corpus*.

De forma semelhante, nos trabalhos de Rana e Cheah (2017), Dragoni *et al.* (2018)

e Tian, Cui e Huang (2018) são propostos métodos para extração automática de termos em diferentes domínios: aspectos relacionados à análise de sentimentos em produtos em inglês, termos relacionados documentos legais baseados no código do consumidor na área de telecomunicações australiana, também é inglês, e termos relacionados a um corpus de aviação, respectivamente. Todos eles abordam o processo em dois passos, sendo o primeiro com regras baseadas em estruturas linguísticas, pondendo ser auxiliadas por dicionários ou estruturas como a *WordNet*. No segundo passo do processo são usadas métricas estatísticas como *C-Value*. Os trabalhos foram avaliados em um *corpora* específico para cada domínio, todos eles atingindo resultados interessantes com relação à quantidade e qualidade dos termos.

Por último, Dai e Song (2019) utiliza o processo de extração de termos baseado em linguística textual para extração de aspectos para treinamento de um modelo usando algoritmo neural para classificação de opiniões de produtos em coleções de documentos de análise de sentimentos. A abordagem denominada **RINANTE** - ***R**ule **I**ncorporated **N**eural **A**spect and **O**pinion **T**erm **E**xtraction* é um método que incorpora extração por regras e redes neurais. O método foi avaliado tanto com relação à extração dos aspectos quanto na classificação dos sentimentos usando as coleções SemEval 2014 e 2015. O melhor resultado para SemEval 2014 ficou em 86,76% de acurácia na identificação de aspectos e 86,34% de acuracia na classificação de opiniões. Para SemEval 2015 melhor 69,90% na identificação de aspectos e 72,09% na classificação de opiniões.

Na Seção 4.3 é apresentado o método de extração de termos do domínio e identificadores de classe baseado em regras linguísticas.

4.3 Extração de termos baseada de regras

Como apresentado no Capítulo 3, em cenários de alto nível de complexidade semântica a representação BoW não é suficiente pra representar e construir eficientes modelos de classificação, com isso existe a necessidade de combinação com representações que agreguem informações enriquecidas, como é o caso da gBoED. Para resolver esse problema, nesta tese de doutorado foi elaborado e avaliado o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio que combina o treinamento de modelos tradicionais de classificação com a melhoria de resultados a partir de predições baseadas em expressões do domínio como informações privilegiadas.

No processo de construção de ambas versões da representação semanticamente enriquecida por expressões do domínio, gBoED_Freq e gBoED_Dist, existe a necessidade da construção de listas de termos do domínio e identificadores de classe por especialistas do domínio do problema. Como apresentado na Seção 3.3, termos do domínio são palavras ou expressões, e seus sinônimos, importantes para o domínio de uma determinadas coleção de documentos e possuem relação direta com a organização ou classificação esperada como resultado do processo de

Mineração de Textos. Já os identificadores de classe são palavras ou expressões, juntamente com seus sinônimos, particularmente relaciondas a uma determinada classe e, assim, são consideradas como termos ou palavras-chaves daquela classe.

O processo de construção das listas de termos, por parte dos especialistas de domínio, é realizado de forma manual com o objetivo de transformar em termos e sinônimos o conhecimento sobre o domínio de um problema de modo a transformá-los em modelos de classificação e extração de conhecimentos. A construção das listas de forma manual é um processo que envolve grande esforço por parte dos especialista e a criação de métodos que automatizem esse processo contribui de forma significativa para a melhoria do trabalho dos especialista e, também, para o avanço do estado-da-arte na literatura voltada para enriquecimento semântico.

A extração de termos baseada em regras morfossintáticas faz parte dos métodos de abordagem linguística. A construção de regras nesse tipo de abordagem possui a necessidade de um tipo de anotação textual específica que identifica as classes morfossintáticas de cada palavra. Essas anotações são denominadas, em inglês, de *Part-of-Speech tags (POS)*. Como cada idioma possui uma construção morfossintática particular, tanto ferramentas que realizam esse tipo de anotação quanto as regras de extração, são dependentes do idioma do domínio. Outra dependência que a abordagem linguística também possui é do próprio domínio, do idioma e tipo de classificação que se deseja realizar. Dependendo do domínio e do tipo de classificação, os termos do domínio e os identificadores de classe podem variar entre substantivos, adjetivos ou verbos.

A implementação do extrator de termos baseado em regras de padrões sintáticos foram realizadas em linguagem Python na versão 3.7. As regras foram implementadas por meio de expressões regulares, na linguagem Python. Também foi utilizada a biblioteca NLTK na versão 3.4.5. Ferramentas de anotação morfossintática *Part-of-Speech (POS)* foram utilizadas em ambos os idiomas. Para anotação morfossintática em inglês foi utilizada a ferramenta "*pos_tag*" contida na biblioteca NLTK. Essa ferramenta segue o padrão de anotações *English Penn Treebank tagset* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993), mais detalhes dos rótulos inseridos por ambas as ferramentas podem ser vistos no Anexo A.

No exemplo do Quadro 5, verificam-se dois cenários distintos para a identificação de padrões textuais e a extração de termos. No cenário 1, é apresentado um exemplo de uma opinião de produto no domínio da análise de sentimentos, em inglês, da classe positiva. Nesse cenário é possível verificar uma sentença que foi rotulada usando um anotador *POS*. Nele, DT é um pronome demonstrativo, NN é um substantivo, VBZ é um verbo e JJ é um adjetivo. Considerando o domínio da classificação de sentimentos, verifica-se que uma possível classe morfossintática para termos do domínio são os substantivos e para os identificadores de classe são os adjetivos. No cenário 3, é apresentado um exemplo de sentença, em português, relacionada ao domínio de classificação de vitória ou derrota de brasileiros em esportes. Na anotação do exemplo, NPROP é um nome próprio (ou substantivo próprio), V é um verbo e ART é um artigo. Considerando

o domínio de classificação de vitória ou derrota de brasileiros em esportes, verifica-se que uma possível classe morfossintática para termos do domínio são os nomes próprios e para os identificadores de classe são os verbos.

Quadro 5 – Exemplos de anotações *Part-of-Speech (POS)*.

Cenário 1: Classificação de sentimentos. Idioma inglês.

Sentença original: This camera is perfect!

Anotação *POS*: This_DT camera_NN is_VBZ perfect_JJ

onde, DT é um pronome demonstrativo, NN é um substantivo, VBZ é um verbo e JJ é um adjetivo.

Termo do domínio: camera_NN
Identificador de classe: perfect_JJ

Cenário 2: Classificação de vitória ou derrota de brasileiros em esportes. Idioma português.

Sentença original: Brasil goleia a Grécia.

Anotação *POS*: Brasil_NPROP goleia_V a_ART Grécia_NPROP .

onde, NPROP é um nome próprio (ou substantivo próprio), V é um verbo e ART é um artigo.

Termo do domínio: Brasil_NPROP
Identificador de classe: goleia_V

Fonte: Elaborada pelo autor.

Como visto no exemplo do Quadro 5, a construção das regras baseadas em anotações morfossintáticas *POS*, é dependente do idioma e do domínio. Portanto, como comentado anteriormente, as regras foram desenvolvidas para os idiomas português e inglês, no domínio da análise e classificação de sentimentos. Na Seção 4.4 são apresentados os conjuntos de regras de extração de termos baseadas em padrões sintáticos para os idiomas português e inglês, respectivamente.

4.4 Regras para análise de sentimentos

Dois conjuntos de regras para diferentes domínios e idiomas foram desenvolvidos com o objetivo de extrair os termos que irão formar as expressões do domínio. O primeiro conjunto

de regras foi desenvolvido para o extração de termos, para o idioma português, no domínio da análise de sentimentos em opiniões de produtos e serviços. O segundo conjunto é composto por 7 regras que correspondem aos padrões encontrados em documentos do idioma inglês, também no domínio da analise de sentimentos.

Na descrição das regras alguns símbolos semelhantes àqueles aplicados em expressões regulares serão utilizados na explicação. Na Tabela 29 são apresentados os principais símbolos utilizados na descrição das regras de extração de termos.

Tabela 29 – Símbolos e descrição.

Símbolo	Descrição
( )	Representa um agrupamento de padrões.
|	Representa uma opção. Semelhante a um operador lógico *OU*.
+	Representa um multiplicador do tipo “um ou muitos”.
*	Representa um multiplicador do tipo “zero ou muitos”.
?	Representa um multiplicador do tipo “um ou nenhum”.

Fonte: Elaborada pelo autor.

Como comentado anteriormente, as regras baseiam-se em estruturas sintáticas identificadas por rótulos *POS*, observados com detalhes no Anexo A. Para que os exemplos de cada regra fique claro, na Tabela 30 é apresentado, de forma resumida, o significado de alguns rótulos para as regras em português e na Tabela 31 é apresentado, de forma resumida, o significado de alguns rótulos para as regras em inglês.

Tabela 30 – Rótulos POS e descrição para o idioma português.

Rótulo POS	Descrição		Rótulo POS	Description
N, NPROP	Substantivo		**ADV**	Advérbio
ADJ	Adjetivo		**PREP**	Preposição
V, VAUX	Verbo		**ART**	Artigo

Fonte: Elaborada pelo autor.

Tabela 31 – Rótulos POS e descrição para o idioma inglês.

POS Tags	Description		POS Tags	Description
NN, NNS	Substantivo		**IN**	Preposição/Conjunção
JJ, JJS	Adjetivo		**DT**	Artigo
VB, VBZ, VBD, VBP	Verbo		**PP, PPZ**	Pronome
RB	Advérbio			

Fonte: Elaborada pelo autor.

Tabela 32 – Regras aplicadas para português e inglês.

Regras para português

#	Padrões	Exemplos de sentenças	Sentenças com rótulos POS	Termos do domínio	Identificadores de classe
1	*(Adjective)*	Ex. 1: Insatisfeito ! Ex. 2: Maravilhoso !	Ex. 1: insatisfeito_ADJ Ex. 2: maravilhoso_ADJ	-	*insatisfeito* *maravilhoso*
2	*(Noum) (Verb)? (Adverb)? (Adjective)*	Ex. 1: Produto muito bom ! Ex. 2: Celular é muito bonito. Ex. 3: Produto é excelente. Ex. 4: Notebook é muito rápido.	Ex. 1: produto_N muito_ADV bom_ADJ Ex. 2: celular_N é_V muito_ADV bonito_ADJ Ex. 3: produto_N é_V excelente_ADJ Ex. 4: notebook_N é_V muito_ADV rápido_ADJ	*produto* *celular* *notebook*	*bom* *bonito* *excelente* *rápido*
3	*(Adverb)? (Noum)? Verb (Adverb)?* *Article (Noum)*	Ex. 1: Vale muito a pena. Ex. 2: Não vale a pena.	Ex. 1: vale_V muito_ADV a_ART pena_N Ex. 2: não_ADV vale_V a_ART pena_N	-	*vale_muito_a_pena* *não_vale_a_pena*
4	*(Noum)+ Verb*	Ex. 1: Som distorce. Ex. 2: Caixa desconecta.	Ex. 1: som_N distorce_V Ex. 2: caixa_N desconecta_V	*som* *caixa*	*distorce* *desconecta*
5	*Adjective (Noum)* (Preposition)* Noum*	Ex. 1: Péssimo pós venda. Ex. 2: Péssima experiência. Ex. 3: Excelente atendimento. Ex. 4: Alta capacidade de armazenamento. Ex. 5: Boa usabilidade. Ex. 6: Péssimas funções.	Ex. 1: péssimo_ADJ pós_PRE venda_N Ex. 2: péssima_ADJ experiência_N Ex. 3: excelente_ADJ atendimento_N Ex. 4: alta_ADJ capacidade_N de_PRE armazenamento_N Ex. 5: boa_ADJ usabilidade_N Ex. 6: péssimas_ADJ funções_N	*pós_venda* *experiência* *atendimento* *capacidade_de_* *armazenamento* *usabilidade* *funções*	*péssimo* *péssima* *péssimas* *excelente* *alta* *boa*
6	*(Noum)? Adverb (Verb)+ (Article)? (Noum)?*	Ex. 1: Não foi entregue. Ex. 2: Produto não foi entregue. Ex. 3: Não recebi o produto. Ex. 4: Eles não respondem!	Ex. 1: não_ADV foi_VAUX entregue_V Ex. 2: produto_N não_ADV foi_VAUX entregue_V Ex. 3: não_ADV recebi_V o_ART produto_N Ex. 4: eles_PRO não_ADV respondem_V	*produto*	*não_foi_entregue* *não_recebi* *não_respondem*
7	*(Adverb)? Verb*	Ex. 1: Super recomendo! Ex. 2: Não recomendo. Ex. 3: Comprei e não recebi.	Ex. 1: super_ADV recomendo_V Ex. 2: não_ADV recomendo_V Ex. 3: comprei_V e_KC não_ADV recebi_V	-	*recomendo* *não_recomendo* *não_recebi*

Regras para inglês

#	Padrão	Exemplos de sentenças	Sentenças com rótulos POS	Termos do domínio	Identificadores de classe
1	*((Noun Conjunction Noun) | Noun)* *Verb ((Adverb)* Adjective)*	Ex. 1: Screen is very large. Ex. 2: Computer is very light. Ex. 3: Ease of use was very impressive. Ex. 4: This player is not worth.	Ex. 1: screen_NN is_VBZ very_RB large_JJ Ex. 2: computer_NN is_V very_RB light_JJ Ex. 3: Ease_NN of_IN use_NN was_V very_RB impressive_JJ Ex. 4: This_PP player_NN is_V not_RB worth_JJ	*screen* *computer* *ease_of_use* *player*	*large* *light* *impressive* *not_worth*
2	*(Adverb)? (Adjective) (Noun)+*	Ex. 1: Good phone. Ex. 2: Horrible screen. Ex. 3: Not a good screen ever.	Ex. 1: good_JJ phone_NN Ex. 2: horrible_JJ screen_NN Ex. 3: not_RB a_DT good_JJ screen_NN ever_NN	phone screen	*good* *horrible* *not_good*
3	*(Noum)+ Verb*	Ex. 1: The software sucks. Ex. 2: The player failed repeatedly.	Ex. 1: the_DT software_NN sucks_VBZ Ex. 2: the_DT player_NN failed_VBD repeatedly_RB	*software* *player*	*sucks* *failed*
4	*(Adjective) ((Noun | Preposition)+ Noun)*	Ex. 1: It's the worst piece of electronics.	Ex. 1: it_PP 's_VBZ the_DT worst_JJS piece_NN of_IN eletronics_NNS	*piece_of_electronics*	*worst*
5	*(Pronoun | Adverb)* ((Verb (Adverb)?)?* *(Verb))? (Adjective)?*	Ex. 1: It works great. Ex. 2: I do not recommend.	Ex. 1: it_PP works_VBZ great_JJ Ex. 2: i_PP do_VBP not_RB recommend_VB	*works*	*great* *do_not_recommend*
6	*(Pronoun | Adverb)* ((Verb (Adverb)?)?* *(Verb))? (Adverb)*	Ex. 1: It just works flawlessly. Ex. 2: It doesn't runs perfectly.	Ex. 1: it_PP just_RB works_VBZ flawlessly_RB Ex. 2: it does_VBZ n't_RB runs_VBZ perfectly_RB	*works* *runs*	*flawlessly* *doesn't_perfectly*
7	*(Noun | Adjective | Verb)$*	Ex. 1: What a disappointment. Ex. 2: It's horrible. Ex. 3: Amazing!	Ex. 1: what_WP a_DT disappointment_NN Ex. 2: it_PRP 's_VBZ horrible_JJ Ex. 3: amazing_JJ	-	*disappointment* *horrible* *amazing*

Fonte: Elaborada pelo autor.

Na Tabela 32 são apresentadas as 7 regras para o português e 7 regras para o inglês, respectivamente, responsáveis por reconhecer a estrutura dos padrões e extrair termos para compor as listas. Na coluna “Padrões” são apresentadas as regras escritas de forma semelhante a uma expressão regular. Na coluna “Exemplos de Sentenças” são apresentados exemplos de sentenças que correspondem a cada padrão. Na coluna “POS Tagged Sentences” as sentenças da coluna anterior são apresentadas com os rótulos POS aplicados e os termos que correspondem aos padrões são apresentados em negrito e sublinhado. As colunas “Termos de Domínio” e “Identificadores de Classe” correspondem aos tipos de termos extraídos pela regra em cada sentença.

As regras para reconhecimento de padrões são apresentadas no domínio da análise de sentimentos. Elas são construídas e aplicadas com base em textos pré-processados, porém, para melhor compreensão, nos exemplos apresentados os textos não são pré-processados. Em exemplos de sentenças que não possuem termos de domínio explícitos, após a extração dos termos, são utilizados um termo neutro, como “item” por exemplo, no lugar do termo de domínio e, assim, formar a expressão de domínio.

A partir das regras apresentadas na Tabela 32, na Seção 4.5 é apresentado o Método de extração de termos baseado em regras morfossintáticas desenvolvido com o objetivo de melhorar o trabalho dos especialistas de domínio com a automatização do processo extração de termos.

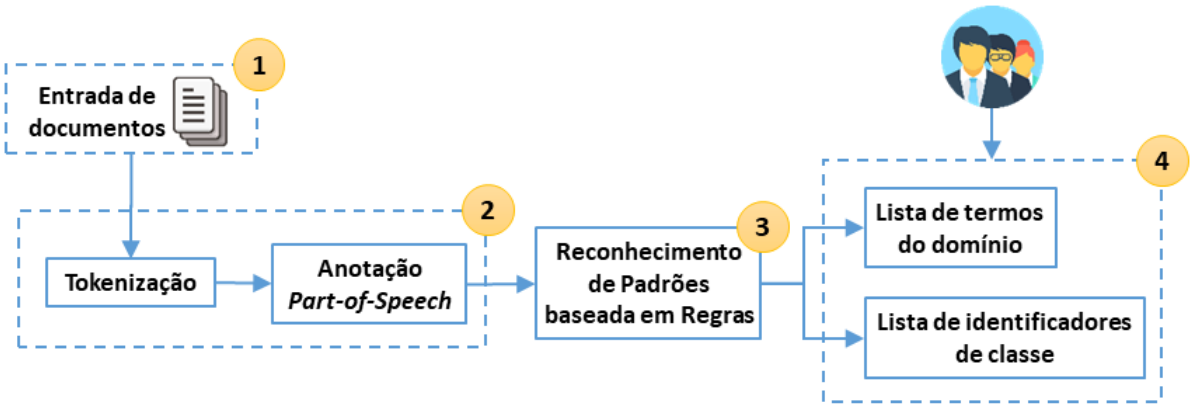
4.5 Método de extração de termos baseado em regras morfossintáticas

O método de extração de termos baseado em regras acontece de maneira simples com a aplicação das regras descritas nas seções anteriores diretamente no conjunto de documentos, visando extrair os termos de domínio e os identificadores de classe. Na Figura 36 é apresentado um diagrama com os passos a serem executados para extração de termos baseado regras morfossintáticas.

Como é possível observar na Figura 36, os documentos são introduzidos e passam por duas técnicas de pré-processamento: tokenização e anotação morfossintática POS. Como apresentada na Subseção 2.1.1, a tokenização é uma técnica de pré-processamento para normalização do texto com o objetivo de quebrar o texto em tokens (símbolos ou palavras). Em seguida, são submetidos às regras de extração de padrões e, por fim, duas listas são geradas: Termos do domínio e Identificadores de classe.

O método como um todo pode ser considerado um processo semiautomático, pois ao final da aplicação das regras de extração de padrões e formação das listas, é necessário um trabalho por parte dos especialistas do domínio para limpar as listas e separar os identificadores em cada tipo de classe.

Figura 36 – Método de extração de termos baseado em regras morfossintáticas.



Fonte: Elaborada pelo autor.

A extração de termos por meio de regras morfossintáticas é aplicada em conjunto com o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. Na Seção 4.6 é apresentada a proposta de uma nova versão do método de classificação enriquecida baseada em regras morfossintáticas. Os resultados serão apresentados em conjunto na Subsubseção 4.6.2.2, aplicados em coleções de documentos referentes a análise de sentimentos em produtos e serviços.

4.6 gBoED_Syntax: gBoED baseada em regras morfossintáticas

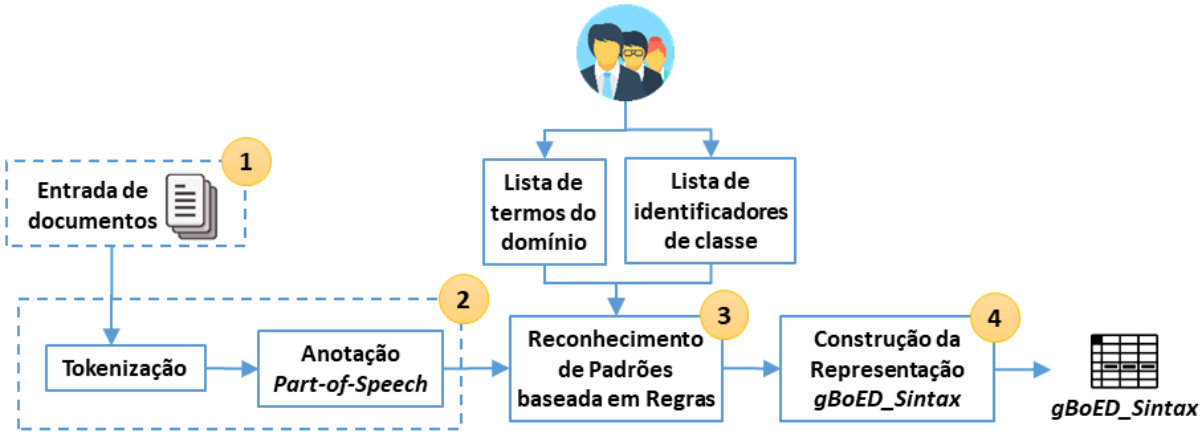
Este método proposto visa a construção de uma nova versão da representação gBoED baseada em padrões morfossintáticos. O objetivo é ampliar ainda mais as opções de representações semanticamente enriquecidas baseadas em expressões do domínio aplicadas ao método de classificação. Outro objetivo é buscar maior qualidade na representação dos documentos, tornando a representanção enriquecida o mais fiel possível ao texto dos documentos. A representação será chamada de "gBoED_Syntax".

4.6.1 Método proposto para construção da representação gBoED_Syntax

Na Figura 37 é apresentado o diagrama do método de construção da representação gBoED baseada em padrões morfossintáticos. O método consiste em 5 etapas principais: 1) Entrada de documentos, 2) Pré-processamento: Tokenização e Anotação morfossintática 3) *Part-of-Speech (POS)* 4) Reconhecimento de padrões por meio de regras e 5) Construção da representação gBoED.

De forma semelhante à extração dos termos, a etapa (1) corresponde à entrada de documentos. Os documentos são submetidos a um pré-processamento composto por tokenização

Figura 37 – Método de construção da representação gBoED baseada em regras morfossintáticas.



Fonte: Elaborada pelo autor.

e anotação morfossintática POS, na etapa (2). Na etapa (3), ao texto tokenizado é aplicada a ferramenta de anotação morfossintática *Part-of-Speech (POS)*. Na etapa (4), os documentos são submetidos ao processo de reconhecimento de padrões por meio de regras, apoiados pelas listas de termos do domínio e identificadores geradas manualmente pelo especialista de domínio ou de forma semiautomática. Os termos encontrados no texto são combinados em expressões do domínio para construção da representação gBoED_Syntax.

Ainda na etapa (5), no processo de construção da gBoED é computada a frequência de cada expressão do domínio. A identificação de negações resulta em expressões do domínio da classe oposta, por exemplo, na sentença “O produto não é bom”, “produto” é um termo do domínio e “bom” é um identificador da classe positiva. Porém, a presença da negação forma uma expressão da classe negativa.

4.6.2 Avaliação experimental do método de construção da representação gBoED_Syntax

Nessa seção é apresentada a avaliação experimental da aplicação da representação gBoED_Syntax no Método de Classificação Semanticamente Enriquecido por Expressões do Domínio. Será analisado o impacto do uso da representação semanticamente enriquecida na melhoria de resultados de classificação de nível semântico. A principal diferença nesse caso são as listas de termos e representação que foram gerados por métodos semiautomáticos.

Como visto anteriormente, a construção de regras baseadas em análise morfossintática possui grande dependência do domínio, idioma e tipo de classificação. Portanto, as coleções de documentos utilizadas na avaliação experimental são as mesmas apresentadas na Subsubseção 4.6.2.1 e serão aplicadas no método de classificação em cenários de análise de sentimentos nos idiomas inglês e português.

4.6.2.1 Coleção de Documentos

As coleções de documentos utilizadas no método de extração de termos aplicado nesse capítulo são as mesmas coleções voltadas para o domínio da análise de sentimentos, tanto em idioma inglês quanto em idioma português, utilizadas nos experimentos do Capítulo 3. A seguir, as coleções serão apresentadas de forma mais sucinta enfatizando características importantes ao método de extração de termos baseado em regras.

B2W Reviews 2019 Info: coleção de documentos composta por 132.374 opiniões de produtos em português separadas em dezenas de categorias (REAL; OSHIRO; MAFRA, 2019). A *B2W Digital* é uma das maiores plataformas de e-commerce da América Latina. Para os experimentos deste trabalho foi selecionada uma categoria relacionada a um único domínio: Informática. Essa categoria é formada por 4262 opiniões de produtos como computadores, notebooks e tablets. A base original possui rotulação por pontos que vão de 1 (ruim) a 5 (excelente). Neste trabalho, por questões de capacidade de processamento, foram selecionadas aleatoriamente 1000 opiniões, 500 opiniões para cada classe. Os rótulos foram modificados para "Positivo" as opiniões pertencentes à pontuação 5 e "Negativo" as opiniões pertencentes às pontuações 1 e 2, de acordo com instruções dos autores.

HuLiu 2004: coleção composta por opiniões de tipos produtos distintos (dois modelos de câmeras digitais, um modelo de telefone celular, um *MP3 Player* e um *DVD player*) (HU; LIU, 2004). A coleção utilizada é um subconjunto da coleção original. Nesta serão considerados 186 opinões positivas e 110 opiniões negativas.

SemEval 2014: coleção composta por opiniões de *laptops* e restaurantes. Inicialmente a coleção foi criada para a competição "*SemEval-2014 Aspect Based Sentiment Analysis task 4*" (PONTIKI *et al.*, 2014). Assim como em HuLiu 2004, nessa coleção será utilizado um subconjunto composto por 1.836 opiniões positivas and 1.073 opiniões negativas, totalizando 2.909 opiniões. Devido a essa coleção possuir dois subdomínios diferentes, serão utilizados divisões da coleção. São elas:

- **SemEval 2014 Laptop**: composta por 619 opiniões positivas e 622 negativas.
- **SemEval 2014 Restaurant**: composta por 1.217 opiniões positivas e 451 negativas.

SemEval 2015: coleção composta por opiniões de hotéis, *laptops* e restaurantes, criada para a competição "*SemEval-2015 Aspect Based Sentiment Analysis task 12*" (PONTIKI *et al.*, 2015). Nessa coleção será utilizado um subconjunto composto por 555 opiniões positivas and 246 opiniões negativas, totalizando 801 opiniões. Assim como SemEval 2014, serão consideradas divisões da coleção por subdomínio. São elas:

- **SemEval 2015 Hotel**: composta por 21 opiniões positivas e 8 negativas.
- **SemEval 2015 Laptop**: composta por 277 opiniões positivas e 151 negativas.

- **SemEval 2015 Restaurant**: composta por 257 opiniões positivas e 87 negativas.

4.6.2.2 *Configuração dos experimentos para validação do método de construção da representação gBoED_Syntax*

O principal foco desta avaliação experimental é analisar o impacto do uso da representação semanticamente enriquecida e gerada por um método semiautomático em comparação com a BoW e os classificadores semanticamente enriquecidos gerados usando listas de termos construídas manualmente. Com isso, a configuração experimental foi realizada de forma semelhante àquela apresentada na Subsubseção 3.4.2.2.

No pré-processamento da etapa 1 do Método de Classificação Semanticamente Enriquecido por Expressões do Domínio, referente à construção da representação *Bag of Words* (BoW) são aplicadas a seguinte sequência de técnicas: 1) Padronização de caixa, 2) Tokenização 3) Remoção de pontuação 4) Remoção de caracteres especiais 5) Remoção de acentos (para as coleções de documentos em português 6) Remoção de números 7) Remoção de *Stopwords* 8) Radicalização.

Como apresentado anteriormente, em coleções de documentos de opiniões de produtos ou serviços cuja classificação está no nível da análise de sentimentos, a negação de aspectos positivos ocorre com bastante frequência. Portanto, na aplicação da técnica de Remoção de *Stopwords* foi desconsiderada a remoção de palavras que trazem tal sentido, como "não" e "nunca" em português e "*not*", "*do not*", "*don't*", "*won't*", "*can't*" , "*can not*", entre outras, em inglês.

Com base nos objetivos desta avaliação experimental, alguns cenários foram idealizados para melhor comparação de resultados:

- **Listas_Orig_100%:** é o mesmo cenário denominado **Listas 100%**, apresentado na Subsubseção 3.4.2.2. É aquele cujas listas de termos foram construídas manualmente pelo especialista a partir de 100% dos documentos da coleção. Ele indica o melhor cenário para o método proposto de modo a comparar seus resultados a um limite superior.

- **Listas_Syntax_100%:** este cenário corresponde às listas de termos do domínio e identificadores de classe gerados a partir de 100% dos textos de cada coleção de documentos, usando o método de extração de termos por regras morfossintáticas.

- **Listas_Orig_10%:** é o mesmo cenário denominado **Listas 10%**, apresentado na Subsubseção 3.4.2.2. Esse cenário possui importância semelhante ao Listas 100%. É aquele cujas listas de termos foram construídas manualmente pelo especialista a partir de apenas 10% dos documentos da coleção. Ele indica o pior cenário para o método proposto de modo a comparar seus resultados a um limite inferior.

Nessa avaliação experimental os cenários de comparação foram definidos com base nas listas de termos geradas a partir de 100% dos documentos de cada coleção, tanto no processo manual, quanto na extração semiautomática por meio de regras. Como é esperado, os processos automatizados tendem a apresentar desempenho menor do que aqueles realizados manualmente. Os cenários construídos a partir de listas formados por 100% dos documentos podem ser considerados como os mais próximos da realidade. Portanto, é a partir deles que serão considerados os principais resultados dos experimentos.

Os algoritmos utilizados são apresentados a seguir, seguidos dos parâmetros de configuração utilizados em cada caso. Nos experimentos executados foi utilizado método de amostragem *10-fold cross validation*. Tais algoritmos e variações são utilizadas na etapa 1 do método.

- **C4.5**, algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor* e critérios para escolha do atributo: Entropia e Gini.
- ***K-nearest neighbor* (KNN)**, algoritmo IBk. Foram utilizadas as opções de voto com peso e voto sem peso. Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: Distância Euclideana e Distância de Cosseno.
- ***Multinomial Naïve Bayes* (MNB)**, algoritmo baseado em Naïve Bayes, com parâmentro alpha(α) considerando os valores: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} e 1.
- ***Suport Vector Machine* (SVM)**, algoritmo *Sequential Minimal Optimization* (SMO). Nesse algoritmo foram consideramos três tipos de kernel: linear, polinomial (expoente=2) e RBF (*Radial Basis Function*). Os valores considerados para cada tipo de kernel foram 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4.

Na avaliação experimental também foram utilizados diferentes valores para o grau de confiança que o método de classificação considera para selecionar os documentos para reclassificação. Nos experimentos realizados o grau de confiança variou entre os valores 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% e 100%.

Nesse experimento foi utilizada a implementação em linguagem Python na versão 3.7, biblioteca Scikit-Learn na versão 0.22.1, NLTK na versão 3.4.5 e Numpy na versão 1.18.1. Na sequência são apresentadas as coleções de documentos utilizadas nessa avaliação experimental.

4.6.2.3 Resultados - validação do método de extração de termos baseado em regras morfossintáticas

Nessa Subsubseção são apresentados os principais resultados obtidos a partir da avaliação experimental do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio Baseada em Regras Morfossintáticas. Os resultados são apresentados individualmente

para cada coleção de textos. Iniciando com as coleções em idioma português e, em seguida, as coleções em inglês, os resultados estão no formato de tabelas contendo as melhores acurácias obtidas pelos modelos gerados por cada algoritmo utilizado.

Nas Tabelas 33, 35, 37, 39, 41, 43, 45, 47 e 49 são apresentadas a representatividade da gBoED nos conjuntos de dados. Nelas são apresentadas a quantidade de documentos que nos quais as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax foram capazes de representar, ou que não puderam representar devido a limitação na abrangência das listas de termos. Além disso, nelas são apresentadas as quantidades de documentos cujas predições a partir de cada uma das representações semanticamente enriquecidas obtiveram acertos e erros. Os documentos que não obtveram representação por parte das gBoEDs ou que obtiveram empate na predição, nestas tabelas estão sendo considerados como neutros e ao serem consultados dentro do método mantém a predição inicial do classificador.

Já nas Tabelas 34, 36, 38, 40, 42, 44, 46, 48 e 50 são apresentados os resultados com as melhores acurácias obtidas em cada algoritmo. Elas apresentam os algoritmos de acordo com um nível decrescente de explicabilidade. Cada tabela de resultados está organizada de maneira que a linha em cinza corresponde ao melhor resultado da gBoED_Syntax obtido para toda a coleção. Na segunda coluna estão as melhores acurácias obtidas a partir dos modelos gerados via BoW. Na sequência, as melhores acurácias obtidas a partir dos modelos gerados pelo método de classificação semanticamente enriquecida incrementado pela representação gBoED_Freq, gBoED_Dist e, por último, as melhores acurácias obtidos dos modelos gerados pelo mesmo método e incrementado pela representação gBoED_Syntax.

Nas tabelas que apresentam as melhores acurácias, os algoritmos estão organizados de acordo com um nível decrescente de explicabilidade. Os resultados obtidos pelo método de classificação enriquecida pela gBoED_Freq, gBoED_Dist e gBoED_Syntax estão divididos em três partes. A primeira coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Orig_100% (aquele cuja gBoED é construída a partir de listas de termos formadas manualmente por 100% dos documentos da coleção). A segunda coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Syntax_100% (aquele cuja gBoED é construída a partir de listas de termos formadas pelo método de extração baseado em regras morfossintáticas em 100% dos documentos da coleção).

A terceira coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Orig_10% (aquele cuja gBoED é construída a partir de listas de termos formadas manualmente por 10% dos documentos da coleção). A leitura da tabela deve ser realizada, primeiramente, observando-se o resultado da melhor acurácia para o cenário Listas_Orig_100%, já que ele corresponde ao cenário mais próximo do real. Em **Acc** é descrita a melhor acurácia e na coluna **Conf** é descrita o grau de confiança em que esse valor de acurácia foi obtido. Em seguida, observa-se o resultado da melhor acurácia para o cenário Listas_Syntax_100% para comparação com o anterior. O cenário Listas_Orig_10% é uma referência de resultado considerando o limite

inferior. Em todos os cenários, logo abaixo do valor da acurácia, estão presentes uma sequência de valores que correspondem à quantidade de documentos enviados para reclassificação devido à confiança ser menor do que o valor descrito, seguido da quantidade desses documentos que foram realmente reclassificados, ou seja, sua predição sofreu alteração de classe.

O primeiro resultado significativo a ser apresentado nesse capítulo está relacionado à representatividade da representação gBoED_Syntax. Nela o conjunto de expressões do domínio que representam cada documento é mais fidedigno do que nas representações gBoED_Freq e gBoED_Dist, ou seja, na construção das representações gBoED_Freq e gBoED_Dist as expressões do domínio são identificadas com base na combinação de todos os termos do domínio com todos os identificadores de classe de uma sentença. Na gBoED_Syntax, as expressões do domínio são formadas a partir das regras baseadas em análise morfossintática que preservam a estrutura do texto. Devido a isso, as expressões do domínio identificadas em cada do documento na gBoED_Syntax são aquelas que realmente estão contidas no texto original, enquanto na gBoED_Freq e gBoED_Dist podem existir expressões que não fazem parte na estrutura semântica original do texto.

No Quadro 6 é apresentado um exemplo de um texto de opinião de serviço de restaurante, em inglês, pertencente à classe *"Positive"*, e quais são as expressões do domínio formadas em cada versão da representação gBoED. As expressões do domínio formadas pelas representações gBoED_Freq e gBoED_Dist são resultados da combinação de todos os termos do domínio e de todos os identificadores de classe em cada sentença. Logo, algumas expressões não correspondem com exatidão a aquilo que está escrito no texto, como é o caso das expressões *fish_tacos_0_cool*, *atmosphere_0_recommendation*, *fire_place_0_recommendation*, *fire_place_0_cool*. Já na representação gBoED_Syntax, como é esperado, são extraídas menos expressões, porém elas são mais fidedignas ao texto e representam com maior precisão aquilo que está escrito. Os números 0 e 1 contidos entre os termos das expressões são referentes às classes positiva (0) e negativa (1)

Iniciando a apresentação dos resultados com a coleção de documentos *B2W Reviews 2019 Info*, são analisados os resultados para a coleção de documentos *B2W Reviews 2019 Info*. Nos gráficos da Figura 38 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 38a pode-se observar as quantidades de termos do domínio, na Figura 38b observa-se as quantidades para identificadores da classe "Positivo" e na Figura 38c observa-se as quantidades para identificadores da classe "Negativo". É possível verificar na Figura 38a que a quantidade de termos do domínio extraídos pelas regras morfossintáticas é similar à quantidade de termos das Listas_Orig_100%. Já nas Figuras 38b e 38c verifica-se que a quantidade de identificadores da classe positiva e negativa é inferior do que aqueles das Listas_Orig_10%.

Na Tabela 33 é apresentada a representatividade para o conjunto de documentos *B2W Reviews 2019 Info*. Nela é possível observar que as Listas_Syntax_100% representam uma quantidade de documentos maior nas três representações comparado às Listas_Orig_100% e,

Quadro 6 – Expressões do domínio formadas na representação gBoED_Syntax.

Documento original:

Great Atmosphere. Went to this place this place on a real cold night and needed a quick meal. I highly recommend the fish tacos, everything else was ok, cool atmosphere, the fire place in the back really ads to it but needs a bit more heat throughout on a cold night.

Classe Original: *Positive*

---

gBoED_Freq e **gBoED_Dist**:

Expressões do Domínio Positivas: *great_0_atmosphere, fish_tacos_0_recommendation, fish_tacos_0_cool, atmosphere_0_cool, atmosphere_0_recommendation, fire_place_0_recommendation, fire_place_0_cool*.

Expressões do Domínio Negativas: *place_1_coldly, meal_1_coldly, fish_tacos_1_coldly, atmosphere_1_coldly, fire_place_1_coldly.*

Predição: *Positive*

---

gBoED_Syntax:

Expressões do Domínio Positivas: *great_0_atmosphere, fish_tacos_0_recommendation, atmosphere_0_cool.*

Expressões do Domínio Negativas: -

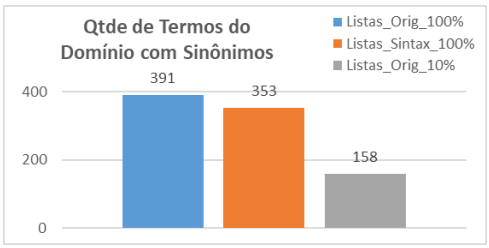
Predição: *Positive*

Fonte: Elaborada pelo autor.

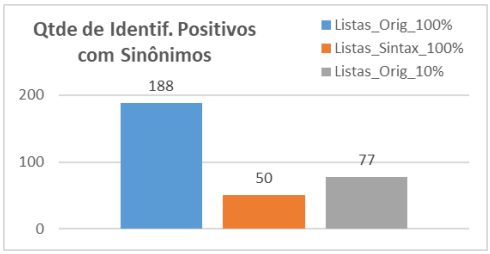
consequentemente, Listas_Orig_10%. Porém, também é possível verificar que mesmo representando mais documentos, a quantidade de acertos na predição usando as Listas_Syntax_100% é menor do que usando as Listas_Orig_100% em todas as representações, manténdo um valor intermediário em cada uma delas. Isso faz com que a relação custo x benefício em todas as representações construídas a partir das Listas_Syntax_100% torne-se interessante.

Na Tabela 34 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *B2W Reviews 2019 Info*. O melhor resultado da gBoED_Syntax e da coleção está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel RBF, com as Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%**. Nesse cenário, a **acurácia obtida foi de 92,200%, Medida-F1 de 92,200%**, $\gamma = 10^{-1}$, grau de confiança de

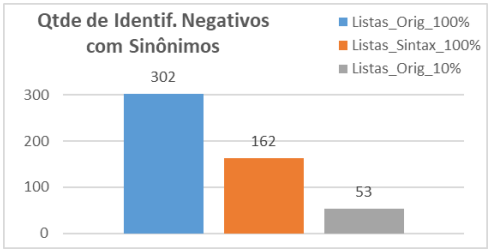
Figura 38 – Gráficos de quantidade de termos por tipo de lista - Coleção *B2W Reviews 2019 Info*.



(a) Termos do Domínio



(b) Identificadores de Classe “Positivo”



(c) Identificadores de Classe “Negativo”

Fonte: Elaborada pelo autor.

Tabela 33 – Representatividade da gBoED no conjunto de dados para *B2W Reviews 2019 Info*.

	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
	Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
Qtde de documentos representados	920	92%	925	92,50%	865	86,50%	920	92%	925	92,50%	865	86,50%	853	85,30%	865	86,50%	791	79,10%
Qtde de documentos sem representação	80	8%	75	7,50%	135	13,50%	80	8%	75	7,50%	135	13,50%	147	14,70%	135	13,50%	209	20,90%
Número de ACERTOS na predição	653	65,30%	637	63,70%	569	56,89%	681	68,10%	677	67,70%	600	60,0%	608	60,80%	602	60,20%	557	55,70%
Número de ERROS na predição	191	19,10%	162	16,20%	216	21,60%	163	16,30%	113	11,30%	185	18,50%	165	16,50%	138	13,80%	221	22,10%
Número de NEUTROS na predição	156	15,60%	201	20,10%	215	21,51%	156	15,60%	210	21%	215	21,50%	227	22,70%	260	26%	222	22,20%

Fonte: Elaborada pelo autor.

50%, com 3 documentos sendo consultados e 3 documentos reclassificados. Nos cenários da representação gBoED_Syntax, Listas_Syntax_100% obteve as melhores acurácias em praticamente todos os algoritmos. De maneira geral, nos cenários das representações gBoED_Freq e gBoED_Dist as melhores acurácias são obtidas Listas_Orig_100%, seguido pelas Listas_Syntax_100% e Listas_Orig_10%. O fato das Listas_Syntax_100% atingirem resultados intermediários em relação as Listas_Orig_100% e Listas_Orig_10%, mostra um custo x benefício interessante em relação ao esforço de construção do conjunto de listas pelo especialista.

Em seguida são analisados os resultados para a coleção de documentos *HuLui 2004*. Nos gráficos da Figura 39 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 39a pode-se observar as quantidades de termos do domínio, na Figura 39b observa-se as quantidades para identificadores da classe “*Positive*” e na Figura 39c observa-se

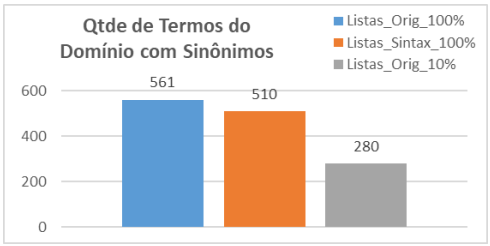
Tabela 34 – Melhores acurácias dos classificadores gerados para *B2W Reviews 2019 Info*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,88600	0,89000 33-14	65%	0,88800 12-6	60%	0,88700 12-6	60%	0,89000 33-14	65%	0,88800 12-6	60%	0,88700 12-6	60%	0,89100 157-55	85%	0,88900 12-7	65%	0,88700 12-5	60%
C4.5-Gini	0,87500	0,87900 175-57	90%	0,87700 1-1	60%	0,87600 1-1	60%	0,87900 175-57	90%	0,87700 1-1	60%	0,87600 1-1	60%	0,88100 225-69	95%	0,87700 1-1	60%	0,87600 1-1	60%
KNN-Cosseno	0,90100	0,90200 66-40 $n = 35$	55%	0,89500 46-32 $n = 35$	55%	0,89300 46-28 $n = 35$	55%	0,90200 66-40 $n = 35$	55%	0,90100 46-26 $n = 35$	55%	0,89300 46-28 $n = 35$	55%	0,90800 128-67 $n = 45$	60%	0,89800 46-27 $n = 35$	55%	0,89700 46-25 $n = 45$	55%
KNN-Euclideana	0,89700	0,89900 67-40 $n = 35$	55%	0,89500 47-33 $n = 35$	55%	0,89000 47-29 $n = 35$	55%	0,89900 67-40 $n = 35$	55%	0,89500 47-33 $n = 35$	55%	0,89000 47-29 $n = 35$	55%	0,90500 70-45 $n = 35$	60%	0,90300 47-36 $n = 35$	55%	0,89400 47-26 $n = 35$	55%
MNB	0,90400	0,89700 50-36 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89200 50-40 $\alpha = 1$	55%	0,89700 50-36 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89200 50-40 $\alpha = 1$	55%	0,90700 123-64 $\alpha = 1$	60%	0,90400 50-42 $\alpha = 1$	55%	0,89700 50-35 $\alpha = 1$	55%
SVM-Linear	0,90900	0,91300 51-33 $\gamma = 10^{-4}$ a 10^4	60%	0,91100 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91300 51-33 $\gamma = 10^{-4}$ a 10^4	60%	0,91100 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91700 73-40 $\gamma = 10^{-4}$ a 10^4	65%	0,91400 73-33 $\gamma = 10^{-4}$ a 10^4	65%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,91500	0,91500 18-14 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91500 18-14 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91900 79-44 $\gamma = 10$	65%	0,91900 79-44 $\gamma = 10$	65%	0,91400 1-1 $\gamma = 10$	50%
SVM-RBF	0,92200	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92200 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%

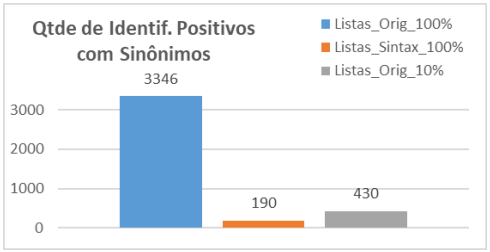
Fonte: Elaborada pelo autor.

as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 39a que a quantidade de termos do domínio extraídos pelas regras morfossintáticas é similar à quantidade de termos das Listas_Orig_100%. Já nas Figuras 39b e 39c verifica-se que a quantidade de identificadores da classe positiva e negativa é menor do que aqueles das Listas_Orig_10%.

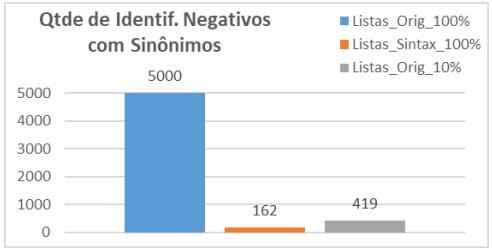
Figura 39 – Gráficos de quantidade de termos por tipo de lista - Coleção *HuLiu 2004*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 35 é apresentada a representatividade para o conjunto de documentos *HuLiu 2004*. Nela é possível observar que as Listas_Syntax_100% representam uma quantidade de documentos maior nas três representações do que Listas_Orig_100% e, consequentemente,

Listas_Orig_10%. Porém, também é possível verificar que mesmo representando mais documentos, a quantidade de acertos na predição usando as Listas_Syntax_100% é menor do que usando as Listas_Orig_100% nas representações gBoED_Freq e gBoED_Dist. Em gBoED_Freq e gBoED_Dist a quantidade de acertos na predição usando as Listas_Syntax_100% também é menor do que usando as Listas_Orig_10%. Em gBoED_Syntax, a quantidade de acertos na predição usando as Listas_Syntax_100% é maior do que usando as Listas_Orig_10%, tornando-se um resultado intermediário e fazendo com que a relação custo x benefício ao utilizar a representação gBoED_Syntax construída a partir das Listas_Syntax_100% torne-se interessante.

Tabela 35 – Representatividade da gBoED no conjunto de dados para *HuLiu 2004*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	285	96,28%	290	97,98%	280	94,60%	285	96,28%	290	97,98%	280	94,60%	272	91,90%	274	92,56%	261	88,18%
Qtde de documentos sem representação	11	3,72%	6	2,02%	16	5,40%	11	3,72%	6	2,02%	16	5,40%	24	8,10%	22	7,44%	35	11,82%
Número de ACERTOS na predição	218	73,64%	211	71,28%	215	72,64%	220	74,32%	213	71,96%	220	74,32%	202	68,24%	197	66,56%	189	63,86%
Número de ERROS na predição	57	19,26%	73	24,66%	54	18,24%	55	18,58%	71	23,98%	49	16,56%	39	13,18%	47	15,88%	46	15,54%
Número de NEUTROS na predição	21	7,10%	12	4,06%	27	9,12%	21	7,10%	12	4,06%	27	9,12%	55	18,58%	52	17,56%	61	20,60%

Fonte: Elaborada pelo autor.

Na Tabela 36 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *HuLiu 2004*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel Polinomial, com Listas_Orig_100%**. Nesse cenário, a **acurácia obtida foi de 84,115%, Medida-F1 de 74,484%**, $\gamma = 10^{-1}$, grau de confiança de 60%, com 32 documentos sendo consultados e 24 documentos reclassificados. O melhor resultado obtido de maneira geral acontece em gBoED_Freq, usando Listas_Orig_100%, com acurácia de 85,816%. Os resultados de gBoED_Syntax usando tanto Listas_Orig_100% quanto Listas_Syntax_100% ainda superam a acurácia obtida pelo classificador gerado pela BoW. Outro resultado de destaque conquistado pelo classificador gerado pelo algoritmo SVM, kernel Polinomial, enriquecido tanto por gBoED_Dist quanto gBoED_Syntax, é que o uso das Listas_Syntax_100% atingiu um resultado intermediário em relação a Listas_Orig_100% e Listas_Orig_10%.

Em seguida são analisados os resultados para a coleção de documentos *SemEval 2014*. Nos gráficos da Figura 40 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 40a pode-se observar as quantidades de termos do domínio, na Figura 40b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 40c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 40a que a quantidade de termos do domínio extraídos pelas regras morfossintáticas é similar à quantidade de termos das Listas_Orig_10%, bem como na Figura 40b com os identificadores da

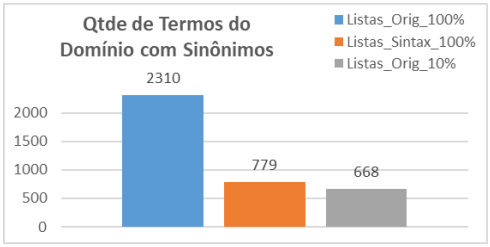
Tabela 36 – Melhores acurácias dos classificadores gerados para *HuLiu 2004*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,68988	0,74057 86-34	90%	0,72264 81-36	80%	0,72678 296-118	100%	0,74390 86-32	90%	0,72276 81-35	80%	0,74345 296-109	100%	0,70218 81-50	85%	0,71253 73-43	75%	0,68207 81-49	85%
C4.5-Gini	0,67609	0,73701 296-114	100%	0,71264 296-96	100%	0,72678 296-111	100%	0,74379 296-118	100%	0,71954 296-97	100%	0,74345 296-106	100%	0,68874 46-23	85%	0,69897 78-39	90%	0,68207 46-23	85%
KNN-Cosseno	0,78712	0,80781 121-57 $n = 9$	70%	0,77701 63-33 $n = 11$	60%	0,79402 63-35 $n = 9$	55%	0,79436 121-55 $n = 9$	70%	0,77713 64-34 $n = 11$	60%	0,79391 63-31 $n = 9$	55%	0,77701 59-37 $n = 13$	60%	0,76356 33-24 $n = 25$	55%	0,77368 59-37 $n = 13$	60%
KNN-Euclideana	0,78712	0,80781 121-57 $n = 9$	70%	0,77701 63-33 $n = 11$	60%	0,79402 63-35 $n = 9$	55%	0,79436 121-55 $n = 9$	70%	0,77713 64-34 $n = 11$	60%	0,79391 63-31 $n = 9$	55%	0,77701 59-37 $n = 13$	60%	0,76356 33-24 $n = 25$	55%	0,77368 59-37 $n = 13$	60%
MNB	0,81402	0,83793 48-29 $\alpha = 10^{-2}$	70%	0,81437 14-6 $\alpha = 10^{-1}$	55%	0,82115 14-5 $\alpha = 10^{-2}$	55%	0,83126 48-31 $\alpha = 10^{-2}$	70%	0,81782 14-5 $\alpha = 10^{-1}$	55%	0,81770 14-8 $\alpha = 10^{-2}$	55%	0,81770 15-9 $\alpha = 10^{-2}$	55%	0,81092 15-8 $\alpha = 10^{-2}$	55%	0,81425 15-10 $\alpha = 10^{-2}$	55%
SVM-Linear	0,82793	0,85161 66-30 $\gamma = 10^{-4}$ a 10^4	70%	0,83414 30-13 $\gamma = 10^{-4}$ a 10^4	60%	0,84149 62-27 $\gamma = 10^{-4}$ a 10^4	70%	0,85149 47-21 $\gamma = 10^{-4}$ a 10^4	65%	0,83414 30-15 $\gamma = 10^{-4}$ a 10^4	60%	0,83804 62-26 $\gamma = 10^{-4}$ a 10^4	70%	0,84103 29-16 $\gamma = 10^{-4}$ a 10^4	60%	0,83414 1-1 $\gamma = 10^{-4}$ a 10^4	55%	0,83414 1-1 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,83436	0,85816 59-21 $\gamma = 10^{-1}$	70%	0,84092 4-3 $\gamma = 10^{-1}$	50%	0,84827 60-17 $\gamma = 10^{-1}$	70%	0,85483 48-21 $\gamma = 10^{-1}$	65%	0,84425 4-2 $\gamma = 10^{-1}$	50%	0,84149 59-20 $\gamma = 10^{-1}$	70%	0,84115 32-18 $\gamma = 10^{-1}$	60%	0,83448 32-24 $\gamma = 10^{-1}$	60%	0,83115 32-22 $\gamma = 10^{-1}$	60%
SVM-RBF	0,81425	0,84149 72-35 $\gamma = 1$	70%	0,82759 36-18 $\gamma = 10^1$	60%	0,83839 69-33 $\gamma = 1$	70%	0,83805 72-29 $\gamma = 1$	70%	0,82747 20-8 $\gamma = 1$	55%	0,82816 69-29 $\gamma = 1$	70%	0,83448 22-15 $\gamma = 1$	55%	0,83448 22-14 $\gamma = 1$	55%	0,82770 22-16 $\gamma = 1$	55%

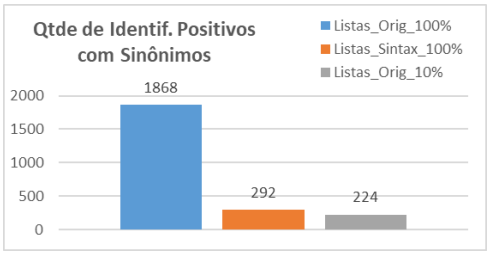
Fonte: Elaborada pelo autor.

classe positiva. Na Figura 40c verifica-se que a quantidade de identificadores da classe negativa é menor do que aqueles das Listas_Orig_10%.

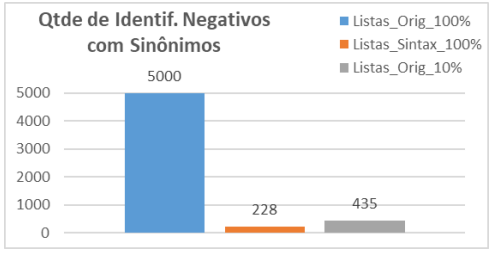
Figura 40 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 37 é apresentada a representatividade para o conjunto de documentos *SemEval 2014*. Nela é possível observar que as representações gBoED_Freq e gBoED_Dist possuem maior quantidade de documentos representados por tipo de lista. As Listas_Orig_100% abrangem uma maior quantidade de documentos e, também, uma maior quantidade de acertos em cada representação. As representações gBoED_Freq e gBoED_Dist construídas a partir de das Listas_Syntax_100% possuem valores intermediários de documentos representados e acertos. A

representação gBoED_Syntax em todas as versões, obtve uma quantidade bastante inferior de documentos representados e de acertos na predição. Nesse caso a representação gBoED_Syntax possui menor representatividade para esta coleção de documentos. Vale lembrar que a coleção *SemEval 2014* é composta por 2 domínios diferentes: *Laptops* e Restaurantes.

Tabela 37 – Representatividade da gBoED no conjunto de dados para *SemEval 2014*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	2186	75,14%	2055	70,64%	1820	62,56%	2186	75,14%	2055	70,64%	1820	62,56%	1134	38,98%	1123	38,60%	861	29,60%
Qtde de documentos sem representação	723	24,86%	854	29,36%	1089	37,44%	723	24,86%	854	29,36%	1089	37,44%	1775	61,02%	1786	61,40%	2048	70,40%
Número de ACERTOS na predição	1719	59,10%	1617	55,58%	1436	49,36%	1830	62,90%	1704	58,58%	1513	52,02%	970	33,34%	971	33,38%	744	25,58%
Número de ERROS na predição	275	9,45%	293	10,08%	247	8,50%	164	5,63%	206	7,08%	170	5,84%	132	4,54%	134	4,60%	100	3,44%
Número de NEUTROS na predição	915	31,45%	999	34,34%	1226	42,14%	915	31,45%	999	34,34%	1226	42,14%	1807	62,12%	1804	62,02%	2065	70,98%

Fonte: Elaborada pelo autor.

Na Tabela 38 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel RBF, com Listas_-Syntax_100%**. Nesse cenário, a **acurácia obtida foi de 80,268%, Medida-F1 de 80,235%**, $\gamma = 1$, grau de confiança de 50%, com 17 documentos sendo consultados e 15 documentos reclassificados. O melhor resultado obtido de maneira geral acontece em gBoED_Dist, usando Listas_Orig_100%, com acurácia de 80,407%. Vale destacar que para os modelos gerados pelos algoritmos C4.5-Gini, KNN-Cosseno, KNN-Euclidiana, MNB, SVM-Linear, SVM-Polinomial e SVM-RBF, enriquecidos por gBoED_Syntax, o uso das Listas_Syntax_100% apresentou os melhores resultados nesses cenários. No enriquecimento pelas representações gBoED_Freq e gBoED_Dist, o uso de Listas_Syntax_100% teve um desempenho intermediário, superando inclusive a acurácia da BoW em SVM-Polinomial e SVM-RBF.

Em seguida são analisados os resultados para a coleção de documentos *SemEval 2014 Laptop*. Nos gráficos da Figura 41 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 41a pode-se observar as quantidades de termos do domínio, na Figura 41b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 41c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 41a que a quantidade de termos do domínio extraídos pelas regras morfossintáticas é similar à quantidade de termos das Listas_Orig_10%, bem como na Figura 41b com os identificadores da classe positiva. Na Figura 41c verifica-se que a quantidade de identificadores da classe negativa é menor do que aqueles das Listas_Orig_10%.

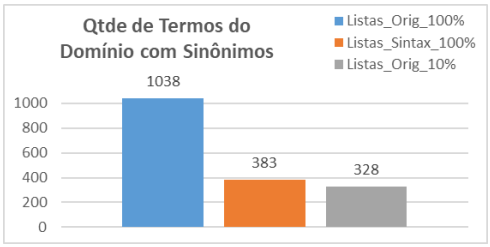
Na Tabela 39 é apresentada a representatividade para o conjunto de documentos *SemEval 2014 Laptop*. Assim como na tabela de representatividade da coleção *SemEval 2014* é possível observar que as representações gBoED_Freq e gBoED_Dist possuem maior quantidade de docu-

Tabela 38 – Melhores acurácias dos classificadores gerados para *SemEval 2014*.

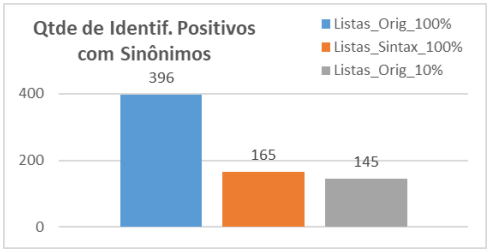
Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,63700	0,63871 2082-1239	55%	0,5995 2088-1162	85%	0,54348 2088-1390	55%	0,66759 2082-1153	55%	0,62048 2088-1115	85%	0,56032 2088-1331	55%	0,63904 1-1	55%	0,45789 2089-1618	85%	0,63904 1-1	50%
C4.5-Gini	0,64525	0,65005 1235-707	55%	0,61978 1432-780	55%	0,57818 27-12	55%	0,66587 1235-654	55%	0,63663 1432-744	55%	0,59090 27-12	55%	0,51459 1432-1142	55%	0,52387 1432-1097	55%	0,46817 1432-1234	55%
KNN-Cosseno	0,77760	0,75973 317-214 $n = 17$	55%	0,75386 340-248 $n = 17$	55%	0,74355 40-29 $n = 17$	55%	0,76419 421-254 $n = 13$	55%	0,7590 340-238 $n = 17$	55%	0,74767 40-29 $n = 17$	55%	0,73015 342-290 $n = 17$	55%	0,73255 340-294 $n = 17$	55%	0,72258 342-306 $n = 17$	55%
KNN-Euclideana	0,77691	0,76076 326-217 $n = 17$	55%	0,75593 346-253 $n = 17$	55%	0,74871 39-28 $n = 17$	55%	0,76523 326-207 $n = 17$	55%	0,75593 346-253 $n = 17$	55%	0,75180 39-28 $n = 17$	55%	0,73668 269-233 $n = 25$	55%	0,73978 264-219 $n = 25$	55%	0,73083 269-242 $n = 25$	55%
MNB	0,80648	0,79823 141-100 $\alpha = 10^{-1}$	55%	0,79409 153-104 $\alpha = 10^{-1}$	55%	0,78584 24-19 $\alpha = 10^{-1}$	55%	0,80063 141-94 $\alpha = 10^{-1}$	55%	0,79546 153-101 $\alpha = 10^{-1}$	55%	0,78687 35-25 $\alpha = 10^{-1}$	55%	0,78068 149-125 $\alpha = 10^{-1}$	55%	0,78343 153-126 $\alpha = 10^{-1}$	55%	0,77793 149-130 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79718	0,79581 25-18 $\gamma = 10^{-4}$ a 10^4	50%	0,79305 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 17-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79615 25-17 $\gamma = 10^{-4}$ a 10^4	50%	0,79305 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 16-8 $\gamma = 10^{-4}$ a 10^4	50%	0,79064 16-16 $\gamma = 10^{-4}$ a 10^4	50%	0,79271 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79030 16-16 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,78824	0,79615 19-13 $\gamma = 10^{-1}$	50%	0,79442 14-11 $\gamma = 10^{-1}$	50%	0,78053 17-14 $\gamma = 10^{-1}$	50%	0,79650 19-12 $\gamma = 10^{-1}$	50%	0,79442 14-11 $\gamma = 10^{-1}$	50%	0,78145 17-14 $\gamma = 10^{-1}$	50%	0,78892 14-13 $\gamma = 10^{-1}$	50%	0,79236 14-12 $\gamma = 10^{-1}$	50%	0,78858 14-14 $\gamma = 10^{-1}$	50%
SVM-RBF	0,79925	0,80235 8-6 $\gamma = 1$	50%	0,80302 17-15 $\gamma = 1$	50%	0,78045 17-14 $\gamma = 1$	50%	0,80407 185-140 $\gamma = 1$	55%	0,80325 17-13 $\gamma = 1$	50%	0,78045 17-14 $\gamma = 1$	50%	0,79683 17-17 $\gamma = 1$	50%	0,80268 17-15 $\gamma = 1$	50%	0,79683 17-17 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

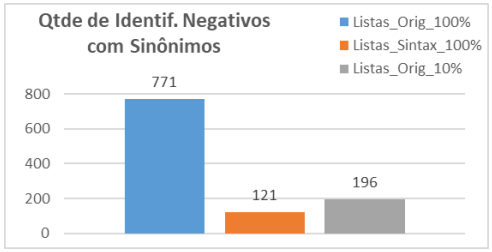
Figura 41 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014 Laptop*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

mentos representados por tipo de lista. As Listas_Orig_100% abrangem uma maior quantidade de documentos e, também, uma maior quantidade de acertos em cada representação, chegando a 61,64% na gBoED_Dist e 57,93% dos documentos na gBoED_Freq. Para as representações gBoED_Freq e gBoED_Dist, as Listas_Syntax_100% possuem valores intermediários de representação e acertos em relação a Listas_Orig_100% e Listas_Orig_10%. Para gBoED_Freq e Listas_Syntax_100% obteve-se 46,09% de acertos e para gBoED_Dist, obteve-se 49,23% de acertos. Verifica-se que gBoED_Syntax possui baixa taxa de representação e acertos em todas as suas versões.

Tabela 39 – Representatividade da gBoED no conjunto de dados para *SemEval 2014 Laptop*.

	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
	Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
Qtde de documentos representados	916	73,82%	752	60,60%	627	50,52%	916	73,82%	752	60,60%	627	50,52%	403	32,48%	403	32,48%	248	19,98%
Qtde de documentos sem representação	325	26,18%	489	39,40%	614	49,48%	325	26,18%	489	39,40%	614	49,48%	838	67,52%	838	67,52%	993	80,02%
Número de ACERTOS na predição	719	57,93%	572	46,10%	485	39,10%	765	61,64%	611	49,24%	503	40,54%	334	26,92%	338	27,24%	202	16,29%
Número de ERROS na predição	120	9,67%	128	10,31%	108	8,70%	74	5,96%	89	7,17%	90	7,25%	57	4,59%	58	4,67%	41	3,30%
Número de NEUTROS na predição	402	32,40%	541	43,59%	648	52,20%	402	32,39%	541	43,59%	648	52,21%	850	68,49%	845	68,09%	998	80,41%

Fonte: Elaborada pelo autor.

Na Tabela 40 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Laptop*. gBoED_Syntax não obteve resultados satisfatórios para essa coleção de documentos, não conseguindo superar os resultados da BoW. Porém, no modelo gerado pelo algoritmo ***Support Vector Machine*, kernel RBF, com Listas_-Syntax_100%**, atingiu resultados superiores a BoW e Listas_Orig_100%. Dentre esses, o melhor resultado obtido foi usando enriquecimento por meio da representação gBoED_Dist, atingindo a **acurácia de 80,334%, Medida-F1 de 80,324%**, $\gamma = 1$, grau de confiança de 50% e com 10 documentos sendo consultados e 9 documentos reclassificados. Quando enriquecido por meio da representação gBoED_Freq, atinge a acurácia de 80,253%, Medida-F1 de 80,224%, $\gamma = 1$, grau de confiança de 50% e com 10 documentos sendo consultados e 10 documentos reclassificados.

Tabela 40 – Melhores acurácias dos classificadores gerados para *SemEval 2014 Laptop*.

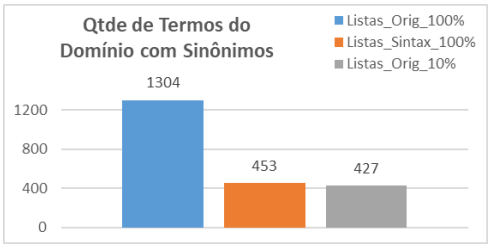
Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,67436	0,65823 174-122	60%	0,64290 174-133	60%	0,62194 174-152	60%	0,66388 174-116	60%	0,65016 174-125	60%	0,62194 174-152	60%	0,61872 174-141	60%	0,61872 174-146	60%	0,59856 174-159	60%
C4.5-Gini	0,68000	0,67355 88-62	60%	0,66548 88-68	60%	0,65500 88-77	60%	0,67677 88-59	60%	0,67032 88-63	60%	0,65500 88-77	60%	0,66467 84-66	60%	0,66145 88-69	60%	0,65258 84-75	60%
KNN-Cosseno	0,77352	0,76869 126-82 $n = 25$	55%	0,75983 126-86 $n = 25$	55%	0,74615 126-100 $n = 25$	55%	0,77352 126-79 $n = 25$	55%	0,76143 126-85 $n = 25$	55%	0,75097 126-98 $n = 25$	55%	0,73890 128-110 $n = 25$	55%	0,74374 126-108 $n = 25$	55%	0,72842 128-119 $n = 25$	55%
KNN-Euclideana	0,77757	0,76790 141-94 $n = 25$	55%	0,75420 141-102 $n = 25$		0,74051 141-115 $n = 25$	55%	0,77353 141-90 $n = 25$	55%	0,75581 141-101 $n = 25$	55%	0,74534 141-113 $n = 25$	55%	0,73408 146-128 $n = 25$	55%	0,73650 141-123 $n = 25$	55%	0,72360 146-136 $n = 25$	55%
MNB	0,79535	0,78727 137-99 $\alpha = 10^{-1}$	60%	0,78243 74-54 $\alpha = 10^{-1}$	55%	0,77519 74-61 $\alpha = 10^{-1}$	55%	0,79291 137-95 $\alpha = 10^{-1}$	60%	0,78727 74-52 $\alpha = 10^{-1}$	55%	0,77519 74-60 $\alpha = 10^{-1}$	55%	0,77277 74-64 $\alpha = 10^{-1}$	55%	0,77518 74-63 $\alpha = 10^{-1}$	55%	0,76712 74-67 $\alpha = 10^{-1}$	55%
SVM-Linear	0,80093	0,79771 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79690 7-7 $\gamma = 10^{-4}$ a 10^4	50%	0,79852 5-5 $\gamma = 10^{-4}$ a 10^4	50%	0,79851 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79690 7-7 $\gamma = 10^{-4}$ a 10^4	50%	0,79932 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79850 9-8 $\gamma = 10^{-4}$ a 10^4	50%	0,79770 5-5 $\gamma = 10^{-4}$ a 10^4	50%	0,79689 9-9 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79610	0,79530 75-59 $\gamma = 10^{-1}$	55%	0,79609 8-7 $\gamma = 10^{-1}$	50%	0,79449 9-9 $\gamma = 10^{-1}$	50%	0,79772 75-54 $\gamma = 10^{-1}$	55%	0,79609 8-7 $\gamma = 10^{-1}$	50%	0,79449 9-9 $\gamma = 10^{-1}$	50%	0,79368 5-5 $\gamma = 10^{-1}$	50%	0,79287 9-8 $\gamma = 10^{-1}$	50%	0,79368 5-5 $\gamma = 10^{-1}$	50%
SVM-RBF	0,80173	0,80173 8-8 $\gamma = 1$	50%	0,80253 10-10 $\gamma = 1$	50%	0,80173 8-6 $\gamma = 1$	50%	0,80254 8-8 $\gamma = 1$	50%	0,80334 10-9 $\gamma = 1$	50%	0,80173 8-6 $\gamma = 1$	50%	0,79850 7-7 $\gamma = 1$	50%	0,80092 8-8 $\gamma = 1$	50%	0,79769 7-7 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

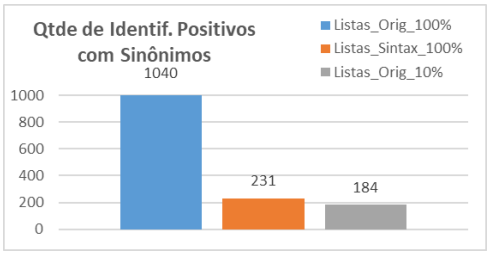
Em seguida são analisados os resultados para a coleção de documentos *SemEval 2014 Restaurant*. Nos gráficos da Figura 42 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 42a pode-se observar as quantidades de termos do domínio, na Figura 42b observa-se as quantidades para identificadores da classe "*Positive*" e

na Figura 42c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 42a que a quantidade de termos do domínio extraídos pelas regras morfossintáticas é similar à quantidade de termos das Listas_Orig_10%, bem como na Figura 42b com os identificadores da classe positiva. Na Figura 42c verifica-se que a quantidade de identificadores da classe negativa é menor do que aqueles das Listas_Orig_10%.

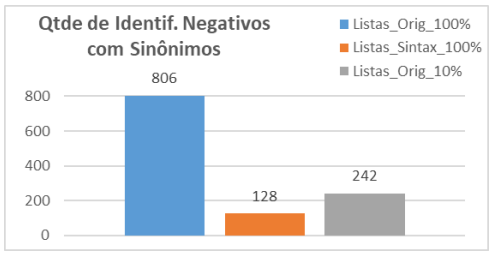
Figura 42 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014 Restaurant*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 41 é apresentada a representatividade para o conjunto de documentos *SemEval 2014 Restaurant*. Nela é possível observar que as representações gBoED_Freq e gBoED_Dist possuem maior quantidade de documentos representados usando Listas_Orig_100% com 77,09%, seguida de Listas_Syntax_100% com 72,12% e, por último Listas_Orig_10%. Com relação à quantidade de acertos na predição, tem-se gBoED_Dist usando Listas_Orig_100% e Listas_Syntax_100%, com 64,80% e 62,94% de acertos respectivamente. Em seguida, gBoED_Freq usando Listas_Orig_100% e Listas_Syntax_100%, com 61,03% e 60,67% de acertos respectivamente. gBoED_Syntax usando Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10% alcançou 38,14%, 36,94% e 28,90% acertos, respectivamente. É importante destacar que nessa coleção de documentos, em todas as representações, Listas_Syntax_100% obteve valores mais baixos de erros na predição. Isso mostra uma melhor representação para essa coleção.

Na Tabela 42 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Restaurant*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel RBF, com Listas_Syntax_100%**. Nesse cenário, a **acurácia obtida foi de 81,657%, Medida-F1 de 81,657%**, $\gamma = 1$, grau de confiança de 50%, com 6 documentos sendo consultados e 6 documentos

Tabela 41 – Representatividade da gBoED no conjunto de dados para *SemEval 2014 Restaurant*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	1286	77,10%	1203	72,13%	1025	61,46%	1286	77,10%	1203	72,13%	1025	61,46%	731	43,82%	688	41,25%	552	33,10%
Qtde de documentos sem representação	382	22,90%	465	27,87%	643	38,54%	382	22,90%	465	27,87%	643	38,54%	937	56,18%	980	58,75	1116	66,90%
Número de ACERTOS na predição	1018	61,04%	1012	60,68%	823	49,35%	1081	64,82%	1050	62,96%	853	51,14%	636	38,14%	616	36,94%	482	28,90%
Número de ERROS na predição	157	9,41%	119	7,13%	140	8,39%	94	5,63%	81	4,85%	110	6,59%	75	4,49%	62	3,71%	62	3,72%
Número de NEUTROS na predição	493	29,55%	537	32,19%	705	42,26%	493	29,55%	537	32,19%	705	42,26%	957	57,37%	990	59,35%	1124	67,38%

Fonte: Elaborada pelo autor.

reclassificados. O melhor resultado obtido de maneira geral acontece tanto em gBoED_Freq quanto gBoED_Dist, usando Listas_Orig_100%, com acurácia de 81,658%, seguido de gBoED_-Freq e gBoED_Dist, usando Listas_Syntax_100%, com acurácia de 81,657%. Vale destacar que para os modelos gerados pelo algoritmo C4.5-Gini, enriquecido por gBoED_Syntax e uso das Listas_Syntax_100% apresentou resultado superior à BoW.

Tabela 42 – Melhores acurácias dos classificadores gerados para *SemEval 2014 Restaurant*.

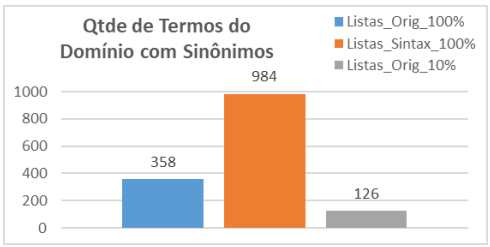
Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**						**gBoED_Syntax**					
		Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75003	0,74940 1-0	50%	0,75000 9-7	65%	0,74940 1-1	50%	0,74940 1-0	50%	0,75000 9-7	65%	0,74940 1-1	50%	0,75120 7-6	60%	0,74940 9-8	65%	0,75180 1-1	55%
C4.5-Gini	0,75543	0,76139 3-3	60%	0,76139 3-3	65%	0,76139 3-3	60%	0,76139 3-3	60%	0,76139 3-3	65%	0,76139 3-3	60%	0,76078 3-3	65%	0,76079 3-3	65%	0,76078 3-3	65%
KNN-Cosseno	0,79196	0,78059 164-108 $n = 13$	55%	0,77580 164-114 $n = 13$	60%	0,76081 164-126 $n = 13$	55%	0,78478 164-101 $n = 13$	55%	0,77819 164-105 $n = 13$	60%	0,76200 164-123 $n = 13$	55%	0,75660 144-124 $n = 15$	55%	0,75600 145-121 $n = 15$	55%	0,74761 144-133 $n = 15$	55%
KNN-Euclideana	0,78959	0,78058 166-110 $n = 13$	55%	0,77579 166-116 $n = 13$	60%	0,76080 166-128 $n = 13$	55%	0,78478 166-103 $n = 13$	55%	0,77819 166-107 $n = 13$	60%	0,76200 166-125 $n = 13$	55%	0,75660 125-106 $n = 19$	55%	0,75660 124-103 $n = 19$	55%	0,74761 146-133 $n = 15$	55%
MNB	0,79856	0,79976 162-108 $\alpha = 10^{-1}$	60%	0,79616 78-62 $\alpha = 10^{-1}$	55%	0,79315 78-65 $\alpha = 10^{-1}$	55%	0,80575 162-95 $\alpha = 10^{-1}$	60%	0,79796 78-57 $\alpha = 10^{-1}$	55%	0,79495 78-62 $\alpha = 10^{-1}$	55%	0,78776 76-63 $\alpha = 10^{-1}$	55%	0,78836 78-69 $\alpha = 10^{-1}$	55%	0,78416 76-69 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79438	0,79438 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79437 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79438 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79437 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79437 8-8 $\gamma = 10^{-4}$ a 10^4	50%	0,79377 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79318 8-8 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79558	0,79558 7-7 $\gamma = 1$	50%	0,79498 7-7 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%	0,79558 7-7 $\gamma = 1$	50%	0,79498 7-7 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%	0,79317 10-8 $\gamma = 1$	50%	0,79558 7-6 $\gamma = 1$	50%	0,79257 10-9 $\gamma = 1$	50%
SVM-RBF	0,81598	0,81658 6-2 $\gamma = 1$	50%	0,81657 6-2 $\gamma = 1$	50%	0,81598 6-3 $\gamma = 1$	50%	0,81658 6-2 $\gamma = 1$	50%	0,81657 6-2 $\gamma = 1$	50%	0,81598 6-3 $\gamma = 1$	50%	0,81538 6-6 $\gamma = 1$	50%	0,81657 6-6 $\gamma = 1$	50%	0,81538 6-6 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

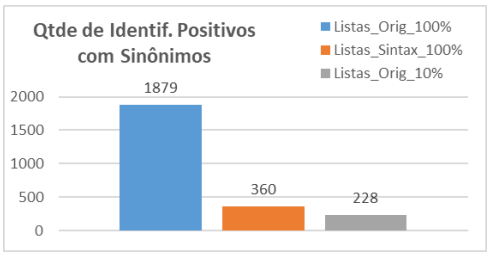
Em seguida são analisados os resultados para a coleção de documentos *SemEval 2015*. Nos gráficos da Figura 43 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_-Orig_10%. No gráfico da Figura 43a pode-se observar as quantidades de termos do domínio, na Figura 43b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 43c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 43a que, diferente das outras coleções de documentos, a quantidade de termos do domínio extraídos pelas regras morfossintáticas é bastante superior à quantidade de termos das Listas_Orig_100% e Listas_Orig_10%. Na Figura 43b verifica-se que para os identificadores da classe positiva, as Listas_Orig_100% possuem uma quantidade de termos bastante superior e,

Listas_Syntax_100% e Listas_Orig_10% possuem uma quantidade semelhante. Na Figura 43c verifica-se que a quantidade de identificadores da classe negativa é menor do que aqueles das Listas_Orig_10%.

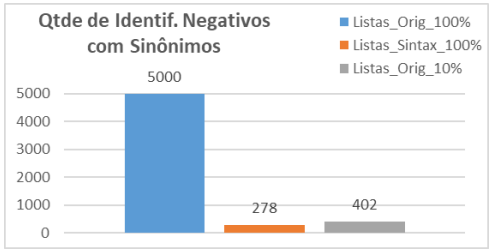
Figura 43 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 43 é apresentada a representatividade para o conjunto de documentos *SemEval 2015*. Nela é possível observar que para essa coleção de documentos, nas três representações Listas_Syntax_100% obteve a maior taxa de documentos representados, com 95,88% para gBoED_Freq e gBoED_Dist e 84,52% para gBoED_Syntax. Considerando as Listas_Orig_100%, gBoED_Syntax obteve 74,29% de documentos representados, gBoED_Freq e gBoED_Dist obtiveram 55,81%. Com relação à quantidade de acertos na predição, nas três representações, Listas_Syntax_100% obteve o maior número de acertos da predição, atingindo 78,52% em gBoED_Dist, 76,28% em gBoED_Freq e 66,92% em gBoED_Syntax. Em gBoED_Syntax com Listas_Syntax_100%, a taxa de erros na predição foi menor do que a taxa de neutros, mostrando melhor qualidade na representação.

Na Tabela 44 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel RBF, com Listas_-Syntax_100%**. Nesse cenário, a **acurácia obtida foi de 88,013%, Medida-F1 de 87,977%**, $\gamma = 10^{-1}$, grau de confiança de 60%, com 68 documentos sendo consultados e 44 documentos reclassificados. Vale destacar que nos experimentos com enriquecimento das representações gBoED_Freq e gBoED_Dist, usando Listas_Orig_100% e Listas_Orig_10% nenhum resultado superou a BoW. Usando Listas_Syntax_100%, os modelos enriquecidos pelas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax atingiram resultados superiores ao da BoW. O

Tabela 43 – Representatividade da gBoED no conjunto de dados para *SemEval 2015*.

	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
	Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
Qtde de documentos representados	447	55,81%	768	95,89%	424	52,93%	447	55,81%	768	95,89%	424	52,93%	595	74,29%	677	84,52%	528	65,92%
Qtde de documentos sem representação	354	44,19%	33	4,11%	377	47,06%	354	44,19%	33	4,11%	377	47,06%	206	25,71%	124	15,48%	273	34,08%
Número de ACERTOS na predição	351	43,83%	611	76,28%	321	40,07%	357	44,58%	629	78,53%	334	41,71%	463	57,80%	536	66,92%	414	51,69%
Número de ERROS na predição	72	8,98%	126	15,73%	79	9,86%	66	8,23%	108	13,48%	66	8,23%	68	8,50%	79	9,86%	64	7,99%
Número de NEUTROS na predição	378	47,19%	64	7,99%	401	50,06%	378	47,19%	64	7,99%	401	50,06%	270	33,70%	186	23,22%	323	40,32%

Fonte: Elaborada pelo autor.

melhor resultado obtido nesse caso, acontece em gBoED_Dist, usando Listas_Syntax_100%, com acurácia de 89,389%, seguido de gBoED_Freq, usando Listas_Syntax_100%, com acurácia de 88,516%. Outro destaque é que todos os modelos enriquecidos pelas representações gBoED_-Freq e gBoED_Dist, formadas pelas Listas_Syntax_100% superaram o resultado da BoW. C4.5-Entropia superou em mais de 3%, C4.5-Gini usando gBoED_Freq superou a BoW de mais de 2%.

Tabela 44 – Melhores acurácias dos classificadores gerados para *SemEval 2015*.

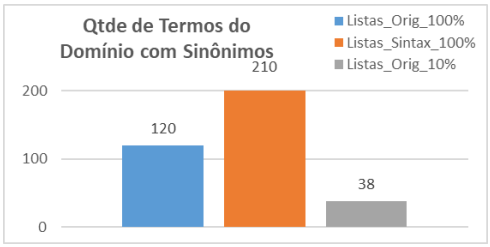
Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75404	0,74904 32-18	60%	0,78773 373-97	95%	0,74529 32-17	55%	0,74779 32-17	60%	0,787713 72-91	95%	0,74529 32-17	60%	0,76657 33-20	60%	0,75279 32-21	60%	0,76407 1-1	50%
C4.5-Gini	0,77528	0,76028 31-21	60%	0,79898 373-92	95%	0,76028 31-20	60%	0,76028 31-20	60%	0,787713 373-92	95%	0,76028 31-20	60%	0,71902 366-208	85%	0,76902 31-18	60%	0,68648 366-217	85%
KNN-Cosseno	0,81526	0,79778 40-30 $n = 25$	55%	0,81645 40-30 $n = 7$	60%	0,79528 40-30 $n = 25$	55%	0,79653 40-29 $n = 25$	55%	0,82018 155-79 $n = 7$	60%	0,79528 40-31 $n = 25$	55%	0,81271 85-56 $n = 13$	55%	0,81778 155-90 $n = 7$	60%	0,79778 40-31 $n = 25$	55%
KNN-Euclideana	0,81651	0,79903 39-29 $n = 25$	55%	0,81770 155-79 $n = 7$	60%	0,79653 39-29 $n = 25$	55%	0,79778 39-28 $n = 25$	55%	0,82143 155-79 $n = 7$	60%	0,79653 39-30 $n = 25$	55%	0,81271 86-57 $n = 13$	55%	0,81902 155-90 $n = 7$	60%	0,79778 39-30 $n = 25$	55%
MNB	0,86145	0,85145 24-19 $\alpha = 10^{-2}$	55%	0,85895 24-13 $\alpha = 10^{-2}$	55%	0,85020 24-18 $\alpha = 10^{-2}$	55%	0,85148 35-25 $\alpha = 10^{-1}$	55%	0,86146 97-39 $\alpha = 10^{-1}$	65%	0,85020 24-17 $\alpha = 10^{-2}$	55%	0,86146 63-44 $\alpha = 10^{-1}$	60%	0,86145 64-37 $\alpha = 10^{-1}$	60%	0,85770 60-44 $\alpha = 10^{-1}$	60%
SVM-Linear	0,86148	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,88146 160-85 $\gamma = 10^{-4}$ a 10^4	75%	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,86398 61-48 $\gamma = 10^{-4}$ a 10^4	60%	0,89020 160-78 $\gamma = 10^{-4}$ a 10^4	75%	0,86148 61-49 $\gamma = 10^{-4}$ a 10^4	60%	0,85523 34-27 $\gamma = 10^{-4}$ a 10^4	55%	0,86020 30-23 $\gamma = 10^{-4}$ a 10^4	55%	0,85148 34-31 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,87270	0,87020 5-3 $\gamma = 1$	50%	0,88895 91-30 $\gamma = 1$	65%	0,87020 5-2 $\gamma = 1$	50%	0,87270 5-4 $\gamma = 1$	50%	0,89518 91-28 $\gamma = 1$	65%	0,87147 30-16 $\gamma = 1$	55%	0,86768 2-2 $\gamma = 10^{-2}$	50%	0,87145 5-4 $\gamma = 1$	50%	0,86768 2-2 $\gamma = 10^{-2}$	50%
SVM-RBF	0,87390	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,88516 129-32 $\gamma = 1$	70%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,89389 192-47 $\gamma = 1$	80%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87140 2-2 $\gamma = 10^{-1}$	50%	0,88013 68-44 $\gamma = 10^{-1}$	60%	0,87015 4-4 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

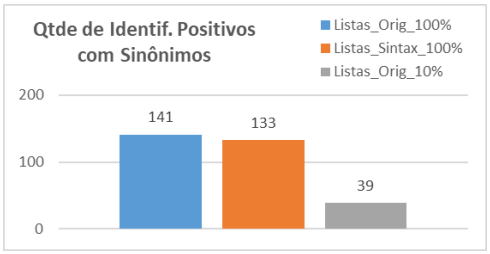
Em seguida são analisados os resultados para a coleção de documentos *SemEval 2015 Hotel*. Nos gráficos da Figura 44 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 44a pode-se observar as quantidades de termos do domínio, na Figura 44b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 44c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 44a que, de forma semlhante à coleção de documentos *SemEval 2015* que possui três diferentes domínio, a quantidade de termos do domínio extraídos pelas regras morfossintáticas é bastante superior à quantidade de termos das Listas_Orig_100% e Listas_-

Orig_10%. Na Figura 44b verifica-se que para os identificadores da classe positiva, as Listas_-Orig_100% possuem uma quantidade de termos um pouco superior às Listas_Syntax_100% e Listas_Orig_10% possui uma queda mais abrupta na quantidade de termos. Na Figura 44c verifica-se que a quantidade de identificadores da classe negativa para Listas_Orig_100% é bastante superior às Listas_Syntax_100% e Listas_Orig_10%.

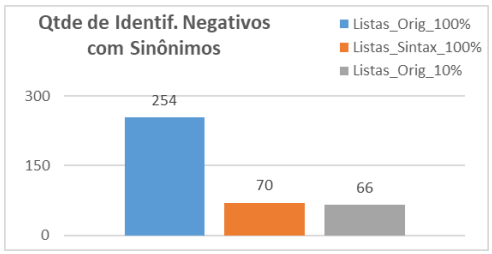
Figura 44 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Hotel*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 45 é apresentada a representatividade para o conjunto de documentos *SemEval 2015 Hotel*. Apesar de ser uma coleção com poucos documentos, nela é possível observar que para essa coleção de documentos, nas representações gBoED_Freq e gBoED_Dist, Listas_Orig_100% e Listas_Syntax_100% obtiveram a maior taxa de documentos representados, com 100%. Na representação gBoED_Syntax, Listas_Orig_100% e Listas_Syntax_100% obtiveram 96,56% de documentos representados. Com relação à quantidade de acertos na predição, na representações gBoED_Dist, Listas_Syntax_100% obteve o maior número de acertos na predição, com 86,22%. Nas outras representações Listas_Orig_100% e Listas_Syntax_100% obtiveram taxa de acertos iguais a 82,77%. Na representação gBoED_Syntax, Listas_Orig_100% e Listas_Syntax_100% obtiveram 79,32% de acertos na predição. Em gBoED_Syntax com Listas_Syntax_100%, a taxa de erros na predição foi menor ou igual à taxa de neutros.

Na Tabela 46 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Hotel*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine*, kernel Polinomial, com Listas_Orig_100%**. Nesse cenário, a **acurácia obtida foi de 86,667%, Medida-F1 de 86,667%**, $\gamma = 1$, grau de confiança de 80%, com 15 documentos sendo consultados e 6 documentos reclassificados. Outro resultado interessante se dá pelo mesmo algorimo com o uso

Tabela 45 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Hotel*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	29	100%	29	100%	28	96,56%	29	100%	29	100%	28	96,56%	28	96,56%	28	96,56%	22	75,86%
Qtde de documentos sem representação	0	0%	0	0%	1	3,44%	0	0%	0	0%	1	3,44%	1	3,44%	1	3,44%	7	24,14%
Número de ACERTOS na predição	24	82,77%	24	82,77%	22	75,87%	24	82,77%	25	86,22%	23	79,32%	23	79,32%	23	79,32%	17	58,63%
Número de ERROS na predição	4	13,79%	3	10,34%	5	17,24%	4	13,79%	2	6,89%	4	13,79%	3	10,34%	2	6,89%	4	13,79%
Número de NEUTROS na predição	1	3,44%	2	6,89%	2	6,89%	1	3,44%	2	6,89%	2	6,89%	3	10,34%	4	13,79%	8	27,58%

Fonte: Elaborada pelo autor.

de gBoED_Syntax e Listas_Syntax_100%. Nesse cenário, a acurácia obtida foi de 83,333%, Medida-F1 de 83,333%, $\gamma = 1$, grau de confiança de 80%, com 15 documentos sendo consultados e 6 documentos reclassificados. Vale destacar que outros algoritmos também obtiveram bons resultados em gBoED_Syntax, tanto usando Listas_Orig_100% quanto Listas_Syntax_100%, como é o caso de KNN-Cosseno, KNN-Euclidiana, MNB e SVM-Linear. Praticamente todos os modelos enriquecidos pelas representações gBoED_Freq e gBoED_Dist geradas com Listas_-Syntax_100% superaram os resultados da BoW, como maior destaque para gBoED_Dist com Listas_Syntax_100% em todos os algoritmos.

Tabela 46 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Hotel*.

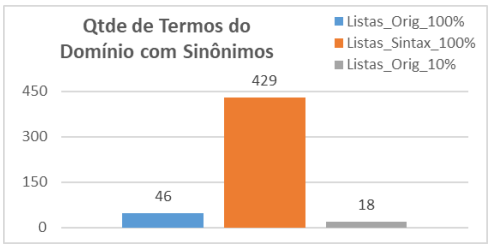
Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**						**gBoED_Syntax**					
		Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,83333	0,83333 29-8	100%	0,83333 29-6	100%	0,76667 29-10	100%	0,83333 29-8	100%	0,86667 29-5	100%	0,80000 29-9	100%	0,80000 29-9	100%	0,80000 29-8	100%	0,60000 29-14	100%
C4.5-Gini	0,83333	0,83333 29-8	100%	0,83333 29-6	100%	0,76667 29-10	100%	0,83333 29-8	100%	0,86667 29-5	100%	0,80000 29-9	100%	0,80000 29-9	100%	0,80000 29-8	100%	0,60000 29-14	100%
KNN-Cosseno	0,80000	0,86667 15-5 $n = 7$	90%	0,83333 11-5 $n = 15$	80%	0,80000 11-4 $n = 3$	70%	0,86667 15-4 $n = 7$	90%	0,90000 11-5 $n = 15$	80%	0,80000 11-4 $n = 3$	70%	0,86667 16-6 $n = 7$	90%	0,83333 11-4 $n = 3$	70%	0,73333 1-1 $n = 11$	55%
KNN-Euclideana	0,80000	0,86667 15-5 $n = 7$	90%	0,83333 11-5 $n = 15$	80%	0,80000 11-4 $n = 3$	70%	0,86667 15-4 $n = 7$	90%	0,90000 11-5 $n = 15$	80%	0,80000 11-4 $n = 3$	70%	0,86667 16-6 $n = 7$	90%	0,83333 11-4 $n = 3$	70%	0,73333 2-1 $n = 7$	60%
MNB	0,73333	0,86667 14-5 $\alpha = 1$	85%	0,83333 14-5 $\alpha = 1$	85%	0,80000 8-3 $\alpha = 1$	80%	0,86667 14-4 $\alpha = 1$	85%	0,90000 14-5 $\alpha = 1$	85%	0,80000 8-3 $\alpha = 1$	80%	0,86667 14-6 $\alpha = 10^{-2}$ a 1	85%	0,83333 14-6 $\alpha = 1$	85%	0,73333 2-2 $\alpha = 10^{-1}$	65%
SVM-Linear	0,73333	0,83333 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,83333 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,80000 8-2 $\gamma = 10^{-4}$ a 10^{4}	70%	0,83333 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,86667 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,80000 8-2 $\gamma = 10^{-4}$ a 10^{4}	70%	0,83333 9-5 $\gamma = 10^{-4}$ a 10^{4}	75%	0,83333 17-6 $\gamma = 10^{-4}$ a 10^{4}	90%	0,73333 2-1 $\gamma = 10^{-4}$ a 10^{4}	60%
SVM-Polinomial	0,73333	0,83333 6-3 $\gamma = 10^{-1}$	70%	0,83333 10-4 $\gamma = 10^{-1}$	75%	0,80000 6-2 $\gamma = 10^{-1}$	70%	0,83333 6-3 $\gamma = 10^{-1}$	70%	0,86667 10-4 $\gamma = 10^{-1}$	75%	0,80000 6-2 $\gamma = 10^{-1}$	70%	0,86667 15-6 $\gamma = 1$	80%	0,83333 15-6 $\gamma = 1$	80%	0,73333 1-0 $\gamma = 1$	55%
SVM-RBF	0,73333	0,83333 21-3 $\gamma = 1$	75%	0,83333 29-5 $\gamma = 10^{-1}$ a 10^{4}	85%	0,76667 21-2 $\gamma = 1$	60%	0,87142 2-2 $\gamma = 10^{-1}$	75%	0,86667 29-6 $\gamma = 10^{-1}$ a 10^{4}	85%	0,76667 21-2 $\gamma = 1$	60%	0,80000 29-8 $\gamma = 10^{-4}$ a 10^{4}	85%	0,80000 29-7 $\gamma = 10^{-1}$ a 10^{4}	85%	0,70000 3-1 $\gamma = 1$	60%

Fonte: Elaborada pelo autor.

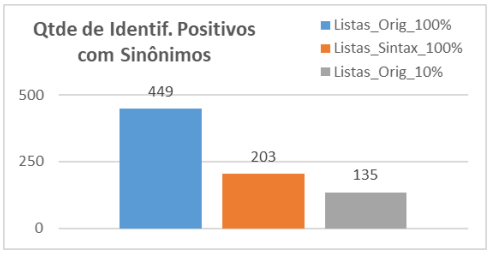
Em seguida são analisados os resultados para a coleção de documentos *SemEval 2015 Laptop*. Nos gráficos da Figura 45 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 45a pode-se observar as quantidades de termos do domínio, na Figura 45b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 45c observa-se as quantidades para identificadores da classe "*Negative*". É possível

verificar na Figura 45a que, de forma semlhante à coleção de documentos *SemEval 2015*, que possui três diferentes domínio, a quantidade de termos do domínio extraídos pelas regras morfossintáticas é bastante superior à quantidade de termos das Listas_Orig_100% e Listas_-Orig_10%. Na Figura 45b verifica-se que para os identificadores da classe positiva, as Listas_-Orig_100% possuem uma quantidade de termos um pouco superior às Listas_Syntax_100% e Listas_Orig_10% possui uma queda mais abrupta na quantidade de termos. Na Figura 45c verifica-se que a quantidade de identificadores da classe negativa para Listas_Orig_100% é bastante superior às Listas_Syntax_100% e Listas_Orig_10%.

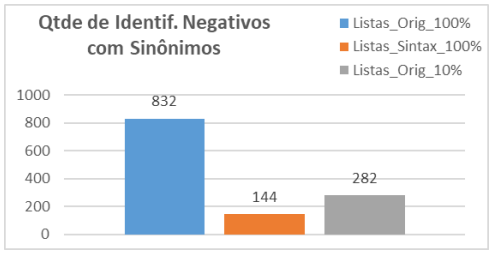
Figura 45 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Laptop*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 47 é apresentada a representatividade para o conjunto de documentos *SemEval 2015 Laptop*. Nela é possível observar que para essa coleção de documentos, nas três representações Listas_Syntax_100% obteve a maior taxa de documentos representados, com 90,66% em gBoED_Freq e gBoED_Dist e 76,64% emgBoED_Syntax. As Listas_Orig_100% e Listas_Orig_10% quando nas representações gBoED_Freq e gBoED_Dist não representaram bem essa coleção de documentos. Com relação à quantidade de acertos na predição, nas três representações Listas_Syntax_100% obteve o maior número de acertos da predição, atingindo 72,66% em gBoED_Freq, 75,93% em gBoED_Dist e 63,78% em gBoED_Syntax. Em gBoED_-Syntax com Listas_Syntax_100%, a taxa de erros na predição foi menor do que a taxa de neutros, demonstrando melhor qualidade na representação e que tanto a representação gBoED_Syntax quanto as Listas_Syntax_100% aplicadas a todas as representações, representaram melhor a coleção.

Na Tabela 48 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Laptop*. O melhor resultado da gBoED_Syntax está des-

Tabela 47 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Laptop*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	90	21,03%	388	90,66%	67	15,66%	90	21,03%	388	90,66%	67	15,66%	271	63,32%	328	76,64%	232	54,21%
Qtde de documentos sem representação	338	78,97%	40	9,34%	361	84,34%	338	78,97%	40	9,34%	361	84,34%	157	36,68%	100	23,36%	196	45,79%
Número de ACERTOS na predição	58	13,56%	311	72,68%	44	10,29%	64	14,96%	325	75,95%	47	10,98%	207	48,38%	273	63,78%	174	40,68%
Número de ERROS na predição	23	5,37%	63	14,71%	19	4,43%	17	3,97%	49	11,44%	16	3,74%	32	7,47%	28	6,55%	30	7%
Número de NEUTROS na predição	347	81,07%	54	12,61%	365	85,28%	347	81,07%	54	12,61%	365	85,28%	189	44,15%	127	29,67%	224	52,33%

Fonte: Elaborada pelo autor.

tacado pela linha cinza e se dá pelo algoritmo ***MNB*, com Listas_Syntax_100%**. Nesse cenário, a **acurácia obtida foi de 87,857%, Medida-F1 de 87,857%**, $\gamma = 10^{-1}$, grau de confiança de 55%, com 15 documentos sendo consultados e 9 documentos reclassificados. Outro resultado interessante se dá pelo mesmo algorimo com o uso de gBoED_Dist e Listas_Syntax_100%. Nesse cenário, a acurácia obtida foi de 87,857%, Medida-F1 de 87,857%, $\gamma = 10^{-1}$, grau de confiança de 55%, com 15 documentos sendo consultados e 8 documentos reclassificados. Vale destacar que outros algoritmos também obtiveram bons resultados em gBoED_Freq e gBoED_Dist, usando Listas_Syntax_100%, como é o caso de C4.5-Entropia e C4.5-Gini.

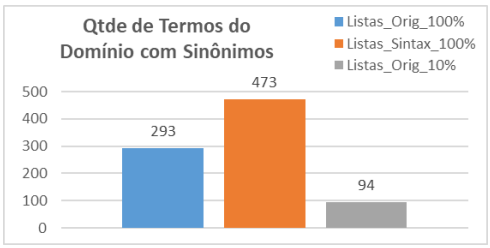
Tabela 48 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Laptop*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**						**gBoED_Syntax**					
		Listas_Orig 100%		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Syntax 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75238	0,71285 41-37	55%	0,76644 146-60	90%	0,71052 41-38	55%	0,71285 41-37	55%	0,77569 146-58	90%	0,71052 41-38	55%	0,73837 27-20	55%	0,74075 41-29	55%	0,71052 41-38	55%
C4.5-Gini	0,75476	0,75011 12-11	55%	0,76173 108-40	75%	0,74779 12-11	55%	0,75011 12-11	55%	0,76185 151-59	90%	0,74779 12-11	55%	0,76871 12-6	55%	0,75011 12-8	55%	0,74779 12-11	55%
KNN-Cosseno	0,85753	0,81085 31-26 $n = 17$	55%	0,84352 62-36 $n = 9$	60%	0,80853 31-26 $n = 17$	55%	0,81085 31-26 $n = 17$	55%	0,85055 79-34 $n = 7$	60%	0,80853 31-26 $n = 17$	55%	0,83892 44-30 $n = 11$	55%	0,85049 43-23 $n = 11$	55%	0,83194 31-23 $n = 17$	55%
KNN-Euclideana	0,85753	0,80847 32-27 $n = 17$	55%	0,84114 63-37 $n = 9$	60%	0,80615 32-27 $n = 17$	55%	0,80847 32-27 $n = 17$	55%	0,84817 80-35 $n = 7$	60%	0,80615 32-27 $n = 17$	55%	0,83654 44-30 $n = 11$	55%	0,84811 44-24 $n = 11$	55%	0,82956 32-24 $n = 17$	55%
MNB	0,87625	0,85997 15-14 $\alpha = 10^{-1}$	55%	0,87392 15-6 $\alpha = 10^{-1}$	55%	0,85997 15-14 $\alpha = 10^{-1}$	55%	0,86229 15-14 $\alpha = 10^{-1}$	55%	0,87857 15-8 $\alpha = 10^{-1}$	55%	0,86229 15-14 $\alpha = 10^{-1}$	55%	0,87392 13-10 $\alpha = 10^{-1}$	55%	0,87857 15-9 $\alpha = 10^{-1}$	55%	0,86926 13-12 $\alpha = 10^{-1}$	55%
SVM-Linear	0,86916	0,86683 2-1 $\gamma = 10^{-4}$ a 10^{4}	50%	0,86910 25-16 $\gamma = 10^{-4}$ a 10^{4}	60%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^{4}	50%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^{4}	50%	0,87602 122-55 $\gamma = 10^{-4}$ a 10^{4}	85%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^{4}	70%	0,86218 2-2 $\gamma = 10^{-4}$ a 10^{4}	50%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^{4}	50%	0,86218 2-2 $\gamma = 10^{-4}$ a 10^{4}	50%
SVM-Polinomial	0,89020	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88316 3-3 $\gamma = 10^{-2}$	50%	0,88555 2-2 $\gamma = 10^{-2}$	50%	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88554 3-2 $\gamma = 10^{-2}$	50%	0,88555 3-2 $\gamma = 10^{-2}$	55%	0,88322 3-2 $\gamma = 10^{-2}$	50%	0,88554 3-2 $\gamma = 10^{-2}$	50%	0,88322 3-2 $\gamma = 10^{-2}$	50%
SVM-RBF	0,88311	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88078 26-10 $\gamma = 1$	60%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88311 26-9 $\gamma = 1$	60%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%

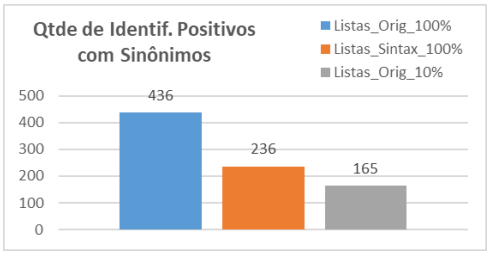
Fonte: Elaborada pelo autor.

Na Tabela 49 é apresentada a representatividade para o conjunto de documentos *SemEval 2015 Restaurant*. Nela é possível observar que as representações gBoED_Freq e gBoED_Dist obtiveram uma quantidade significativa de documentos representados, 96,81% usando Listas_Syntax_100% e 95,35% usando Listas_Orig_100%. Nas representação gBoED_Syntax, a quantidade de documentos representados usando Listas_Orig_100% e Listas_Syntax_100% atingiram valores próximos entre si, de 86,63% e 84,89% respectivamente. Com relação à quantidade de acertos na predição, nas três representações Listas_Syntax_100% obteve o maior

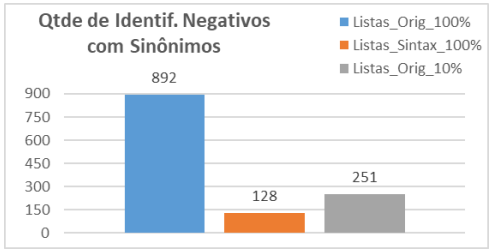
Figura 46 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Restaurant*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

número de acertos da predição, atingindo 81,68% em gBoED_Dist, 79,80% em gBoED_Freq e 69,76% em gBoED_Syntax. As taxas de acertos na predição, nas três representações, usando Listas_Syntax_100% obtiveram valores próximos àqueles obtidos ao utilizar Listas_Orig_100%, com atingindo 81,68% em , 79,20% em gBoED_Freq e gBoED_Dist e 67,45% em gBoED_-Syntax. Em gBoED_Syntax com Listas_Syntax_100%, a taxa de erros na predição foi menor do que a taxa de neutros, demonstrando melhor qualidade na representação.

Tabela 49 – Representatividade da gBoED no conjunto de dados para *SemEval 2015 Restaurant*.

	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
	Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
Qtde de documentos representados	328	95,35%	333	96,81%	312	90,70%	328	95,35%	333	96,80%	312	90,70%	298	86,63%	292	84,89%	266	77,33%
Qtde de documentos sem representação	16	4,65%	11	3,19%	32	9,30%	16	4,65%	11	3,19%	32	9,30%	46	13,37%	52	15,11%	78	22,67%
Número de ACERTOS na predição	269	78,20%	272	79,08%	247	71,81%	269	78,20%	281	81,70%	255	74,13%	232	67,45%	240	69,76%	210	61,06%
Número de ERROS na predição	45	13,08%	51	14,82%	51	14,82%	45	13,08%	42	12,20%	43	12,50%	34	9,88%	29	8,44%	33	9,59%
Número de NEUTROS na predição	30	8,72%	21	6,10%	46	13,37%	30	8,72%	21	6,10%	46	13,37%	78	22,67%	75	21,80%	101	29,36%

Fonte: Elaborada pelo autor.

Na Tabela 50 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Restaurant*. O melhor resultado da gBoED_Syntax está destacado pela linha cinza e se dá pelo algoritmo ***Support Vector Machine* - SVM, com kernel Polinomial, com Listas_Orig_100%**. Nesse cenário, a **acurácia obtida foi de 88,369%, Medida-F1 de 88,369%**, $\gamma = 1$, grau de confiança de 50%, com 2 documentos sendo consultados e 1 documento reclassificado. Outro resultado interessante se dá pelo mesmo algorimo

com o uso de gBoED_Dist e Listas_Syntax_100%. Nesse cenário, a acurácia obtida foi de 89,226%, Medida-F1 de 89,226%, $\gamma = 10^{-1}$, grau de confiança de 55%, com 20 documentos sendo consultados e 4 documentos reclassificados. Vale destacar que todos algoritmos enriquecidos por gBoED_Syntax e gBoED_Dist, usando Listas_Syntax_100%, obtiveram resultados superiores à BoW. Nessas representações, em sua maiorias os resultados usando Listas_Syntax_-100% obtiveram resultados intermediários às representações formadas pelas Listas_Orig_100% e Listas_Orig_10%.

Tabela 50 – Melhores acurácias dos classificadores gerados para *SemEval 2015 Restaurant*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Syntax 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,81723	0,81403 30-8	85%	0,81714 30-3	85%	0,81118 30-7	55%	0,81983 30-7	85%	0,82865 158-19	90%	0,81697 30-6	60%	0,81117 30-11	85%	0,81117 30-11	85%	0,78789 1-1	85%
C4.5-Gini	0,77328	0,78168 344-96	100%	0,79050 344-81	100%	0,72672 272-46	95%	0,81983 30-7	85%	0,81672 344-74	100%	0,74126 272-46	100%	0,78789 1-1	85%	0,81403 1-1	75%	0,81403 1-1	75%
KNN-Cosseno	0,81378	0,84303 81-29 $n = 11$	65%	0,82563 46-16 $n = 9$	60%	0,81378 28-16 $n = 13$	55%	0,78176 344-97 $n = 11$	100%	0,82571 126-38 $n = 55$	75%	0,82563 96-38 $n = 5$	60%	0,83411 79-42 $n = 11$	65%	0,83403 28-16 $n = 13$	55%	0,81092 39-22 $n = 11$	55%
KNN-Euclideana	0,81378	0,84303 81-29 $n = 11$	65%	0,82563 46-16 $n = 9$	60%	0,81378 28-16 $n = 13$	55%	0,85168 81-30 $n = 11$	65%	0,83151 96-30 $n = 5$	60%	0,82563 96-38 $n = 5$	55%	0,83411 79-42 $n = 11$	65%	0,83403 28-16 $n = 13$	55%	0,81092 39-22 $n = 11$	55%
MNB	0,86908	0,88084 34-12 $\alpha = 10^{-2}$	65%	0,87210 34-13 $\alpha = 10^{-2}$	65%	0,87773 11-5 $\alpha = 10^{-2}$	55%	0,88076 34-10 $\alpha = 10^{-2}$	65%	0,87504 34-11 $\alpha = 10^{-2}$	65%	0,87773 11-5 $\alpha = 10^{-2}$	55%	0,88067 11-7 $\alpha = 10^{-2}$	55%	0,88369 34-14 $\alpha = 10^{-2}$	65%	0,87478 11-7 $\alpha = 10^{-2}$	55%
SVM-Linear	0,87765	0,89513 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,88647 42-8 $\gamma = 10^{-4}$ a 10^4	65%	0,88933 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,89513 50-14 $\gamma = 10^{-4}$ a 10^4	70%	0,89226 42-9 $\gamma = 10^{-4}$ a 10^4	70%	0,86398 50-14 $\gamma = 10^{-4}$ a 10^4	70%	0,85453 92-50 $\gamma = 10^{-4}$ a 10^4	80%	0,88050 20-9 $\gamma = 10^{-4}$ a 10^4	55%	0,85436 15-10 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,88059	0,89504 50-13 $\gamma = 10^{-1}$	70%	0,88655 26-4 $\gamma = 1$	60%	0,88639 50-13 $\gamma = 10^{-1}$	70%	0,89218 50-10 $\gamma = 10^{-1}$	70%	0,89226 20-4 $\gamma = 10^{-1}$	55%	0,89210 50-10 $\gamma = 10^{-1}$	55%	0,88369 2-1 $\gamma = 1$	50%	0,88344 17-11 $\gamma = 10^{-1}$	55%	0,88084 2-2 $\gamma = 1$	50%
SVM-RBF	0,82261	0,83983 139-29 $\gamma = 1$	90%	0,83689 140-20 $\gamma = 1$	90%	0,82563 27-3 $\gamma = 1$	60%	0,84588 166-30 $\gamma = 1$	95%	0,84882 172-26 $\gamma = 1$	95%	0,82857 27-4 $\gamma = 1$	60%	0,82277 16-8 $\gamma = 1$	55%	0,82546 3-2 $\gamma = 1$	50%	0,82268 1-1 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

4.7 Considerações finais

Neste capítulo foi realizada a proposta e aplicação de um método para extração e classificação semanticamente enriquecida a partir de padrões morfossintáticos. As regras baseadas em padrões morfossintáticos são dependentes do idioma, do domínio e do tipo de classificação que se deseja realizar. Portanto, para a implementação do método proposto foram escolhidos os idiomas português e inglês e as regras construídas e aplicadas em coleções de documentos do domínio da análise de sentimentos.

Inicialmente, foram desenvolvidos dois conjuntos de regras para extração de termos do domínio e identificadores de classe, para os idiomas português e inglês. Ao todo são 7 regras para português e 7 regras para o idioma inglês. O método de extração termos pode ser considerado um método semiautomático, pois ainda mantém certa dependência do especialista do domínio. Essa característica é bastante importante pois o conhecimento do especialista do domínio é parte fundamental para a escolha e construção de expressões do domínio que realmente possam representar o conjunto de textos.

No método de extração termos por meio de regras morfossintáticas existe uma etapa de

anotação de rótulos que identificam as palavras do texto com suas características sintáticas. Essa etapa é de fundamental importância no processo pois é a partir dos rótulos que os padrões são identificados e os termos extraídos. Ao final do processo, o especialista realiza uma limpeza nos termos extraídos e identifica quais são os termos identificadores de cada classe. O método foi aplicado em diversas coleções de documentos. Como o objetivo do método é tornar o processo de extração mais automatizado, auxiliando o trabalho dos especialistas, o método foi aplicado em diversas coleções de documentos, em 100% dos documentos contidos em cada coleção, e as listas geradas pelo metodo recebe o nome de Listas_Syntax_100%.

Em seguida, utilizando o mesmo conjunto de regras desenvolvidas para o método de extração de termos foi gerada uma nova versão da representação semanticamente enriquecida por expressões do domínio, cuja construção baseia-se nos padrões morfossintáticos implementados. Essa versão da representação enriquecida foi nomeada como gBoED_Syntax.

Os experimentos foram conduzidos de modo a evidenciar características das listas de termos geradas a partir dos padrões morfossintáticos como a quantidade de termos. Foram extraídos conjuntos de medidas para identificar o nível de representatividade tanto das Listas_-Syntax_100% quanto da representação gBoED_Syntax em cada coleção. Por último, foram realizados experimentos que usando o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio, utilizando as Listas_Syntax_100% e a representação gBoED_Syntax como enriquecimento semântico. Os resultados foram coletados e avaliados com relação ao desempenho e também comparados com as outras versões da representação gBoED.

Com relação à extração de termos é possível avaliar que, de modo geral, as listas de termos geradas usando o método de extração baseado em regras obtiveram uma quantidade de termos bastante semelhante às Listas_Orig_10%, construídas de forma manual a partir de apenas 10% dos documentos de cada coleção.

Ao analisar a representatividade das listas e das representações verifica-se na maioria dos casos que o uso das Listas_Syntax_100% para construir as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax posiciona-se em um nível intermediário entre a representatividade obtida pelas representações formadas pelas Listas_Orig_100% e Listas_Orig_10%. Alguns destaques importantes são relacionados às coleções *B2W Reviews 2019 Info*, *HuLiu 2004* e *SemEval 2015 Restaurant* que obtiveram uma melhor representatividade ao utilizar as Listas_-Syntax_100%. As coleções *SemEval 2015* e *SemEval 2015 Laptop* obtiveram um aumento bastante significativo de representatividade, tanto na quantidade de documentos quanto na taxa de acertos na predição, ao utilizas as Listas_Syntax_100% e a representação gBoED_-Syntax. Outro destaque são as coleções *SemEval 2014*, *SemEval 2014 Laptop* e *SemEval 2014 Restaurant* que, apesar de não superarem, a quantidade de documentos representados e a taxa de acertos obtidas usando Listas_Syntax_100% ficaram bastante próximas daquelas atingidas usando Listas_Orig_100%, mostrando que o método automatizado pode melhorar o trabalho dos especialistas e atingir uma representatividade aproximada.

Nos experimentos de utilizando o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio foram avaliados diferentes cenários com as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax, construídas a partir das Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. Os resultados foram avaliados comparando as melhores acurácias obtidas nos diferentes cenário por modelos construídos por 4 algoritmos principais. Assim como nos experimentos do Capítulo 3, os melhores resultados utilizando enriquecimento por gBoED_Syntax ocorreram nos modelos gerados pelo algoritmo SVM. De modo semelhante ao nível de representatividade, na maioria dos casos os modelos enriquecidos pelas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax construídas a partir das Listas_Syntax_100% apresentaram resultados intermediários às representações construídas a partir das Listas_Orig_100% e Listas_Orig_10%. Os principais destaques são os resultados obtidos nas coleções *B2W Reviews 2019 Info*, *SemEval 2014*, *SemEval 2014 Restaurant*, *SemEval 2015* e *SemEval 2015 Laptop* com enriquecimento pelas representações construídas a partir de Listas_Syntax_100%. Nesses casos, além dos resultados obtidos pelo algoritmo SVM, diversos outros algoritmos, considerados mais explicáveis, obtiveram resultados significativos com o enriquecimento pelas representações construídas a partir de Listas_Syntax_100%.

Nesse capítulo, foi apresentado um método semiautomático baseado em regras morfossintáticas para extração de termos do domínio e identificadores de classe, uma nova versão da representação semanticamente enriquecida gBoED baseada em regras morfossintáticas, denominada gBoED_Syntax, e análise de impacto do uso das listas extraídas pelo método semiautomáticos no Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. Esse resultado está relacionado às questões de pesquisa **Q1** e **Q2** e aos objetivos específicos **O1**, **O2** e **O3**.

No Capítulo 5 será apresentada a construção e aplicação de um método de extração termos e classificação de documentos baseado em modelo de linguagem BERT.

CAPÍTULO

5

EXTRAÇÃO DE TERMOS E CLASSIFICAÇÃO SEMÂNTICA USANDO MODELOS DE LINGUAGEM BERT

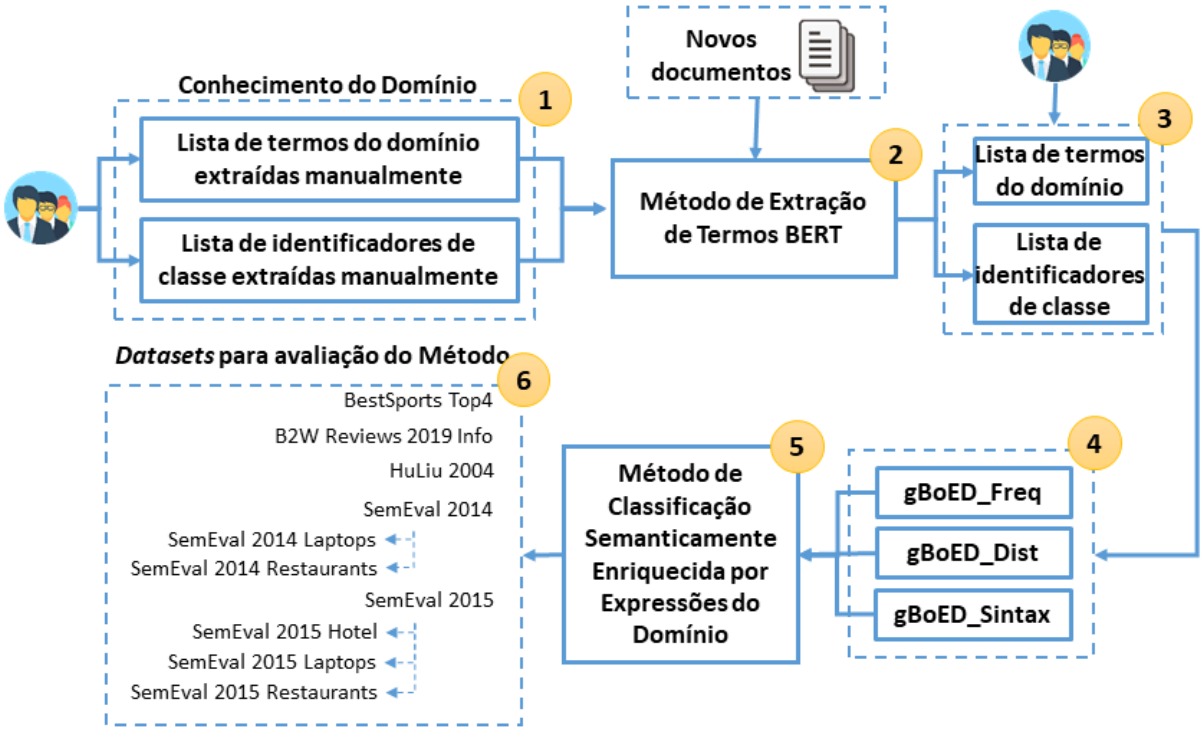
5.1 Considerações iniciais

Como apresentado no Capítulo 3, em cenários de alto nível de complexidade semântica a representação BoW não é suficiente pra representar e construir eficientes modelos de classificação, com isso existe a necessidade de combinação com representações que agreguem informações enriquecidas, como é o caso da gBoED. Como uma possível solução, foi proposto o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio que combina o treinamento de modelos tradicionais de classificação com a melhoria de resultados a partir de predições baseadas em expressões do domínio como informações privilegiadas. Ainda no Capítulo 3, foram apresentadas as representações semanticamente enriquecidas por expressões do domínio, gBoED_Freq e gBoED_Dist, e no Capítulo 4 foi proposta a representação baseada em análise morfossintática gBoED_Syntax, bem como o Método de extração de termos baseado em regras morfossintáticas afim de tornar o processo de extração de termos e construção das representações mais automatizado.

Neste capítulo é apresentada a aplicação de um método semiautomático para extração de termos do domínio e identificadores de classe e, consequentemente, para construção das representações semanticamente enriquecidas (gBoEDs) apresentadas nos capítulos anteriores. Com base no objetivo específico 3, definido na Seção 1.3, que visa aplicar e analisar o impacto de soluções de extração de termos para a construção de listas de termos do domínio e identificadores de classe, o principal objetivo deste capítulo é verificar se a aplicação um método de extração de termos semiautomático baseado em modelo de linguagem BERT é capaz de extrair termos, construir representações semanticamente enriquecidas e melhorar resultados de

classificação com desempenho melhor ou semelhante àqueles realizados com as listas extraídas manualmente pelos especialistas do domínio. Portanto, o Capítulo 5 está organizado de acordo com o diagrama Figura 47 que ilustra as 6 principais etapas de desenvolvimento e validação.

Figura 47 – Diagrama das etapas de desenvolvimento do Capítulo 5.



Fonte: Elaborada pelo autor.

Na etapa 1, os especialistas de domínio preparam listas com termos extraídos manualmente, que servem de treinamento para construção de um modelo de extração baseado em BERT, ilustrado na etapa 2 e apresentado na Seção 5.3. Novos documentos são inseridos na etapa 2 para que mais termos possam ser extraídos e as Listas de termos do domínio e identificadores de classe tornem-se mais completas. Na etapa 3, os especialistas realizam um processo de limpeza e organização das novas listas de termos. Na etapa 4, as representações semanticamente enriquecidas gBoED_Freq, gBoED_Dist e gBoED_Syntax são construídas a partir das novas listas de termos.

As etapas 5 e 6 fazem parte do processo de validação dos modelos de extração de termos BERT, apresentado na Subseção 5.3.2. A validação ocorre por meio do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, apresentado no Capítulo 3. O método de classificação faz uso das diferentes representações gBoED construídas a partir das listas de termos extraídas pelo método BERT e aplicadas em 10 coleções de documentos distintas. Os resultados da extração de termos, representatividade e classificação são apresentados com detalhes na Subsubseção 5.3.2.3.

5.2 Trabalhos relacionados

O uso dos modelos de linguagens BERT em diferentes tarefas vem sendo utilizado de mais modo cada vez mais intenso na literatura e extração de informações é uma dessas tarefas. Trabalhos que envolvem a tarefa de extração aspectos também podem ser considerados como relacionados à essa pesquisa. A seguir, são apresentados alguns trabalhos que utilizam modelos de linguagens BERT na extração de termos e aspectos.

Em Hoang, Bihorac e Rouces (2019) são propostos três modelos distintos usando BERT, aplicados em bases de idioma inglês, para as tarefas de classificação de sentimentos, classificação de aspectos e um modelo combinado entre os dois. O primeiro é um modelo de predição de sentimentos em textos. O treinamento ocorre com dados no formato sentença e um aspecto, bem como o formato de entrada para os documentos a serem preditos. A saída é a predição dentre os seguintes sentimentos: *"positive"*, *"negative"*, *"neutral"* e *"conflict"*, sendo que *"conflict"* significa que o texto possui partes onde o aspecto é julgado como positivo e partes onde é julgado como negativo. O segundo é um modelo de extração de aspectos com treinamento no formato sentença e um aspecto. Este modelo é usado para predizer um determinado aspecto está ou não relacionado ao texto. Predição ocorre a partir da classificação de um aspecto como *"related"* ou *"unrelated"*. O último modelo desenvolvido neste trabalho combina os dois anteriores em um classificador multi-classe para predizer aspecto e sentimento.

Os modelos são avaliados utilizando a coleção de documentos *SemEval-2016 Task 5*, mais especificamente: classificação de aspectos nas subtarefas 1 e 2, e polaridade de sentimentos na sutarefa 1 e 2. *SemEval-2016 Task 5* pode ser comparado ao uso da coleção *SemEval-2015* pois utilizam o mesmo conjuto de documentos de opiniões. A diferença é que em *SemEval-2016 Task 5* há um conjunto de testes adicional. Os modelos foram avaliados no conjunto de dados separados por domínio, já que *SemEval-2016 Task 5* também contém textos do domínio *Hotel*, *Laptops* e *Restaurants*. No modelo de classificação de aspectos, os melhores resultados atingiram 87,3% de acurácia no conjunto do domínio *Hotel*, 78,7% de acurácia no conjunto do domínio *Laptop* e 87,5% de acurácia na coleção do domínio *Restaurant*.

No trabalho de Xu *et al.* (2019) é proposto um novo ramo de pesquisa denominado *Review Reading Comprehension* (RRC) que baseia-se possui como base o *Machine Reading Comprehension* (MRC). O objetivo em MRC é a construção de bases de conhecimento a partir de documentos formais. Em RRC, o desafio é a construção de uma fonte de conhecimentos com base opiniões de produtos e serviços. Como base para a construção desta fonte de conhecimentos os autores exploram um método para extração de aspectos e sentimentos em opiniões de produtos. Neste trabalho as principais contribuições foram (1) a proposta de um novo ramo de pesquisa, o RRC (2) Construção de uma base de dados anotada para problemas do tipo RRC (3) Proposta de um novo método de pós-treinamento usando BERT para RRC, extração de aspectos e classificação de aspectos em sentimentos. Na extração de aspectos foram utilizados como base de treinamento aspectos encontrados na coleção de documentos *HuLiu 2004* (HU; LIU, 2004).

A validação do método é realizada sobre a coleção de documentos *SemEval 2014 Task 4*. Como resultados a acurácia obtida na classificação de sentimentos ao utilizar o conjunto de aspectos foi de 78,07% para o conjunto *Laptop* e 84,95% para o conjunto *Restaurant*.

Em Yanuar e Shiramatsu (2020) é construído um modelo baseado em BERT para extração de aspectos relacionados a opiniões de pontos turísticos, em idioma Indonésio. Nesse trabalho foi utilizada um coleção de documentos relacionadas a opiniões turísticas do portal *TripAdvisor*. O principal diferencial deste trabalho é a construção de uma fase de pré-processamento específica para o idioma indonésio e uma fase de pré-treinamento adional, além de um *fine-tunning* usando 501 sentenças relacionadas. A fase de pré-treinamento consiste em associar ao textos um conjunto de dados do mesmo domínio e idioma a fim de melhorar a performance antes da etapa de *fine-tunning*, na qual o modelo irá aprender a tarefa de extração de aspectos. Os resultados apresentaram um acurácia na extração dos aspectos de 82,4%.

No trabalho de Santos, Marcacini e Rezende (2021) é proposta uma abordagem de extração de aspectos usando BERT em um contexto de aprendizado multi-domínio. Como o processo de rotulação de aspectos e termos é uma atividade bastante onerosa, o trabalho visa criar um método de extração baseado em BERT, treinado a partir de aspectos anotados em diferentes domínios. A abordagem, denominada *MDAE-BERT (Multi-Domain Aspect Extraction using Bidirectional Encoder Representations from Transformers)* é focado nos padrões linguísticos existentes nas camadas do BERT e relaciona a inconsistência de aspectos entre domínios diferentes. No modelo é realizado um treinamento utilizando aspectos de coleções de documentos de diferentes domínios e aplicados a um conjunto de testes de cada domínio individualmente. A avaliação dos resultados foi realizada em diferentes cenários, sendo o mais interessante dentre os trabalhos relacionados, a comparação entre as melhores acurácias de um modelo BERT de domínio específico e a abordagem multi-domínio. De maneira geral, os modelos aplicados a domínios específicos obtiveram uma média de 90% de acurácia, enquanto a abordagem multi-domínio obteve em média 85% de acurácia.

Com base nos trabalhos descritos anteriormente, na Seção 5.3 é apresentado o Método de extração de termos baseado em modelo de linguagem BERT.

5.3 Método de extração de termos baseado em modelo de linguagem BERT

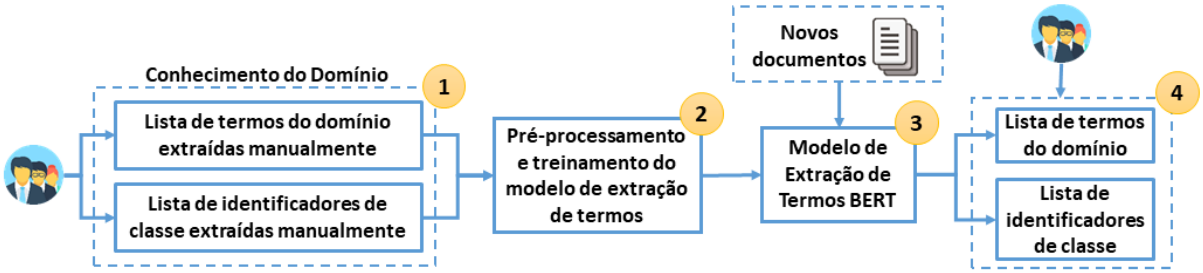
O objetivo deste capítulo é aplicação, validação e avaliação do impacto do uso de listas de termos geradas por métodos automatizados de extração de termos em diferentes cenários de classificação semântica. Com base no objetivo apresentado, foi adotado o método de classificação de aspectos de Hoang, Bihorac e Rouces (2019). A partir deste método foi realizado uma adaptação e implementação com o objetivo de torná-lo um método de extração de termos de domínio e identificadores de classe que possa ser aplicado em múltiplos cenários e aos idiomas

português e inglês.

5.3.1 Descrição do método de extração de termos baseado em modelo de linguagem BERT

Como apresentado na Seção 2.4, os modelos de linguagem têm como base o paradigma das redes neurais, portanto são modelos que aplicam **Aprendizado supervisionado**. No diagrama ilustrativo da Figura 48 são apresentadas as 4 etapas principais de desenvolvimento do método proposto.

Figura 48 – Diagrama ilustrativo do Método Semiautomático de Extração de Termos Usando Modelo de Linguagem BERT.



Fonte: Elaborada pelo autor.

A etapa 1 corresponde à fase de preparação do conjunto de dados para treinamento do modelo de extração de termos. Neste método são utilizados as listas construídas manualmente pelos próprios especialistas de domínio. Nelas estão contidos o **conhecimento de domínio** do especialista. Na etapa de validação deste método e das listas geradas automáticamente pelo modelo baseado em BERT, será utilizado o cenário mais próximo da realidade possível. Como o objetivo do método é diminuir o esforço do especialista, são utilizadas as Listas_Orig_10%, aquelas cujos termos são extraídos de apenas 10% dos textos de cada coleção, que servem como base de treinamento para construção do modelo.

Na etapa 2 é realizado o trabalho de pré-processamento dos textos, preparação e treinamento dos modelos. O pré-processamento dos textos é realizado com a preparação de um arquivo de entrada para treinamento de cada modelo BERT. Os arquivos são construídos para cada tipo de lista que será extraída pelo modelo (por exemplo, termos do domínio, termos positivos ou termos negativos). Esses arquivos foram construídos com base em uma estrutura simples que entrega ao algoritmo uma sentença, seguida de uma palavra ou expressão, seguida de um valor (por exemplo, 0 ou 1) que indica se aquela palavra ou expressão é um termo "válido" ou "inválido" em determinado tipo de lista. Entende-se por "válido" ou "inválido" aquele o tipo de termo que se deseja extrair, por exemplo um termo do domínio.

Nos experimentos realizados utilizou-se a separação por palavras simples (ou unigramas) para indicar os termos válidos e inválidos, com base nas listas anotadas inicialmente pelos

especialistas. As sentenças que indicam palavras válidas ou inválidas são inseridas no arquivo de forma balanceada, ou seja, existe a mesma quantidade de termos válidos e termos inválidos para cada sentença. No Quadro 7 observa-se um exemplo de estrutura de arquivo para treinamento do modelo BERT para extração de termos do domínio construída a partir de um documento da coleção *Best Sports Top 4*.

Quadro 7 – Exemplo de senteças para treinamento do modelo BERT para extração de termos do domínio na Coleção *Best Sports Top 4*.

Documento original:

Brasil vence a Bolívia em Sao Paulo e segue invicto nas Eliminatorias. A seleção brasileira de Futebol jogou completa neste domingo (05/09) e conseguiu uma vitoria tranquila sobre a Bolívia, pelas Eliminatorias Sul-Americanas da Copa do Mundo de 2006.

Termos de domínio da lista original:

- *Brasil*
- *seleção brasileira*

**Lembrete:* Os termos do domínio na coleção *Best Sports Top 4* são palavras ou expressões que indicam esportistas ou entidades do esporte brasileiro.

Exemplo do conjunto de treinamento:

- Brasil vence a Bolívia em Sao Paulo e segue invicto nas Eliminatorias | Brasil | 1

- Brasil vence a Bolívia em Sao Paulo e segue invicto nas Eliminatorias | Bolívia | 0

- A seleção brasileira de Futebol jogou completa neste domingo (05/09) e conseguiu uma vitoria tranquila sobre a Bolívia, pelas Eliminatorias Sul-Americanas da Copa do Mundo de 2006 | brasileira | 1

- A seleção brasileira de Futebol jogou completa neste domingo (05/09) e conseguiu uma vitoria tranquila sobre a Bolívia, pelas Eliminatorias Sul-Americanas da Copa do Mundo de 2006 | jogou | 0

Fonte: Elaborada pelo autor.

Na etapa 3, tem-se o Modelo de Extração de Termos BERT para cada tipo de lista. Nesta etapa novos documentos são inseridos ao processo de modo que sejam extraídos os termos para cada tipo de lista. Os textos são processados e inseridos no modelo de extração em um formato pré-definido. Esse formato corresponde a um par (sentença/palavra) e o modelo realiza a predição

das classes 0 para “Não é termo” e 1 para “É termo”. De forma complementar ao Quadro 7, no Quadro 8 é apresentado um exemplo com sentenças no formato de entrada para que sejam extraídos os termos contidos nelas.

Quadro 8 – Exemplo de senteças no formato de entrada submetidos ao modelo BERT para extração de termos do domínio na Coleção *Best Sports Top 4*.

Documento original:

Seleção feminina de Futebol do Brasil perde para a Alemanha por 2 a 1 nas Olimpíadas.

Termos de domínio esperados:

- Seleção feminina
- Brasil

Exemplo do conjunto de entrada para predição:

- Seleção feminina de Futebol do Brasil perde para a Alemanha por 2 a 1 nas Olimpiadas | Seleção

- Seleção feminina de Futebol do Brasil perde para a Alemanha por 2 a 1 nas Olimpiadas | feminina

- Seleção feminina de Futebol do Brasil perde para a Alemanha por 2 a 1 nas Olimpiadas | Alemanha

- Seleção feminina de Futebol do Brasil perde para a Alemanha por 2 a 1 nas Olimpiadas | perde

Fonte: Elaborada pelo autor.

Na etapa 4, as listas são formadas e os especialistas do domínio, realizam um proceso de limpeza e organização das listas. Ainda na etapa 4, como forma de auxiliar o trabalho de organização das listas por parte do especilista, foi utilizada automatizada que realiza a “radicalização” das palavras (*stemming*) e junta as palavras semelhantes em sinônimos. As listas geradas são utilizadas para gerar as representações semanticamente enriquecidas, gBoED, e aplicadas ao Método de Classificação Semanticamente Enriquecidas por Expressões do Domínio para melhoria de resultados de classificação.

5.3.2 Avaliação experimental do método de extração de termos baseado em modelo de linguagem BERT

Nessa seção é apresentada a avaliação experimental da aplicação da representação gBoED_Bert no Método de Classificação Semanticamente Enriquecido por Expressões do Domínio. Será analisado o impacto do uso da representação semanticamente enriquecida na melhoria de resultados de classificação de nível semântico. A principal diferença nesse caso são as listas de termos geradas pelo método semiautomático baseado em BERT.

5.3.2.1 Coleção de documentos

O objetivo das listas de termos extraídas pelo método semiautomático baseado em BERT é servir como recurso para a construção das representações semanticamente enriquecidas e, consequentemente, a melhoria dos resultados de classificação usando o método do capítulo 3. O método de extração proposto nesse capítulo não é extritamente dependente de domínio. Portanto, as coleções de documentos utilizadas no processo de extração de termos aplicado nesse capítulo são as mesmas utilizadas na avaliação experimental da Subseção 3.4.2. Elas possuem variação de idioma (inglês e português) e também de domínio.

B2W Reviews 2019 Info: coleção de documentos composta por 132.374 opiniões de produtos em português separadas em dezenas de categorias (REAL; OSHIRO; MAFRA, 2019). A *B2W Digital* é uma das maiores plataformas de e-commerce da América Latina. Para os experimentos deste trabalho foi selecionada uma categoria relacionada a um único domínio: Informática. Essa categoria é formada por 4262 opiniões de produtos como computadores, notebooks e tablets. A base original possui rotulação por pontos que vão de 1 (ruim) a 5 (excelente). Neste trabalho, por questões de capacidade de processamento, foram selecionadas aleatoriamente 1000 opiniões, 500 opiniões para cada classe. Os rótulos foram modificados para "Positivo" as opiniões pertencentes à pontuação 5 e "Negativo" as opiniões pertencentes às pontuações 1 e 2, de acordo com instruções dos autores.

BestSports Top4: é um conjunto de notícias de esportes escritas em língua portuguesa extraídas do *website BEST Sports*[1] e preparados para execução em diferentes níveis semânticos (SINOARA; REZENDE, 2018). A configuração utilizada nessa avaliação experimental é referente ao segundo nível de complexidade semântica, com um total de 181 notícias, sendo 93 notícias da classe "Brasileiro venceu" e 88 da classe "Brasileiro não venceu".

HuLiu 2004: coleção composta por opiniões de tipos produtos distintos (dois modelos de câmeras digitais, um modelo de telefone celular, um *MP3 Player* e um *DVD player*) (HU; LIU, 2004). A coleção utilizada é um subconjunto da coleção original. Nesta serão considerados 186 opinões positivas e 110 opiniões negativas.

[1] *BEST sports* - Arquivo de notícias: <http://bestsports.com.br/db/notarqhome.php>

SemEval 2014: coleção composta por opiniões de *laptops* e restaurantes. Inicialmente a coleção foi criada para a competição "*SemEval-2014 Aspect Based Sentiment Analysis task 4*" (PONTIKI *et al.*, 2014). Assim como em HuLiu 2004, nessa coleção será utilizado um subconjunto composto por 1.836 opiniões positivas and 1.073 opiniões negativas, totalizando 2.909 opiniões. Devido a essa coleção possuir dois subdomínios diferentes, serão utilizados divisões da coleção. São elas:

- **SemEval 2014 Laptop**: composta por 619 opiniões positivas e 622 negativas.
- **SemEval 2014 Restaurant**: composta por 1.217 opiniões positivas e 451 negativas.

SemEval 2015: coleção composta por opiniões de hotéis, *laptops* e restaurantes, criada para a competição "*SemEval-2015 Aspect Based Sentiment Analysis task 12*" (PONTIKI *et al.*, 2015). Nessa coleção será utilizado um subconjunto composto por 555 opiniões positivas and 246 opiniões negativas, totalizando 801 opiniões. Assim como SemEval 2015, serão consideradas divisões da coleção por subdomínio. São elas:

- **SemEval 2015 Hotel**: composta por 21 opiniões positivas e 8 negativas.
- **SemEval 2015 Laptop**: composta por 277 opiniões positivas e 151 negativas.
- **SemEval 2015 Restaurant**: composta por 257 opiniões positivas e 87 negativas.

5.3.2.2 *Configuração dos experimentos para validação do método de extração de termos baseado em modelo de linguagem BERT*

O principal foco desta avaliação experimental é analisar o impacto do uso das listas de termos geradas pelo método semiautomático de extração de termos baseado em modelo de linguagem BERT aplicados ao Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, em comparação com os resultados de classificadores gerados pela BoW e classificadores semanticamente enriquecidos gerados usando listas de termos construídas manualmente. Com isso, a configuração experimental ocorre de forma semelhante àquela aplicada na Subsubseção 3.4.2.2.

Com base nos objetivos desta avaliação experimental, alguns cenários foram idealizados para melhor comparação de resultados:

- **Listas_Orig_100%:** é o mesmo cenário denominado **Listas 100%**, apresentado na Subsubseção 3.4.2.2. É aquele cujas listas de termos foram construídas manualmente pelo especialista a partir de 100% dos documentos da coleção. Ele indica o melhor cenário para o método proposto de modo a comparar seus resultados a um limite superior.
- **Listas_Bert_100%:** este cenário corresponde às listas de termos do domínio e identificadores de classe gerados a partir do método de extração de termos baseado em BERT

e aplicado em 100% dos textos de cada coleção de documentos. Como é uma método semiautomático que visa diminuir o esforço dos especialistas de domínio, para que o cenário esteja o mais próximo possível da realidade, o treinamento do extrator de termos é realizado com apenas 10% do conjunto de textos anotados manualmente e é aplicado em 100% dos documentos. Nesse cenário é esperado que haja um aumento na quantidade de termos extraídos a partir de um conjunto de treinamento mais enxuto.

- **Listas_Syntax_100%:** este cenário corresponde às listas de termos do domínio e identificadores de classe gerados a partir de 100% dos textos de cada coleção de documentos, usando o método de extração de termos por regras morfossintáticas.

- **Listas_Orig_10%:** é o mesmo cenário denominado **Listas 10%**, apresentado na Subsubseção 3.4.2.2. Esse cenário possui importância semelhante ao Listas 100%. É aquele cujas listas de termos foram construídas manualmente pelo especialista a partir de apenas 10% dos documentos da coleção. Ele indica o pior cenário para o método proposto de modo a comparar seus resultados a um limite inferior.

Nessa avaliação experimental os cenários de comparação foram definidos com base nas listas de termos geradas a partir de 100% dos documentos de cada coleção, tanto no processo manual, quanto na extração semiautomática baseada em BERT. Como é esperado, os processos automatizados tendem a apresentar desempenho menor do que aqueles realizados manualmente pelos especialistas do domínio. Os cenários construídos a partir de listas formados por 100% dos documentos podem ser considerados como os mais próximos da realidade. Portanto, é a partir deles que serão considerados os principais resultados dos experimentos.

Os algoritmos utilizados são apresentados a seguir, seguidos dos parâmetros de configuração utilizados em cada caso. Nos experimentos executados foi utilizado método de amostragem *10-fold cross validation*. Tais algoritmos e variações são utilizadas para construção dos modelos no Método de Classificação Semanticamente Enriquecido por Expressões do Domínio.

- **C4.5**, algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor* e critérios para escolha do atributo: Entropia e Gini.

- ***K-nearest neighbor* (KNN)**, algoritmo IBk. Foram utilizadas as opções de voto com peso e voto sem peso. Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: Distância Euclideana e Distância de Cosseno.

- ***Multinomial Naïve Bayes* (MNB)**, algoritmo baseado em Naïve Bayes, com parâmentro alpha(α) considerando os valores: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} e 1.

- ***Suport Vector Machine* (SVM)**, algoritmo *Sequential Minimal Optimization* (SMO). Nesse algoritmo foram consideramos três tipos de kernel: linear, polinomial (expoente=2)

e RBF (*Radial Basis Function*). Os valores considerados para cada tipo de kernel foram 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}.

Na avaliação experimental também foram utilizados diferentes valores para o grau de confiança que o método de classificação considera para selecionar os documentos para reclassificação. Nos experimentos realizados o grau de confiança variou entre os valores 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% e 100%.

Nesse experimento foi utilizada a implementação em linguagem Python versão 3.7, no ambiente Google Colaboratory, biblioteca Scikit-Learn na versão 1.0.2, PyTorch 1.12.1 e Numpy na versão 1.21.6. Modelos pré-treinados foram utilizados para os idiomas português e inglês. Para o idioma português foi utilizada a biblioteca de modelos *BERTimbau*, na versão *"bert-base-portuguese-cased"* (SOUZA; NOGUEIRA; LOTUFO, 2020). Para o idioma inglês foi utilizada a biblioteca de modelos *Hugging Face Transformers*, na versão *"bert-base-uncased"* (DEVLIN *et al.*, 2018). Na Subsubseção 5.3.2.3 são apresentados os resultados obtidos nessa avaliação experimental.

5.3.2.3 Resultados - validação do método de extração de termos baseada em modelos de linguagem BERT

Nessa Subsubseção são apresentados os principais resultados obtidos a partir da avaliação experimental do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio Usando Modelos de Linguagem BERT. Os resultados são apresentados individualmente para cada coleção de textos. Iniciando com as coleções em idioma português e, em seguida, as coleções em inglês, os resultados estão no formato de tabelas contendo as melhores acurácias obtidas pelos modelos gerados por cada algoritmo utilizado.

Nesse capítulo, devido a grande quantidade de resultados a serem comparados, as tabelas que apresentam os valores de representatividade da gBoED nos conjuntos de dados e as tabelas com melhores acurácias de cada algoritmo foram divididas em duas partes. Em todas elas serão comparados os valores de representatividade em relação às representações gBoED_Freq, gBoED_Dist e gBoED_Syntax. A diferença entre cada grupo de tabelas está nas listas de termos que geram cada representação.

Nas Tabelas 51, 55, 57, 61, 65, 69, 73, 77, 81 e 85 são apresentados valores de representatividade em cada conjunto de dados, quando utilizadas as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax geradas a partir das Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Portanto, elas visam comparar a representatividade das gBoEDs geradas pela Listas_Bert_100% com as listas originais geradas manualmente.

Nas Tabelas 52, 58, 62, 66, 70, 74, 78, 82 e 86 são apresentados valores de representatividade em cada conjunto de dados, quando utilizadas as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax geradas a partir das Listas_Syntax_100% e Listas_Bert_100%. Portanto, elas

visam comparar a representatividade das gBoEDs geradas a partir das listas construídas pelos métodos semiautomáticos.

Nas tabelas citadas nos dois parágrafos anteriores é apresentada a quantidade de documentos cujas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax são capazes de representar e que não são capazes de representar devido a limitação na abrangência das listas de termos. Além disso, nelas são apresentadas as quantidades de documentos cujas predições a partir de cada uma das representações semanticamente enriquecidas obtiveram acertos e erros. Os documentos que não obtveram representação por parte das gBoEDs ou que obtiveram empate na predição, nestas tabelas estão sendo considerados como neutros e ao serem consultados dentro do método mantém a predição inicial do classificador.

Nas Tabelas 53, 56, 59, 63, 67, 71, 75, 79, 83 e 87 são apresentados o resultados dos modelos com as melhores acurácias obtidas em cada algoritmo, enriquecida pelas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax geradas a partir das Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Portanto, elas visam comparar as acurácias dos modelos enriquecidos pelas gBoEDs geradas a partir da Listas_Bert_100% com as listas originais geradas manualmente.

Nas tabelas citadas no parágrafo anterior, os algoritmos são apresentados de acordo com um nível decrescente de explicabilidade. Cada tabela de resultados está organizada de maneira que a linha em cinza corresponde ao melhor resultado obtido pelo modelos enriquecido pela Listas_Bert_100%. Na segunda coluna estão as melhores acurácias obtidas a partir dos modelos gerados via BoW. Na sequência, as melhores acurácias obtidas a partir dos modelos gerados pelo método de classificação semanticamente enriquecida incrementado pela representação gBoED_Freq, gBoED_Dist e gBoED_Syntax.

Em cada uma das representações, os resultados obtidos estão divididos em três colunas. A primeira coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Orig_100% (aquele cuja gBoED é construída a partir de listas de termos formadas manualmente por 100% dos documentos da coleção). A segunda coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Bert_100% (aquele cuja gBoED é construída a partir de listas de termos formadas pelo método de extração baseado em BERT, treinado com termos extraídos de 10% dos documentos e aplicado em 100% dos documentos da coleção). A terceira coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Orig_10% (aquele cuja gBoED é construída a partir de listas de termos formadas manualmente por 10% dos documentos da coleção). A leitura das tabelas deve ser realizada, primeiramente, observando-se o resultado da melhor acurácia para o cenário Listas_Orig_100%, já que ele corresponde ao cenário mais próximo do real. Em **Acc** é descrita a melhor acurácia e na coluna **Conf** é descrita o grau de confiança em que esse valor de acurácia foi obtido. Em seguida, observa-se o resultado da melhor acurácia para o cenário Listas_Bert_100% para comparação com o anterior. O cenário Listas_Orig_10% é uma referência de resultado

considerando o limite inferior. Em todos os cenários, logo abaixo do valor da acurácia, estão presentes uma sequência de valores que correspondem à quantidade de documentos enviados para reclassificação devido à confiança ser menor do que o valor descrito, seguido da quantidade desses documentos que foram realmente reclassificados, ou seja, sua predição sofreu alteração de classe.

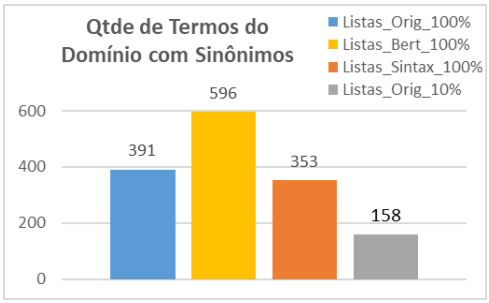
Nas Tabelas 54, 60, 64, 68, 72, 76, 80, 84 e 88 são apresentados o resultados com as melhores acurácias obtidas em cada algoritmo, enriquecida pelas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax geradas a partir das Listas_Syntax_100% e Listas_Bert_100%. Portanto, elas visam comparar as melhores acurácias obtidas pelos modelos enriquecidos pelas gBoEDs geradas a partir das listas construídas pelos métodos semiautomáticos.

Nas tabelas citadas no parágrafo anterior, os algoritmos também são apresentados de acordo com um nível decrescente de explicabilidade. Cada tabela está organizada de maneira que a linha em cinza corresponde ao melhor resultado obtido pelo modelos enriquecido pela Listas_Bert_100%. Na segunda coluna estão as melhores acurácias obtidas a partir dos modelos gerados via BoW. Na sequência, as melhores acurácias obtidas a partir dos modelos gerados pelo método de classificação semanticamente enriquecida incrementado pela representação gBoED_Freq, gBoED_Dist e gBoED_Syntax. Em cada umas delas, os resultados obtidos estão divididos em duas colunas. A primeira coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Syntax_100% (aquele cuja gBoED é construída a partir de listas de termos formadas pelo método de extração baseado em regras morfossintáticas em 100% dos documentos da coleção). A segunda coluna de cada representação contém os resultados do experimento aplicado no cenário Listas_Bert_100%. A distribuição das informações pelas tabelas ocorre da mesma forma que no primeiro conjunto de tabelas com melhores acurácias.

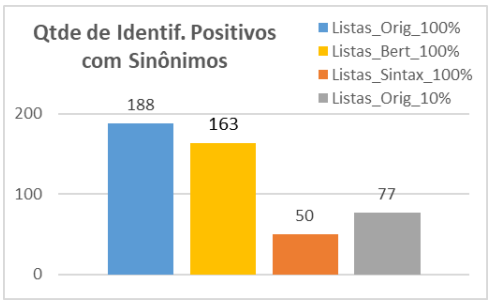
Iniciando a apresentação dos resultados são analisadas as resultados as quantidades de termos extraídas para a coleção de documentos *B2W Reviews 2019 Info*. Nos gráficos da Figura 49 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 49a pode-se observar as quantidades de termos do domínio, na Figura 49b observa-se as quantidades para identificadores da classe “Positivo” e na Figura 49c observa-se as quantidades para identificadores da classe “Negativo”. É possível verificar na Figura 49a que a quantidade de termos do domínio extraída pelas regras morfossintáticas é superior à quantidade de termos das Listas_Orig_100%. Já nas Figuras 49b e 49c verifica-se que a quantidade de identificadores da classe positiva e negativa é bastante próxima à quantidade de termos das Listas_Orig_100%.

Nas Tabelas 51 e 52 é apresentada a representatividade para o conjunto de documentos *B2W Reviews 2019 Info*. Na Tabela 51 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 52 é feita a comparação entre a representatividade das gBoEDs construídas a partir

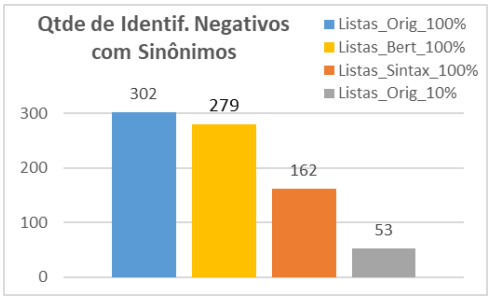
Figura 49 – Gráficos de quantidade de termos por tipo de lista - Coleção *B2W Reviews 2019 Info*.



(a) Termos do Domínio



(b) Identificadores de Classe “Positivo”



(c) Identificadores de Classe “Negativo”

Fonte: Elaborada pelo autor.

da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos, as representações formadas a partir das Listas_Bert_100% obtiveram níveis intermediários de representatividade. Em todos os casos a quantidade de documentos representados foi menor do que Listas_Orig_100% e Listas_Syntax_100%, e maior do que Listas_Orig_10%. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 90,70% dos documentos representados, Listas_Orig_100% obteve 92%, Listas_Syntax_-100% obteve 92,50% e Listas_Orig_10% obteve 86,50%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 80,10% dos documentos representados, Listas_Orig_100% obteve 85,30%, Listas_Syntax_100% obteve 86,50% e Listas_Orig_10% obteve 79,10%.

Com relação à quantidade de acertos na predição, na representação gBoED_Freq, Listas_-Bert_100% obteve 62,50%, Listas_Orig_100% obteve 65,30%, Listas_Syntax_100% obteve 63,70% e Listas_Orig_10% obteve 56,89%. Na representação gBoED_Dist, Listas_Bert_100% obteve 65,80%, Listas_Orig_100% obteve 68,10%, Listas_Syntax_100% obteve 67,70% e Listas_Orig_10% obteve 60%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 53,50%, Listas_Orig_100% obteve 60,80%, Listas_Syntax_100% obteve 60,20% e Listas_Orig_-10% obteve 55,70%. De maneira geral, o método de extração de termos usando BERT, para essa coleção de documentos, permitiu extrair uma quantidade maior de termos a partir de 10% de documentos anotados, porém as expressões de domínio formadas não foram tão representativas.

Nas Tabelas 53 e 54 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *B2W Reviews 2019 Info*. Em ambas, os resultados de maior

Tabela 51 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *B2W Reviews 2019 Info*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	920	92%	907	90,70%	865	86,50%	920	92%	907	90,70%	865	86,50%	853	85,30%	801	80,10%	791	79,10%
Qtde de documentos sem representação	80	8%	93	9,30%	135	13,50%	80	8%	93	9,30%	135	13,50%	147	14,70%	199	19,90%	209	20,90%
Número de ACERTOS na predição	653	65,30%	625	62,50%	569	56,89%	681	68,10%	658	65,80%	600	60%	608	60,80%	535	53,50%	557	55,70%
Número de ERROS na predição	191	19,10%	172	17,20%	216	21,60%	163	16,30%	130	13%	185	18,50%	165	16,50%	252	25,20%	221	22,10%
Número de NEUTROS na predição	156	15,60%	203	20,30%	215	21,51%	156	15,60%	212	21,20%	215	21,50%	227	22,70%	213	21,30%	222	22,20%

Fonte: Elaborada pelo autor.

Tabela 52 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *B2W Reviews 2019 Info*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	925	92,50%	907	90,70%	925	92,50%	907	90,70%	865	86,50%	801	80,10%
Qtde de documentos sem representação	75	7,50%	93	9,30%	75	7,50%	93	9,30%	135	13,50%	199	19,90%
Número de ACERTOS na predição	637	63,70%	625	62,50%	677	67,70%	658	65,80%	602	60,20%	535	53,50%
Número de ERROS na predição	162	16,20%	172	17,20%	113	11,30%	130	13%	138	13,80%	252	25,20%
Número de NEUTROS na predição	201	20,10%	203	20,30%	210	21%	212	21,20%	260	26%	213	21,30%

Fonte: Elaborada pelo autor.

destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF**. Para esse algoritmo os melhores resultados foram obtidos pelo enriquecimento a partir da representação gBoED_Syntax construída a partir das Listas_Bert_100%, com 92,200% de acurácia, Medida-F1 de 92,200%, $\gamma = 10^{-1}$, grau de confiança de 50%, com 3 documentos sendo consultados e 2 documentos reclassificados. Considerando outros resultados enriquecidos pelas Listas_-Bert_100%, vale destacar o modelo obtido por C4.5-Gini com gBoED_Syntax que obteve a acurácia de 89,000%, superando Listas_Orig_100% e Listas_Syntax_100% e superando BoW em aproximadamente 1,5%.

Em seguida são analisados os resultados para a coleção de documentos *Best Sports Top 4*. Nos gráficos da Figura 50 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_-Syntax_100% e Listas_Orig_10%. No gráfico da Figura 50a pode-se observar as quantidades de termos do domínio, na Figura 50b observa-se as quantidades para identificadores da classe "Brasileiro venceu" e na Figura 50c observa-se as quantidades para identificadores da classe "Brasileiro não venceu". É possível verificar na Figura 50a que a quantidade de termos do do-

Tabela 53 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *B2W Reviews 2019 Info*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,88600	0,89000 33-14	65%	0,88700 12-6	60%	0,88700 12-6	60%	0,89000 33-14	65%	0,88700 12-6	60%	0,88700 12-6	60%	0,89100 157-55	85%	0,88700 12-5	60%	0,88700 12-5	60%
C4.5-Gini	0,87500	0,87900 175-57	90%	0,87600 1-1	60%	0,87600 1-1	60%	0,87900 175-57	90%	0,87600 1-1	60%	0,87600 1-1	60%	0,88100 225-69	95%	0,89000 157-58	85%	0,87600 1-1	60%
KNN-Cosseno	0,90100	0,90200 66-40 $n = 35$	55%	0,89300 46-31 $n = 35$	55%	0,89300 46-28 $n = 35$	55%	0,90200 66-40 $n = 35$	55%	0,90000 46-26 $n = 35$	55%	0,89300 46-28 $n = 35$	55%	0,90800 128-67 $n = 45$	60%	0,89700 46-25 $n = 45$	55%	0,89700 46-25 $n = 45$	55%
KNN-Euclideana	0,89700	0,89900 67-40 $n = 35$	55%	0,89300 47-34 $n = 35$	55%	0,89000 47-29 $n = 35$	55%	0,89900 67-40 $n = 35$	55%	0,89300 47-34 $n = 35$	55%	0,89000 47-29 $n = 35$	55%	0,90500 70-45 $n = 35$	60%	0,89900 47-30 $n = 35$	55%	0,89400 47-26 $n = 35$	55%
MNB	0,90400	0,89700 50-36 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89200 50-40 $\alpha = 1$	55%	0,89700 50-36 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89200 50-40 $\alpha = 1$	55%	0,90700 123-64 $\alpha = 1$	60%	0,90300 50-41 $\alpha = 1$	55%	0,89700 50-35 $\alpha = 1$	55%
SVM-Linear	0,90900	0,91300 51-33 $\gamma = 10^{-4}$ a 10^4	60%	0,91000 3-1 $\gamma = 10^{-4}$ a 10^4	50%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91300 51-33 $\gamma = 10^{-4}$ a 10^4	60%	0,91000 3-1 $\gamma = 10^{-4}$ a 10^4	50%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91700 73-40 $\gamma = 10^{-4}$ a 10^4	65%	0,91600 73-37 $\gamma = 10^{-4}$ a 10^4	65%	0,90900 3-2 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,91500	0,91500 18-14 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91500 18-14 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91900 79-44 $\gamma = 10$	65%	0,91600 18-16 $\gamma = 10$	55%	0,91400 1-1 $\gamma = 10$	50%
SVM-RBF	0,92200	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

Tabela 54 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *B2W Reviews 2019 Info*.

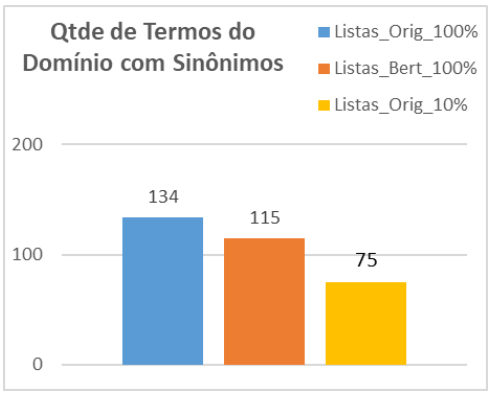
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,88600	0,88800 12-6	60%	0,88700 12-6	60%	0,88800 12-6	60%	0,88700 12-6	60%	0,88900 12-7	65%	0,88700 12-5	60%
C4.5-Gini	0,87500	0,87700 1-1	60%	0,87600 1-1	60%	0,87700 1-1	60%	0,87600 1-1	60%	0,87700 1-1	60%	0,89000 157-58	85%
KNN-Cosseno	0,90100	0,89500 46-32 $n = 35$	55%	0,89300 46-31 $n = 35$	55%	0,90100 46-26 $n = 35$	55%	0,90000 46-26 $n = 35$	55%	0,89800 46-27 $n = 35$	55%	0,89700 46-25 $n = 45$	55%
KNN-Euclideana	0,89700	0,89500 47-33 $n = 35$	55%	0,89300 47-34 $n = 35$	55%	0,89500 47-33 $n = 35$	55%	0,89300 47-34 $n = 35$	55%	0,90300 47-36 $n = 35$	55%	0,89900 47-30 $n = 35$	55%
MNB	0,90400	0,89500 50-45 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,89500 50-45 $\alpha = 1$	55%	0,90400 50-42 $\alpha = 1$	55%	0,90300 50-41 $\alpha = 1$	55%
SVM-Linear	0,90900	0,91100 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91000 3-1 $\gamma = 10^{-4}$ a 10^4	50%	0,91100 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,91000 3-1 $\gamma = 10^{-4}$ a 10^4	50%	0,91400 73-33 $\gamma = 10^{-4}$ a 10^4	65%	0,91600 73-37 $\gamma = 10^{-4}$ a 10^4	65%
SVM-Polinomial	0,91500	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91400 1-1 $\gamma = 10$	50%	0,91900 79-44 $\gamma = 10$	65%	0,91600 18-16 $\gamma = 10$	55%
SVM-RBF	0,92200	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92200 3-3 $\gamma = 10^{-1}$	50%	0,92100 3-3 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%	0,92200 3-2 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

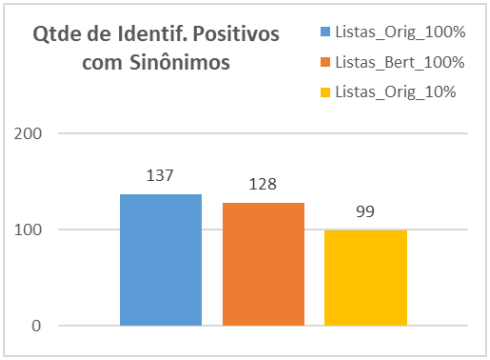
mínio extraída pelo método BERT atinge uma quantidade intermediária de termos em relação às Listas_Orig_100% e Listas_Orig_10%. Da mesma forma, na Figura 50b verifica-se que a quantidade de identificadores da classe positiva extraída pelo método BERT atinge uma quantidade intermediária de termos em relação às Listas_Orig_100% e Listas_Orig_10%, porém mais

próxima das Listas_Orig_100%. Já na Figura 50c verifica-se que a quantidade de identificadores da classe negativa é maior do que aqueles das Listas_Orig_100% e Listas_Orig_10%.

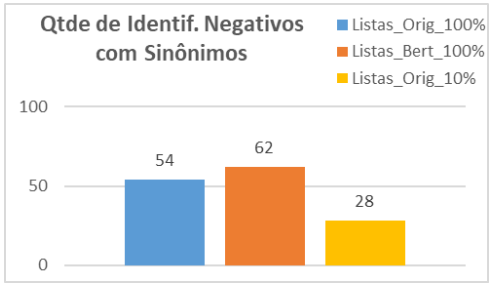
Figura 50 – Gráficos de quantidade de termos por tipo de lista - Coleção *Best Sports Top 4*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Na Tabela 55 é apresentada a representatividade para o conjunto de documentos *Best Sports Top 4* e a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Como a coleção de documentos *Best Sports Top 4* não foi utilizada nos experimentos realizados com o método de extração de termos baseado em regras morfossintáticas e as regras morfossintáticas que geram a representação semanticamente enriquecida gBoED_Syntax não foram desenvolvidas para este domínio, nesta coleção não há resultados para gBoED_Syntax e Listas_Syntax_100%.

Nesta tabela é possível observar que, para essa coleção de documentos, as representações formadas a partir das Listas_Bert_100% obtiveram níveis semelhantes de representatividade aos das Listas_Orig_100%. Em todos os casos a quantidade de documentos representados foi igual aos das Listas_Orig_100%. Em ambas das representações, Listas_Bert_100% e Listas_Orig_-100% obteveram 89,50% dos documentos representados e Listas_Orig_10% obteve 88,95%. Com relação à quantidade de acertos na predição, na representação gBoED_Freq, Listas_Bert_-100% obteve 48,07%, Listas_Orig_100% obteve 52,48% e Listas_Orig_10% obteve 50,27%. Na representação gBoED_Dist, Listas_Bert_100% obteve 54,14%, Listas_Orig_100% obteve 56,35% e Listas_Orig_10% obteve 53,59%. Em todas as representações construídas a partir das

Listas_Bert_100%, a taxa de erros foi próxima às das outras listas.

Tabela 55 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *Best Sports Top 4*.

	gBoED_Freq						**gBoED_Dist**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	162	89,50%	162	89,50%	161	88,95%	162	89,50%	162	89,50%	161	88,95%
Qtde de documentos sem representação	19	10,49%	19	10,49%	20	11,04%	19	10,49%	19	10,49%	20	11,04%
Número de ACERTOS na predição	95	52,48%	87	48,07%	91	50,27%	102	56,35%	98	54,14%	97	53,59%
Número de ERROS na predição	54	29,83%	52	28,73%	54	29,83%	47	25,96%	52	28,73%	48	26,51%
Número de NEUTROS na predição	32	15,60%	42	23,20%	36	19,88%	32	17,67%	31	17,13%	36	19,88%

Fonte: Elaborada pelo autor.

Na Tabela 56 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *Best Sports Top 4*. De forma semelhante à Tabela 55, como a coleção de documentos *Best Sports Top 4* não foi utilizada nos experimentos realizados com o método de extração de termos baseado em regras morfossintáticas e as regras morfossintáticas que geram a representação semanticamente enriquecida gBoED_Syntax não foram desenvolvidas para este domínio, nesta coleção não há resultados para gBoED_Syntax e Listas_Syntax_100%. Ao observar esta tabela é possível observar que, para esta coleção de documentos, o uso do método de extração de termos baseado em BERT não obteve boa representação e consequentemente bons resultados de acurácia. Em sua maioria, os resultados obtidos pelas representações construídas a partir das Listas_Bert_100% atingiram valores abaixo ou semelhantes à BoW. Dentre o conjunto de resultados, é possível destacar os modelos obtidos por C4.5-Entropia com gBoED_Dist, atingiu uma acurácia de 53,922%, valor intermediário entre Listas_Orig_100% e Listas_Orig_10%.

Em seguida são analisados os resultados para a coleção de documentos *HuLiu 2004*. Nos gráficos da Figura 51 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 51a pode-se observar as quantidades de termos do domínio, na Figura 51b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 51c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 51a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT é bastante superior à quantidade de termos das Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. Na Figura 51b verifica-se que para os identificadores da classe positiva, a quantidade de termos foi de aproximadamente $1/3$ da quantidade de termos das Listas_Orig_100%, porém bastante superior às Listas_Syntax_100% e Listas_Orig_10%. Na Figura 51c verifica-se que a quantidade de identificadores da

Tabela 56 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *Best Sports Top 4*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,51871	0,52953 19-4	60%	0,52083 19-4	60%	0,53508 46-7	90%	0,56432 181-85	100%	0,53922 181-81	100%	0,53654 181-82	100%
C4.5-Gini	0,58596	0,59707 52-12	95%	0,59707 52-12	95%	0,60263 52-11	95%	0,60818 52-11	95%	0,60797 52-11	95%	0,61374 52-10	95%
KNN-Cosseno	0,70643	0,67368 14-7 $n = 25$	55%	0,67233 14-6 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,67233 14-6 $n = 25$	55%	0,67368 14-7 $n = 25$	55%
KNN-Euclideana	0,70643	0,67368 14-7 $n = 25$	55%	0,67233 14-6 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,67368 14-7 $n = 25$	55%	0,67233 14-6 $n = 25$	55%	0,67368 14-7 $n = 25$	55%
MNB	0,69064	0,67368 22-10 $\alpha = 1$	55%	0,67233 22-9 $\alpha = 1$	55%	0,67368 16-12 $\alpha = 1$	60%	0,67368 22-11 $\alpha = 1$	55%	0,67233 22-9 $\alpha = 1$	55%	0,67368 22-11 $\alpha = 1$	55%
SVM-Linear	0,71315	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,70934 4-1 $\gamma = 10^{-4}$ a 10^4	50%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%	0,70934 4-1 $\gamma = 10^{-4}$ a 10^4	50%	0,71315 4-0 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,71286	0,70730 2-1 $\gamma = 10^{-1}$	50%	0,70580 2-2 $\gamma = 10^{-1}$	50%	0,71286 2-0 $\gamma = 10^{-1}$	50%	0,70730 2-1 $\gamma = 10^{-1}$	50%	0,70580 2-2 $\gamma = 10^{-1}$	50%	0,71286 2-0 $\gamma = 10^{-1}$	50%
SVM-RBF	0,70175	0,69619 18-5 $\gamma = 1$	55%	0,69619 18-5 $\gamma = 1$	55%	0,69619 18-5 $\gamma = 1$	55%	0,69619 18-4 $\gamma = 1$	55%	0,69619 18-4 $\gamma = 1$	55%	0,69619 18-4 $\gamma = 1$	55%

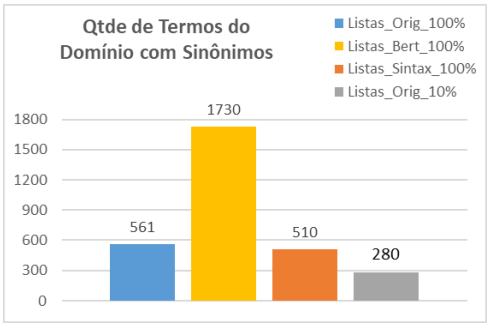
Fonte: Elaborada pelo autor.

classe negativa segue uma quantidade semelhante às listas da classe positiva. Listas_Bert_100% atingem valores inferiores Listas_Orig_100%, porém bastante superior às Listas_Syntax_100% e Listas_Orig_10%.

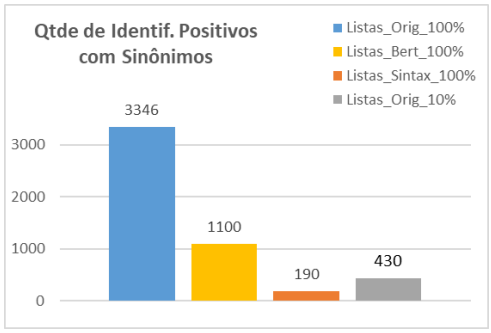
Nas Tabelas 57 e 58 é apresentada a representatividade para o conjunto de documentos *HuLiu 2004*. Na Tabela 57 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 58 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que para essa coleção de documentos as representações formadas a partir das Listas_Bert_100% obtiveram níveis baixos de representatividade. Em todos os casos a quantidade de documentos representados foi menor do que Listas_Orig_100%, Listas_Syntax_100% e, também, Listas_Orig_10%. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 92,57% dos documentos representados, Listas_Orig_100% obteve 96,28%, Listas_Syntax_100% obteve 97,98% e Listas_Orig_10% obteve 94,60%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 88,18% dos documentos representados, Listas_Orig_100% obteve 91,90%, Listas_Syntax_100% obteve 92,56% e Listas_Orig_10% obteve 88,18%, empatando com Listas_Bert_100%.

Com relação à quantidade de acertos na predição, na representação gBoED_Freq, Listas_-Bert_100% obteve 62,50%, Listas_Orig_100% obteve 65,30%, Listas_Syntax_100% obteve

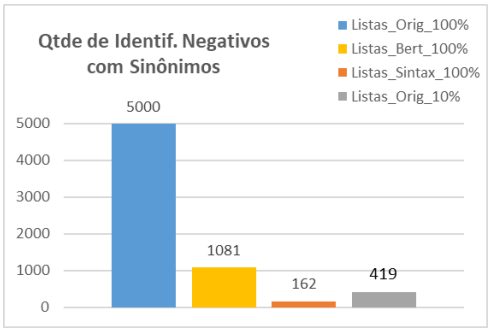
Figura 51 – Gráficos de quantidade de termos por tipo de lista - Coleção *HuLiu 2004*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

63,70% e Listas_Orig_10% obteve 56,89%. Na representação gBoED_Dist, Listas_Bert_100% obteve 65,80%, Listas_Orig_100% obteve 68,10%, Listas_Syntax_100% obteve 67,70% e Listas_Orig_10% obteve 60%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 53,50%, Listas_Orig_100% obteve 60,80%, Listas_Syntax_100% obteve 60,20% e Listas_Orig_-10% obteve 55,70%. De maneira geral, o método de extração de termos usando BERT, para essa coleção de documentos, permitiu extrair uma quantidade maior de termos a partir de 10% de documentos anotados, porém as expressões de domínio formadas não foram tão representativas.

Tabela 57 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *HuLiu 2004*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	285	96,28%	274	92,57%	280	94,60%	285	96,28%	274	92,57%	280	94,60%	272	91,90%	261	88,18%	261	88,18%
Qtde de documentos sem representação	11	3,72%	22	7,43%	16	5,40%	11	3,72%	22	7,43%	16	5,40%	24	8,10%	35	11,82%	35	11,82%
Número de ACERTOS na predição	218	73,64%	196	66,21%	215	72,64%	220	74,32%	197	66,55%	220	74,32%	202	68,24%	175	59,13%	189	63,86%
Número de ERROS na predição	57	19,26%	74	25%	54	18,24%	55	18,58%	73	24,66%	49	16,56%	39	13,18%	78	26,35%	46	15,54%
Número de NEUTROS na predição	21	7,10%	26	8,79%	27	9,12%	21	7,10%	26	8,79%	27	9,12%	55	18,58%	43	14,52%	61	20,60%

Fonte: Elaborada pelo autor.

Nas Tabelas 59 e 60 são apresentadas as melhores acurácias obtidas em cada algoritmo

Tabela 58 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *HuLiu 2004*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	290	97,98%	274	92,57%	290	97,98%	274	92,57%	274	92,56%	261	88,18%
Qtde de documentos sem representação	6	2,02%	22	7,43%	6	2,02%	22	7,43%	22	7,44%	35	11,82%
Número de ACERTOS na predição	211	71,28%	196	66,21%	213	71,96%	197	66,55%	197	66,56%	175	59,13%
Número de ERROS na predição	73	24,66%	74	25%	71	23,98%	73	24,66%	47	15,88%	78	26,35%
Número de NEUTROS na predição	12	4,06%	26	8,79%	12	4,06%	26	8,79%	52	17,56%	43	14,52%

Fonte: Elaborada pelo autor.

aplicado à coleção de documentos *HuLiu 2004*. Em ambas, os resultados de maior destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel Polinomial**. Para esse algoritmo os melhores resultados foram obtidos pelo enriquecimento a partir da representação gBoED_Dist construída a partir das Listas_Orig_100%, com 85,483% de acurácia, seguida por gBoED_Freq Listas_Orig_100% com 85,816% de acurácia. Em todas as representações construídas com Listas_Bert_100% a acurácia atingiu um resultado intermediário de 84,872%, Medida-F1 de 84,872%, $\gamma = 10^{-1}$, grau de confiança de 50%, com 3 documentos sendo consultados e 1 documentos reclassificados. Isso significa que, mesmo não obtendo o melhor nível de representatividade no conjunto de termos extraídos, as representações construídas a partir das Listas_Bert_100% para a coleção de documentos *HuLiu 2004* obteve uma contribuição interessante, superando a BoW e aproximando-se dos resultados obtidos pelo cenário Listas_Orig_100%, considerado o limite superior.

Considerando outros resultados enriquecidos pelas Listas_Bert_100%, vale destacar o modelo obtido por C4.5-Gini com gBoED_Syntax que obteve a acurácia de 72,632%, superando Listas_Orig_100% e Listas_Syntax_100%, e superando BoW em aproximadamente 4%. Outros destaques são os modelos que obtiveram resultados intermediários como C4.5-Entropia (com gBoED_Freq e gBoED_Syntax) e MNB (com gBoED_Freq, gBoED_Dist).

Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2014*. Nos gráficos da Figura 52 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_-Syntax_100% e Listas_Orig_10%. No gráfico da Figura 52a pode-se observar as quantidades de termos do domínio, na Figura 52b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 52c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 52a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT aproxima-se da quantidade de

Tabela 59 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *HuLiu 2004*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,68988	0,74057 86-34	90%	0,73678 74-49	70%	0,72678 296-118	100%	0,74390 86-32	90%	0,71977 74-50	70%	0,74345 296-109	100%	0,70218 81-50	85%	0,69908 38-29	65%	0,68207 81-49	85%
C4.5-Gini	0,67609	0,73701 296-114	100%	0,73655 95-24	90%	0,72678 296-111	100%	0,74379 296-118	100%	0,72966 72-22	100%	0,74345 296-106	100%	0,68874 46-23	85%	0,72632 1-1	600%	0,68207 46-23	85%
KNN-Cosseno	0,78712	0,80781 121-57 $n = 9$	70%	0,77747 65-36 $n = 9$	60%	0,79402 63-35 $n = 9$	55%	0,79436 121-55 $n = 9$	70%	0,77747 65-34 $n = 9$	60%	0,79391 63-31 $n = 9$	55%	0,77701 59-37 $n = 13$	60%	0,75989 26-14 $n = 25$	55%	0,77368 59-37 $n = 13$	60%
KNN-Euclideana	0,78712	0,80781 121-57 $n = 9$	70%	0,77747 65-36 $n = 9$	60%	0,79402 63-35 $n = 9$	55%	0,79436 121-55 $n = 9$	70%	0,77747 65-34 $n = 9$	60%	0,79391 63-31 $n = 9$	55%	0,77701 59-37 $n = 13$	60%	0,75989 26-14 $n = 25$	55%	0,77368 59-37 $n = 13$	60%
MNB	0,81402	0,83793 48-29 $\alpha = 10^{-2}$	70%	0,82782 19-9 $\alpha = 10^{-1}$	55%	0,82115 14-5 $\alpha = 10^{-2}$	55%	0,83126 48-31 $\alpha = 10^{-2}$	70%	0,82448 19-10 $\alpha = 10^{-1}$	55%	0,81770 14-8 $\alpha = 10^{-2}$	55%	0,81770 15-9 $\alpha = 10^{-2}$	55%	0,81092 19-14 $\alpha = 10^{-1}$	55%	0,81425 15-10 $\alpha = 10^{-2}$	55%
SVM-Linear	0,82793	0,85161 66-30 $\gamma = 10^{-4}$ a 10^4	70%	0,83816 49-26 $\gamma = 10^{-4}$ a 10^4	65%	0,84149 62-27 $\gamma = 10^{-4}$ a 10^4	70%	0,85149 47-21 $\gamma = 10^{-4}$ a 10^4	65%	0,83782 20-12 $\gamma = 10^{-4}$ a 10^4	55%	0,83804 62-26 $\gamma = 10^{-4}$ a 10^4	70%	0,84103 29-16 $\gamma = 10^{-4}$ a 10^4	60%	0,82747 2-2 $\gamma = 10^{-4}$ a 10^4	55%	0,83414 1-1 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,83436	0,85816 59-21 $\gamma = 10^{-1}$	70%	0,84872 3-1 $\gamma = 10^{-1}$	50%	0,84827 60-17 $\gamma = 10^{-1}$	70%	0,85483 48-21 $\gamma = 10^{-1}$	65%	0,84782 3-1 $\gamma = 10^{-1}$	50%	0,84149 59-20 $\gamma = 10^{-1}$	70%	0,84115 32-18 $\gamma = 10^{-1}$	60%	0,84782 3-1 $\gamma = 10^{-1}$	50%	0,83115 32-22 $\gamma = 10^{-1}$	60%
SVM-RBF	0,81425	0,84149 72-35 $\gamma = 1$	70%	0,82793 34-15 $\gamma = 1$	60%	0,83839 69-33 $\gamma = 1$	70%	0,83805 72-29 $\gamma = 1$	70%	0,82448 10-5 $\gamma = 1$	55%	0,82816 69-29 $\gamma = 1$	70%	0,83448 22-15 $\gamma = 1$	55%	0,82103 10-7 $\gamma = 1$	55%	0,82770 22-16 $\gamma = 1$	55%

Fonte: Elaborada pelo autor.

Tabela 60 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *HuLiu 2004*.

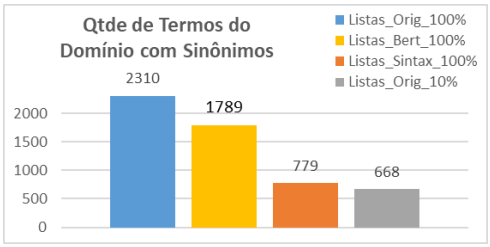
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,68988	0,72264 81-36	80%	0,73678 74-49	70%	0,72276 81-35	80%	0,71977 74-50	70%	0,71253 73-43	75%	0,69908 38-29	65%
C4.5-Gini	0,67609	0,71264 296-96	100%	0,73655 95-24	90%	0,71954 296-97	100%	0,72966 72-22	100%	0,69897 78-39	90%	0,72632 1-1	600%
KNN-Cosseno	0,78712	0,77701 63-33 $n = 11$	60%	0,77747 65-36 $n = 9$	60%	0,77713 64-34 $n = 11$	60%	0,77747 65-34 $n = 9$	60%	0,76356 33-24 $n = 25$	55%	0,75989 26-14 $n = 25$	55%
KNN-Euclideana	0,78712	0,77701 63-33 $n = 11$	60%	0,77747 65-36 $n = 9$	60%	0,77713 64-34 $n = 11$	60%	0,77747 65-34 $n = 9$	60%	0,76356 33-24 $n = 25$	55%	0,75989 26-14 $n = 25$	55%
MNB	0,81402	0,81437 14-6 $\alpha = 10^{-1}$	55%	0,82782 19-9 $\alpha = 10^{-1}$	55%	0,81782 14-5 $\alpha = 10^{-1}$	55%	0,82448 19-10 $\alpha = 10^{-1}$	55%	0,81092 15-8 $\alpha = 10^{-2}$	55%	0,81092 19-14 $\alpha = 10^{-1}$	55%
SVM-Linear	0,82793	0,83414 30-13 $\gamma = 10^{-4}$ a 10^4	60%	0,83816 49-26 $\gamma = 10^{-4}$ a 10^4	65%	0,83414 30-15 $\gamma = 10^{-4}$ a 10^4	60%	0,83782 20-12 $\gamma = 10^{-4}$ a 10^4	55%	0,83414 1-1 $\gamma = 10^{-4}$ a 10^4	55%	0,82747 2-2 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,83436	0,84092 4-3 $\gamma = 10^{-1}$	50%	0,84872 3-1 $\gamma = 10^{-1}$	50%	0,84425 4-2 $\gamma = 10^{-1}$	50%	0,84782 3-1 $\gamma = 10^{-1}$	50%	0,83448 32-24 $\gamma = 10^{-1}$	60%	0,84782 3-1 $\gamma = 10^{-1}$	50%
SVM-RBF	0,81425	0,82759 36-18 $\gamma = 10^1$	60%	0,82793 34-15 $\gamma = 1$	60%	0,82747 20-8 $\gamma = 1$	55%	0,82448 10-5 $\gamma = 1$	55%	0,83448 22-14 $\gamma = 1$	55%	0,82103 10-7 $\gamma = 1$	55%

Fonte: Elaborada pelo autor.

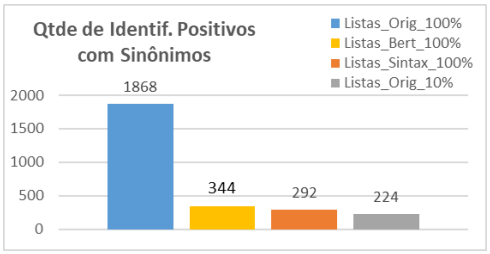
termos contida nas Listas_Orig_100% e supera a quantidade de termos em Listas_Syntax_100% e Listas_Orig_10%. Na Figura 52b verifica-se que para os identificadores da classe positiva, a quantidade é levemente superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima das Listas_Orig_100%. Na Figura 52c verifica-se que a quantidade de identificadores da classe negativa é superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima

das Listas_Orig_100%.

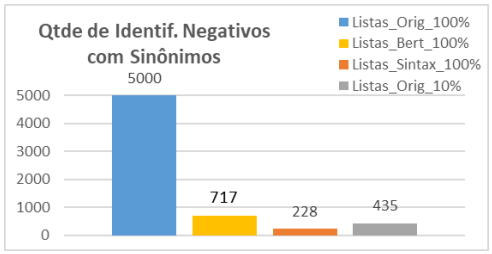
Figura 52 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Nas Tabelas 61 e 62 é apresentada a representatividade para o conjunto de documentos *SemEval 2014*. Na Tabela 61 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 62 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Vale lembrar que o conjunto de documentos *SemEval 2014* corresponde à união de dois domínios diferentes: *Laptops* e *Restaurants*. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram níveis intermediários de representatividade, enquanto gBoED_Syntax formada a partir das Listas_Bert_100% obtive nível baixo de representatividade. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 73,20% dos documentos representados, Listas_Orig_100% obteve 77,10%, Listas_Syntax_100% obteve 72,13% e Listas_Orig_10% obteve 61,46%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 26,98% dos documentos representados, Listas_Orig_100% obteve 43,82%, Listas_Syntax_100% obteve 41,25% e Listas_Orig_10% obteve 33,10%, empatando com Listas_Bert_100%.

Com relação à quantidade de acertos na predição, na representação gBoED_Dist, Listas_Bert_100% obteve 55,03%, Listas_Orig_100% obteve 64,82%, Listas_Syntax_100% obteve 62,96% e Listas_Orig_10% obteve 51,14%. Na representação gBoED_Freq, Listas_Bert_100% obteve 52,10%, Listas_Orig_100% obteve 61,04%, Listas_Syntax_100% obteve 60,68% e Listas_Orig_10% obteve 49,35%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 22,30%, Listas_Orig_100% obteve 38,14%, Listas_Syntax_100% obteve 41,25% e Listas_Orig_-

Tabela 61 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	1286	77,10%	1221	73,20%	1025	61,46%	1286	77,10%	1221	73,20%	1025	61,46%	731	43,82%	450	26,98%	552	33,10%
Qtde de documentos sem representação	382	22,90%	447	26,80%	643	38,54%	382	22,90%	447	26,80%	643	38,54%	937	56,18%	1218	73,02%	1116	66,90%
Número de ACERTOS na predição	1018	61,04%	869	52,10%	823	49,35%	1081	64,82%	918	55,03%	853	51,14%	636	38,14%	372	22,30%	482	28,90%
Número de ERROS na predição	157	9,41%	257	15,40%	140	8,39%	94	5,63%	208	12,47%	110	6,59%	75	4,49%	75	4,50%	62	3,72%
Número de NEUTROS na predição	493	29,55%	542	32,50%	705	42,26%	493	29,55%	542	32,49%	705	42,26%	957	57,37%	1221	73,20%	1124	67,38%

Fonte: Elaborada pelo autor.

10% obteve 28,90%. Em todas as representações construídas a partir das Listas_Bert_100%, observa-se que mesmo extraindo uma maior quantidade de termos em relação ao método baseado em análise morfossintática, para essa coleção de documentos, os termos não foram tão representativos.

Tabela 62 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2014*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	1203	72,13%	1221	73,20%	1203	72,13%	1221	73,20%	688	41,25%	450	26,98%
Qtde de documentos sem representação	465	27,87%	447	26,80%	465	27,87%	447	26,80%	980	58,75	1218	73,02%
Número de ACERTOS na predição	1012	60,68%	869	52,10%	1050	62,96%	918	55,03%	616	36,94%	372	22,30%
Número de ERROS na predição	119	7,13%	257	15,40%	81	4,85%	208	12,47%	62	3,71%	75	4,50%
Número de NEUTROS na predição	537	32,19%	542	32,50%	537	32,19%	542	32,49%	990	59,35%	1221	73,20%

Fonte: Elaborada pelo autor.

Nas Tabelas 63 e 64 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014*. Em ambas, os resultados de maior destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, enriquecido pela representação gBoED_Dist**. Nesse cenário os resultados obtidos foram de 79,993% de acurácia, Medida-F1 de 79,993%, $\gamma = 1$, grau de confiança de 55%, com 181 documentos sendo consultados e 103 documentos reclassificados. Para esse mesmo modelo, o uso de Listas_Bert_100% atingiu um valor menor do que Listas_Orig_100% e Listas_Syntax_100%. Outro cenário de destaque para Listas_Bert_100% é o resultado do modelos obtido por ***Support Vector Machine* – SVM, kernel Polinomial, enriquecido pela representação gBoED_Dist**. Nesse cenário os resultados obtidos foram de 79,751% de acurácia, Medida-F1 de 79,631%, $\gamma = 10^{-1}$, grau de confiança de 50%, com 20 documentos sendo consultados e 12 documentos reclassificados. Para

esse mesmo modelo, o uso de Listas_Bert_100% superou Listas_Orig_100% em gBoED_Freq com 79,682% de acurácia, e gBoED_Syntax com 79,442% de acurácia.

Outros resultados enriquecidos pelas Listas_Bert_100%, que valem ser destacados são os modelos obtidos por C4.5-Entropia gBoED_Dist que obteve a acurácia de 67,103%, superando Listas_Orig_100% e Listas_Syntax_100%, e superando BoW em aproximadamente 4%. Outros destaques são os modelos que obtiveram resultados intermediários como C4.5-Gini (com gBoED_Dist), KNN-Cosseno e KNN-Euclidiana (com gBoED_Dist) e MNB (com gBoED_Dist). Esses modelos possuem um maior grau de explicabilidade e obtiveram melhores resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT.

Tabela 63 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,63700	0,63871 2082-1239	55%	0,61053 2088-1321	75%	0,54348 2088-1390	55%	0,66759 2082-1153	55%	0,67103 2088-1143	75%	0,56032 2088-1331	55%	0,63904 1-1	55%	0,37642 2084-1805	55%	0,63904 1-1	50%
C4.5-Gini	0,64525	0,65005 1235-707	55%	0,61499 1649-1050	55%	0,57818 27-12	55%	0,66587 1235-654	55%	0,66896 2057-1112	60%	0,59090 27-12	55%	0,51459 1432-1142	55%	0,43483 1649-1424	55%	0,46817 1432-1234	55%
KNN-Cosseno	0,77760	0,75973 317-214 $n = 17$	55%	0,76317 437-273 $n = 13$	55%	0,74355 40-29 $n = 17$	55%	0,76419 421-254 $n = 13$	55%	0,77451 437-245 $n = 13$	55%	0,74767 40-29 $n = 17$	55%	0,73015 342-290 $n = 17$	55%	0,73531 239-210 $n = 25$	55%	0,72258 342-306 $n = 17$	55%
KNN-Euclideana	0,77691	0,76076 326-217 $n = 17$	55%	0,76488 256-181 $n = 25$	55%	0,74871 39-28 $n = 17$	55%	0,76523 326-207 $n = 17$	55%	0,77416 453-258 $n = 13$	55%	0,75180 39-28 $n = 17$	55%	0,73668 269-233 $n = 25$	55%	0,73566 256-227 $n = 25$	55%	0,73083 269-242 $n = 25$	55%
MNB	0,80648	0,79823 141-100 $\alpha = 10^{-1}$	55%	0,79925 148-100 $\alpha = 10^{-1}$	55%	0,78584 24-19 $\alpha = 10^{-1}$	55%	0,80063 141-94 $\alpha = 10^{-1}$	55%	0,80269 148-90 $\alpha = 10^{-1}$	55%	0,78687 35-25 $\alpha = 10^{-1}$	55%	0,78068 149-125 $\alpha = 10^{-1}$	55%	0,78206 148-136 $\alpha = 10^{-1}$	55%	0,77793 149-130 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79718	0,79581 25-18 $\gamma = 10^{-4}$ a 10^4	50%	0,79477 16-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 17-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79615 25-17 $\gamma = 10^{-4}$ a 10^4	50%	0,79511 16-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79168 16-8 $\gamma = 10^{-4}$ a 10^4	50%	0,79064 16-16 $\gamma = 10^{-4}$ a 10^4	50%	0,79305 16-15 $\gamma = 10^{-4}$ a 10^4	50%	0,79030 16-16 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,78824	0,79615 19-13 $\gamma = 10^{-1}$	50%	0,79682 20-12 $\gamma = 10^{-1}$	50%	0,78053 17-14 $\gamma = 10^{-1}$	50%	0,79650 19-12 $\gamma = 10^{-1}$	50%	0,79751 20-12 $\gamma = 10^{-1}$	50%	0,78145 17-14 $\gamma = 10^{-1}$	50%	0,78892 14-13 $\gamma = 10^{-1}$	50%	0,79442 20-19 $\gamma = 10^{-1}$	50%	0,78858 14-14 $\gamma = 10^{-1}$	50%
SVM-RBF	0,79925	0,80235 8-6 $\gamma = 1$	50%	0,79821 18-15 $\gamma = 1$	50%	0,78045 17-14 $\gamma = 1$	50%	0,80407 185-140 $\gamma = 1$	55%	0,79993 181-103 $\gamma = 1$	55%	0,78045 17-14 $\gamma = 1$	50%	0,79683 17-17 $\gamma = 1$	50%	0,79752 18-17 $\gamma = 1$	50%	0,79683 17-17 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2014 Laptop*. Nos gráficos da Figura 53 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 54a pode-se observar as quantidades de termos do domínio, na Figura 53b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 53c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 53a que a quantidade de termos do domínio e identificadores da classe positiva extraída pelo método de extração de termos baseado em modelos de linguagem BERT aproxima-se da quantidade de termos contida nas Listas_Orig_100% e supera a quantidade de termos em Listas_Syntax_100% e Listas_Orig_10%. Na Figura 53c verifica-se que a quantidade de identificadores da classe negativa é levemente superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima das Listas_Orig_100%.

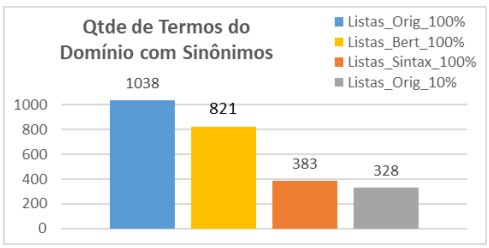
Nas Tabelas 65 e 66 é apresentada a representatividade para o conjunto de documentos

Tabela 64 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2014*.

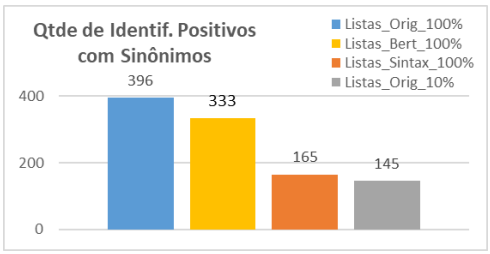
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,63700	0,5995 2088-1162	85%	0,61053 2088-1321	75%	0,62048 2088-1115	85%	0,67103 2088-1143	75%	0,45789 2089-1618	85%	0,37642 2084-1805	55%
C4.5-Gini	0,64525	0,61978 1432-780	55%	0,61499 1649-1050	55%	0,63663 1432-744	55%	0,66896 2057-1112	60%	0,52387 1432-1097	55%	0,43483 1649-1424	55%
KNN-Cosseno	0,77760	0,75386 340-248 $n = 17$	55%	0,76317 437-273 $n = 13$	55%	0,7590 340-238 $n = 17$	55%	0,77451 437-245 $n = 13$	55%	0,73255 340-294 $n = 17$	55%	0,73531 239-210 $n = 25$	55%
KNN-Euclideana	0,77691	0,75593 346-253 $n = 17$	55%	0,76488 256-181 $n = 25$	55%	0,75593 346-253 $n = 17$	55%	0,77416 453-258 $n = 13$	55%	0,73978 264-219 $n = 25$	55%	0,73566 256-227 $n = 25$	55%
MNB	0,80648	0,79409 153-104 $\alpha = 10^{-1}$	55%	0,79925 148-100 $\alpha = 10^{-1}$	55%	0,79546 153-101 $\alpha = 10^{-1}$	55%	0,80269 148-90 $\alpha = 10^{-1}$	55%	0,78343 153-126 $\alpha = 10^{-1}$	55%	0,78206 148-136 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79718	0,79305 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79477 16-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79305 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79511 16-13 $\gamma = 10^{-4}$ a 10^4	50%	0,79271 16-14 $\gamma = 10^{-4}$ a 10^4	50%	0,79305 16-15 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,78824	0,79442 14-11 $\gamma = 10^{-1}$	50%	0,79682 20-12 $\gamma = 10^{-1}$	50%	0,79442 14-11 $\gamma = 10^{-1}$	50%	0,79751 20-12 $\gamma = 10^{-1}$	50%	0,79236 14-12 $\gamma = 10^{-1}$	50%	0,79442 20-19 $\gamma = 10^{-1}$	50%
SVM-RBF	0,79925	0,80302 17-15 $\gamma = 1$	50%	0,79821 18-15 $\gamma = 1$	50%	0,80325 17-13 $\gamma = 1$	50%	0,79993 181-103 $\gamma = 1$	55%	0,80268 17-15 $\gamma = 1$	50%	0,79752 18-17 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

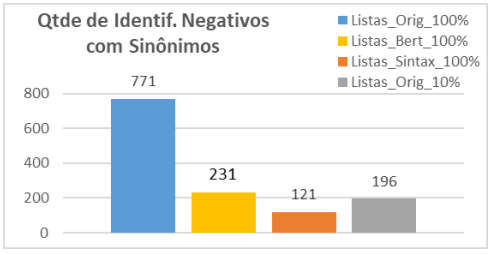
Figura 53 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014 Laptop*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

SemEval 2014 Laptop. Na Tabela 65 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 66 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_-Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa

coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram os níveis mais altos de representatividade, enquanto gBoED_-Syntax formada a partir das Listas_Bert_100% obtive o nível mais baixo de representatividade em todos os cenários. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 78,97% dos documentos representados, Listas_Orig_100% obteve 73,82%, Listas_Syntax_100% obteve 60,60% e Listas_Orig_10% obteve 50,52%. Na representação gBoED_Syntax, Listas_-Bert_100% obteve 19,33% dos documentos representados, Listas_Orig_100% obteve 32,48%, Listas_Syntax_100% obteve 32,48% e Listas_Orig_10% obteve 19,98%.

Tabela 65 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014 Laptop*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	916	73,82%	980	78,97%	627	50,52%	916	73,82%	980	78,97%	627	50,52%	403	32,48%	240	19,33%	248	19,98%
Qtde de documentos sem representação	325	26,18%	261	21,03%	614	49,48%	325	26,18%	261	21,03%	614	49,48%	838	67,52%	1001	80,66%	993	80,02%
Número de ACERTOS na predição	719	57,93%	540	43,51%	485	39,10%	765	61,64%	585	47,14%	503	40,54%	334	26,92%	173	13,94%	202	16,29%
Número de ERROS na predição	120	9,67%	382	30,78%	108	8,70%	74	5,96%	337	27,16%	90	7,25%	57	4,59%	65	5,28%	41	3,30%
Número de NEUTROS na predição	402	32,40%	319	25,70%	648	52,20%	402	32,39%	319	25,70%	648	52,21%	850	68,49%	1003	80,82%	998	80,41%

Fonte: Elaborada pelo autor.

Com relação à quantidade de acertos na predição, na representação gBoED_Dist, Listas_-Bert_100% obteve 47,14%, Listas_Orig_100% obteve 61,64%, Listas_Syntax_100% obteve 49,24% e Listas_Orig_10% obteve 40,54%. Na representação gBoED_Freq, Listas_Bert_100% obteve 43,51%, Listas_Orig_100% obteve 57,93%, Listas_Syntax_100% obteve 46,10% e Listas_Orig_10% obteve 39,10%. Na representação a quantidade de acertos na predição atingiu os níveis mais baixos. Em gBoED_Syntax, Listas_Bert_100% obteve 13,94%, Listas_Orig_-100% obteve 26,92%, Listas_Syntax_100% obteve 27,24% e Listas_Orig_10% obteve 16,29%. Portanto, nas representações gBoED_Freq e gBoED_Dist, construídas a partir das Listas_Bert_-100%, observa-se grande representatividade da coleção de documentos, porém em gBoED_-Syntax os termos não foram tão representativos.

Nas Tabelas 67 e 68 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Laptop*. Em ambas, os resultados de maior destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, enriquecido pela representação gBoED_Dist**. Nesse cenário os resultados obtidos foram de 80,338% de acurácia, Medida-F1 de 80,337%, $\gamma = 1$, grau de confiança de 55%, com 67 documentos sendo consultados e 56 documentos reclassificados. Para esse mesmo modelo, o uso de Listas_Bert_-100% atingiu um valor maior do que Listas_Orig_100% e Listas_Syntax_100%. Outros destaques que obtiveram os melhores resultados ao utilizar Listas_Bert_100% são os modelos gerados por C4.5-Gini (com gBoED_Freq, gBoED_Dist e gBoED_Syntax). Esses modelos possuem

Tabela 66 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2014 Laptop*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	752	60,60%	980	78,97%	752	60,60%	980	78,97%	403	32,48%	240	19,33%
Qtde de documentos sem representação	489	39,40%	261	21,03%	489	39,40%	261	21,03%	838	67,52%	1001	80,66%
Número de ACERTOS na predição	572	46,10%	540	43,51%	611	49,24%	585	47,14%	338	27,24%	173	13,94%
Número de ERROS na predição	128	10,31%	382	30,78%	89	7,17%	337	27,16%	58	4,67%	65	5,28%
Número de NEUTROS na predição	541	43,59%	319	25,70%	541	43,59%	319	25,70%	845	68,09%	1003	80,82%

Fonte: Elaborada pelo autor.

um maior grau de explicabilidade e obtiveram melhores resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT. Bons resultados também foram obtidos pelos modelos gerados por SVM-Linear e SVM-Polinomial (com gBoED_Freq, gBoED_Dist e gBoED_Syntax).

Tabela 67 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014 Laptop*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**						**gBoED_Syntax**					
		Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,67436	0,65823 174-122	60%	0,63580 192-186	60%	0,62194 174-152	60%	0,66388 174-116	60%	0,64145 192-179	60%	0,62194 174-152	60%	0,61872 174-141	60%	0,60193 192-186	60%	0,59856 174-159	60%
C4.5-Gini	0,68000	0,67355 88-62	60%	0,67694 1-0	55%	0,65500 88-77	60%	0,67677 88-59	60%	0,67694 1-0	55%	0,65500 88-77	60%	0,66467 84-66	60%	0,67613 1-1	55%	0,65258 84-75	60%
KNN-Cosseno	0,77352	0,76869 126-82 $n = 25$	55%	0,73567 134-84 $n = 25$	55%	0,74615 126-100 $n = 25$	55%	0,77352 126-79 $n = 25$	55%	0,74050 134-83 $n = 25$	55%	0,75097 126-98 $n = 25$	55%	0,73890 128-110 $n = 25$	55%	0,70988 134-126 $n = 25$	55%	0,72842 128-119 $n = 25$	55%
KNN-Euclideana	0,77757	0,76790 141-94 $n = 25$	55%	0,73730 143-92 $n = 25$	55%	0,74051 141-115 $n = 25$	55%	0,77353 141-90 $n = 25$	55%	0,74294 143-91 $n = 25$	55%	0,74534 141-113 $n = 25$	55%	0,73408 146-128 $n = 25$	55%	0,71231 143-135 $n = 25$	55%	0,72360 146-136 $n = 25$	55%
MNB	0,79535	0,78727 137-99 $\alpha = 10^{-1}$	60%	0,78482 69-42 $\alpha = 10^{-1}$	55%	0,77519 74-61 $\alpha = 10^{-1}$	55%	0,79291 137-95 $\alpha = 10^{-1}$	60%	0,78966 136-70 $\alpha = 1$	55%	0,77519 74-60 $\alpha = 10^{-1}$	55%	0,77277 74-64 $\alpha = 10^{-1}$	55%	0,77031 69-62 $\alpha = 10^{-1}$	55%	0,76712 74-67 $\alpha = 10^{-1}$	55%
SVM-Linear	0,80093	0,79771 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,80256 10-10 $\gamma = 10^{-4}$ a 10^4	50%	0,79852 5-5 $\gamma = 10^{-4}$ a 10^4	50%	0,79851 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,80256 10-10 $\gamma = 10^{-4}$ a 10^4	50%	0,79932 5-4 $\gamma = 10^{-4}$ a 10^4	50%	0,79850 9-8 $\gamma = 10^{-4}$ a 10^4	50%	0,79934 10-10 $\gamma = 10^{-4}$ a 10^4	50%	0,79689 9-9 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79610	0,79530 75-59 $\gamma = 10^{-1}$	55%	0,80014 11-11 $\gamma = 10^{-1}$	50%	0,79449 9-9 $\gamma = 10^{-1}$	50%	0,79772 75-54 $\gamma = 10^{-1}$	55%	0,80014 11-11 $\gamma = 10^{-1}$	50%	0,79449 9-9 $\gamma = 10^{-1}$	50%	0,79368 5-5 $\gamma = 10^{-1}$	50%	0,79772 11-11 $\gamma = 10^{-1}$	50%	0,79368 5-5 $\gamma = 10^{-1}$	50%
SVM-RBF	0,80173	0,80173 8-8 $\gamma = 1$	50%	0,80257 67-57 $\gamma = 1$	55%	0,80173 8-6 $\gamma = 1$	50%	0,80254 8-8 $\gamma = 1$	50%	0,80338 67-56 $\gamma = 1$	55%	0,80173 8-6 $\gamma = 1$	50%	0,79850 7-7 $\gamma = 1$	50%	0,79613 10-10 $\gamma = 1$	50%	0,79769 7-7 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2014 Restaurant*. Nos gráficos da Figura 54 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_-100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 54a pode-se observar as quantidades de termos do domínio, na Figura 54b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 54c observa-se as quantidades para identificadores da classe

Tabela 68 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2014 Laptop*.

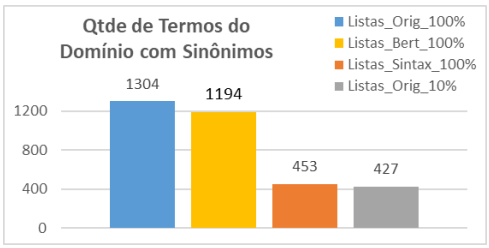
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,67436	0,64290 174-133	60%	0,63580 192-186	60%	0,65016 174-125	60%	0,64145 192-179	60%	0,61872 174-146	60%	0,60193 192-186	60%
C4.5-Gini	0,68000	0,66548 88-68	60%	0,67694 1-0	55%	0,67032 88-63	60%	0,67694 1-0	55%	0,66145 88-69	60%	0,67613 1-1	55%
KNN-Cosseno	0,77352	0,75983 126-86 $n = 25$	55%	0,73567 134-84 $n = 25$	55%	0,76143 126-85 $n = 25$	55%	0,74050 134-83 $n = 25$	55%	0,74374 126-108 $n = 25$	55%	0,70988 134-126 $n = 25$	55%
KNN-Euclideana	0,77757	0,75420 141-102 $n = 25$		0,73730 143-92 $n = 25$	55%	0,75581 141-101 $n = 25$	55%	0,74294 143-91 $n = 25$	55%	0,73650 141-123 $n = 25$	55%	0,71231 143-135 $n = 25$	55%
MNB	0,79535	0,78243 74-54 $\alpha = 10^{-1}$	55%	0,78482 69-42 $\alpha = 10^{-1}$	55%	0,78727 74-52 $\alpha = 10^{-1}$	55%	0,78966 136-70 $\alpha = 1$	55%	0,77518 74-63 $\alpha = 10^{-1}$	55%	0,77031 69-62 $\alpha = 10^{-1}$	55%
SVM-Linear	0,80093	0,79690 7-7 $\gamma = 10^{-4}$ a 10^4	50%	0,80256 10-10 $\gamma = 10^{-4}$ a 10^4	50%	0,79690 7-7 $\gamma = 10^{-4}$ a 10^4	50%	0,80256 10-10 $\gamma = 10^{-4}$ a 10^4	50%	0,79770 5-5 $\gamma = 10^{-4}$ a 10^4	50%	0,79934 10-10 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79610	0,79609 8-7 $\gamma = 10^{-1}$	50%	0,80014 11-11 $\gamma = 10^{-1}$	50%	0,79609 8-7 $\gamma = 10^{-1}$	50%	0,80014 11-11 $\gamma = 10^{-1}$	50%	0,79287 9-8 $\gamma = 10^{-1}$	50%	0,79772 11-11 $\gamma = 10^{-1}$	50%
SVM-RBF	0,80173	0,80253 10-10 $\gamma = 1$	50%	0,80257 67-57 $\gamma = 1$	55%	0,80334 10-9 $\gamma = 1$	50%	0,80338 67-56 $\gamma = 1$	55%	0,80092 8-8 $\gamma = 1$	50%	0,79613 10-10 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

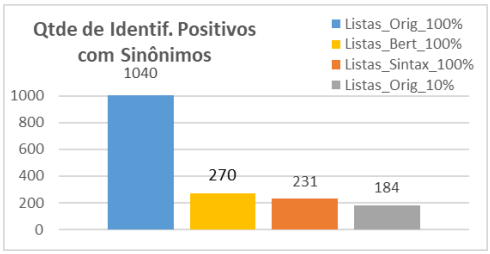
"*Negative*". É possível verificar na Figura 54a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT aproxima-se da quantidade de termos contida nas Listas_Orig_100% e supera a quantidade de termos em Listas_Syntax_100% e Listas_Orig_10%. Na Figura 54b verifica-se que para os identificadores da classe positiva, a quantidade é levemente superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima das Listas_Orig_100%. Na Figura 54c verifica-se que a quantidade de identificadores da classe negativa é superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima das Listas_Orig_100%.

Nas Tabelas 69 e 70 é apresentada a representatividade para o conjunto de documentos *SemEval 2014*. Na Tabela 69 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 70 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram níveis intermediários de representatividade, enquanto gBoED_Syntax formada a partir das Listas_Bert_100% obteve nível baixo de representatividade. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 73,20% dos documentos representados, Listas_Orig_100% obteve 77,10%, Listas_Syntax_100% obteve 72,13% e Listas_Orig_10% obteve 61,46%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 26,98%

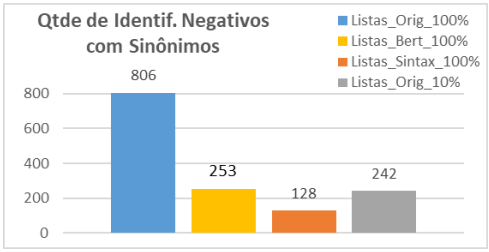
Figura 54 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2014 Restaurant*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

dos documentos representados, Listas_Orig_100% obteve 43,82%, Listas_Syntax_100% obteve 41,25% e Listas_Orig_10% obteve 33,10%.

Com relação à quantidade de acertos na predição, na representação gBoED_Dist, Listas_-Bert_100% obteve 55,03%, Listas_Orig_100% obteve 64,82%, Listas_Syntax_100% obteve 62,96% e Listas_Orig_10% obteve 51,14%. Na representação gBoED_Freq, Listas_Bert_100% obteve 52,10%, Listas_Orig_100% obteve 61,04%, Listas_Syntax_100% obteve 60,68% e Listas_Orig_10% obteve 49,35%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 22,30%, Listas_Orig_100% obteve 38,14%, Listas_Syntax_100% obteve 41,25% e Listas_Orig_-10% obteve 28,90%.

Tabela 69 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014 Restaturant*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	1286	77,10%	1221	73,20%	1025	61,46%	1286	77,10%	1221	73,20%	1025	61,46%	731	43,82%	450	26,98%	552	33,10%
Qtde de documentos sem representação	382	22,90%	447	26,80%	643	38,54%	382	22,90%	447	26,80%	643	38,54%	937	56,18%	1218	73,02%	1116	66,90%
Número de ACERTOS na predição	1018	61,04%	869	52,10%	823	49,35%	1081	64,82%	918	55,03%	853	51,14%	636	38,14%	372	22,30%	482	28,90%
Número de ERROS na predição	157	9,41%	257	15,40%	140	8,39%	94	5,63%	208	12,47%	110	6,59%	75	4,49%	75	4,50%	62	3,72%
Número de NEUTROS na predição	493	29,55%	542	32,50%	705	42,26%	493	29,55%	542	32,49%	705	42,26%	957	57,37%	1221	73,20%	1124	67,38%

Fonte: Elaborada pelo autor.

Nas Tabelas 71 e 72 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2014 Restaurant*. Em ambas, os resultados de maior

Tabela 70 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2014 Restaurant*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	1203	72,13%	1221	73,20%	1203	72,13%	1221	73,20%	688	41,25%	450	26,98%
Qtde de documentos sem representação	465	27,87%	447	26,80%	465	27,87%	447	26,80%	980	58,75	1218	73,02%
Número de ACERTOS na predição	1012	60,68%	869	52,10%	1050	62,96%	918	55,03%	616	36,94%	372	22,30%
Número de ERROS na predição	119	7,13%	257	15,40%	81	4,85%	208	12,47%	62	3,71%	75	4,50%
Número de NEUTROS na predição	537	32,19%	542	32,50%	537	32,19%	542	32,49%	990	59,35%	1221	73,20%

Fonte: Elaborada pelo autor.

destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel Polinomial, enriquecido pela representação gBoED_Syntax**. Nesse cenário os resultados obtidos foram de 80,339% de acurácia, Medida-F1 de 80,339%, $\gamma = 1$, grau de confiança de 50%, com 9 documentos sendo consultados e 9 documentos reclassificados. Outros destaques são para o mesmo algoritmo, porém enriquecidos com as representações gBoED_Freq e gBoED_Dist. Em ***Support Vector Machine* – SVM, kernel RBF, enriquecido pelas representações gBoED_-Freq, gBoED_Dist e gBoED_Syntax** obtiveram os maiores resultados, porém intermediários em relação às Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%.

Tabela 71 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2014 Restaurant*.

		gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
Algoritmos	**BoW**	**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75003	0,74940 1-0	50%	0,74823 2-2	50%	0,74940 1-1	50%	0,74940 1-0	50%	0,74823 2-2	50%	0,74940 1-1	50%	0,75120 7-6	60%	0,74823 2-2	50%	0,75180 1-1	55%
C4.5-Gini	0,75543	0,76139 3-3	60%	0,76081 3-3	65%	0,76139 3-3	60%	0,76139 3-3	60%	0,76081 3-3	65%	0,76139 3-3	60%	0,76078 3-3	65%	0,76021 3-3	65%	0,76078 3-3	65%
KNN-Cosseno	0,79196	0,78059 164-108 $n = 13$	55%	0,76261 199-131 $n = 11$	55%	0,76081 164-126 $n = 13$	55%	0,78478 164-101 $n = 13$	55%	0,76681 199-125 $n = 11$	55%	0,76200 164-123 $n = 13$	55%	0,75660 144-124 $n = 15$	55%	0,74405 136-124 $n = 17$	55%	0,74761 144-133 $n = 15$	55%
KNN-Euclideana	0,78959	0,78058 166-110 $n = 13$	55%	0,76082 201-133 $n = 11$	55%	0,76080 166-128 $n = 13$	55%	0,78478 166-103 $n = 13$	55%	0,76501 201-127 $n = 13$	55%	0,76200 166-125 $n = 13$	55%	0,75660 125-106 $n = 19$	55%	0,74345 130-120 $n = 19$	55%	0,74761 146-133 $n = 15$	55%
MNB	0,79856	0,79976 162-108 $\alpha = 10^{-1}$	60%	0,80277 57-40 $\alpha = 10^{-3}$	55%	0,79315 78-65 $\alpha = 10^{-1}$	55%	0,80575 162-95 $\alpha = 10^{-1}$	60%	0,80457 57-36 $\alpha = 10^{-3}$	55%	0,79495 78-62 $\alpha = 10^{-1}$	55%	0,78776 76-63 $\alpha = 10^{-1}$	55%	0,79677 57-52 $\alpha = 10^{-3}$	55%	0,78416 76-69 $\alpha = 10^{-1}$	55%
SVM-Linear	0,79438	0,79438 7-6 $\gamma = 10^{-4}$ a 10^{4}	50%	0,78779 94-84 $\gamma = 10^{-4}$ a 10^{4}	55%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^{4}	50%	0,79438 7-6 $\gamma = 10^{-4}$ a 10^{4}	50%	0,79019 94-80 $\gamma = 10^{-4}$ a 10^{4}	55%	0,79378 7-6 $\gamma = 10^{-4}$ a 10^{4}	50%	0,79437 8-8 $\gamma = 10^{-4}$ a 10^{4}	50%	0,78361 11-11 $\gamma = 10^{-4}$ a 10^{4}	50%	0,79318 8-8 $\gamma = 10^{-4}$ a 10^{4}	50%
SVM-Polinomial	0,79558	0,79558 7-7 $\gamma = 1$	50%	0,80279 9-9 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%	0,79558 7-7 $\gamma = 1$	50%	0,80279 9-8 $\gamma = 1$	50%	0,79499 7-7 $\gamma = 1$	50%	0,79317 10-8 $\gamma = 1$	50%	0,80339 9-9 $\gamma = 1$	50%	0,79257 10-9 $\gamma = 1$	50%
SVM-RBF	0,81598	0,81658 6-2 $\gamma = 1$	50%	0,81537 5-1 $\gamma = 1$	50%	0,81598 6-3 $\gamma = 1$	50%	0,81658 6-2 $\gamma = 1$	50%	0,81537 6-2 $\gamma = 1$	50%	0,81598 5-1 $\gamma = 1$	50%	0,81538 6-6 $\gamma = 1$	50%	0,81417 5-3 $\gamma = 1$	50%	0,81538 6-6 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

Outros destaques são os modelos que obtiveram resultados intermediários e que superam a BoW, ocorre em MNB (com gBoED_Freq, gBoED_Dist e gBoED_Syntax). Esse modelo

possui um maior grau de explicabilidade e obtiveram melhores resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT.

Tabela 72 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2014 Restaurant*.

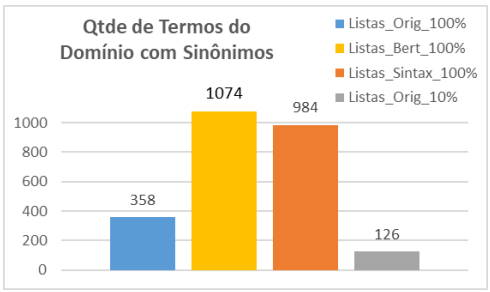
Algoritmos	**BoW**	**gBoED_Freq**				**gBoED_Dist**				**gBoED_Syntax**			
		Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75003	0,75000 9-7	65%	0,74823 2-2	50%	0,75000 9-7	65%	0,74823 2-2	50%	0,74940 9-8	65%	0,74823 2-2	50%
C4.5-Gini	0,75543	0,76139 3-3	65%	0,76081 3-3	65%	0,76139 3-3	65%	0,76081 3-3	65%	0,76079 3-3	65%	0,76021 3-3	65%
KNN-Cosseno	0,79196	0,77580 164-114 $n = 13$	60%	0,76261 199-131 $n = 11$	55%	0,77819 164-105 $n = 13$	60%	0,76681 199-125 $n = 11$	55%	0,75600 145-121 $n = 15$	55%	0,74405 136-124 $n = 17$	55%
KNN-Euclideana	0,78959	0,77579 166-116 $n = 13$	60%	0,76082 201-133 $n = 11$	55%	0,77819 166-107 $n = 13$	60%	0,76501 201-127 $n = 13$	55%	0,75660 124-103 $n = 19$	55%	0,74345 130-120 $n = 19$	55%
MNB	0,79856	0,79616 78-62 $\alpha = 10^{-1}$	55%	0,80277 57-40 $\alpha = 10^{-3}$	55%	0,79796 78-57 $\alpha = 10^{-1}$	55%	0,80457 57-36 $\alpha = 10^{-3}$	55%	0,78836 78-69 $\alpha = 10^{-1}$	55%	0,79677 57-52 $\alpha = 10^{-3}$	55%
SVM-Linear	0,79438	0,79437 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,78779 94-84 $\gamma = 10^{-4}$ a 10^4	55%	0,79437 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,79019 94-80 $\gamma = 10^{-4}$ a 10^4	55%	0,79377 7-6 $\gamma = 10^{-4}$ a 10^4	50%	0,78361 11-11 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,79558	0,79498 7-7 $\gamma = 1$	50%	0,80279 9-9 $\gamma = 1$	50%	0,79498 7-7 $\gamma = 1$	50%	0,80279 9-8 $\gamma = 1$	50%	0,79558 7-6 $\gamma = 1$	50%	0,80339 9-9 $\gamma = 1$	50%
SVM-RBF	0,81598	0,81657 6-2 $\gamma = 1$	50%	0,81537 5-1 $\gamma = 1$	50%	0,81657 6-2 $\gamma = 1$	50%	0,81537 6-2 $\gamma = 1$	50%	0,81418 6-6 $\gamma = 1$	50%	0,81417 5-3 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

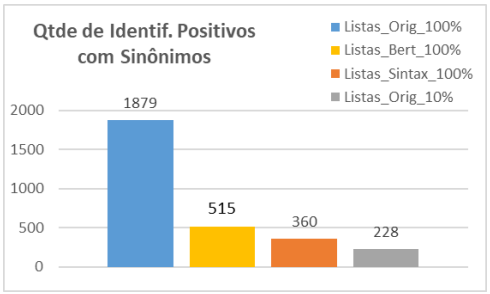
Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2015*. Nos gráficos da Figura 55 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_-Syntax_100% e Listas_Orig_10%. No gráfico da Figura 55a pode-se observar as quantidades de termos do domínio, na Figura 55b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 55c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 55a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT, de modo semelhante à quantidade termos do domínio extraída pelo método de extração baseado em análise morfossintática (Listas_-Syntax_100%), é bastante superior à quantidade de termos do domínio contida nas Listas_Orig_-100%. Na Figura 55b e Figura 55c verifica-se que para os identificadores da classe positiva e identificadores da classe negativa respectivamente, a quantidade é levemente superior às Listas_Syntax_100% e Listas_Orig_10%, porém não se aproxima das Listas_Orig_100%.

Nas Tabelas 73 e 74 é apresentada a representatividade para o conjunto de documentos *SemEval 2015*. Na Tabela 73 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 74 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_-

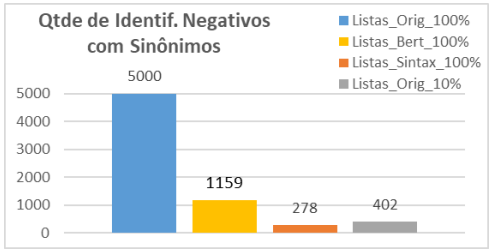
Figura 55 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram níveis altos de representatividade, aproximando-se das versões geradas por Listas_Syntax_100%. Para gBoED_Syntax formada a partir das Listas_Bert_100% foram obtidos níveis intermediários de representatividade. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 95% dos documentos representados, Listas_Orig_100% obteve 55,81%, Listas_Syntax_100% obteve 95,89% e Listas_Orig_10% obteve 52,93%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 68,04% dos documentos representados, Listas_Orig_100% obteve 74,29%, Listas_Syntax_100% obteve 84,52% e Listas_Orig_10% obteve 65,92%.

Com relação à quantidade de acertos na predição, na representação gBoED_Dist, Listas_-Bert_100% obteve 71,91%, Listas_Orig_100% obteve 44,58%, Listas_Syntax_100% obteve 78,53% e Listas_Orig_10% obteve 41,71%. Na representação gBoED_Freq, Listas_Bert_100% obteve 70,66%, Listas_Orig_100% obteve 43,83%, Listas_Syntax_100% obteve 76,28% e Listas_Orig_10% obteve 40,07%. Na representação gBoED_Syntax, Listas_Bert_100% obteve o melhor nível de acertos com 44,32%, Listas_Orig_100% obteve 57,80%, Listas_Syntax_100% obteve 66,92% e Listas_Orig_10% obteve 51,60%.

Nas Tabelas 75 e 76 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015*. Em ambas, os resultados de maior destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, enriquecido pela**

Tabela 73 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	447	55,81%	761	95%	424	52,93%	447	55,81%	761	95%	424	52,93%	595	74,29%	545	68,04%	528	65,92%
Qtde de documentos sem representação	354	44,19%	40	5%	377	47,06%	354	44,19%	40	5%	377	47,06%	206	25,71%	256	31,96%	273	34,08%
Número de ACERTOS na predição	351	43,83%	566	70,66%	321	40,07%	357	44,58%	576	71,91%	334	41,71%	463	57,80%	355	44,32%	414	51,69%
Número de ERROS na predição	72	8,98%	163	20,35%	79	9,86%	66	8,23%	153	19,10%	66	8,23%	68	8,50%	131	16,35%	64	7,99%
Número de NEUTROS na predição	378	47,19%	72	8,99%	401	50,06%	378	47,19%	72	8,99%	401	50,06%	270	33,70%	315	39,33%	323	40,32%

Fonte: Elaborada pelo autor.

Tabela 74 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2015*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	768	95,89%	761	95%	768	95,89%	761	95%	677	84,52%	545	68,04%
Qtde de documentos sem representação	33	4,11%	40	5%	33	4,11%	40	5%	124	15,48%	256	31,96%
Número de ACERTOS na predição	611	76,28%	566	70,66%	629	78,53%	576	71,91%	536	66,92%	355	44,32%
Número de ERROS na predição	126	15,73%	163	20,35%	108	13,48%	153	19,10%	79	9,86%	131	16,35%
Número de NEUTROS na predição	64	7,99%	72	8,99%	64	7,99%	72	8,99%	186	23,22%	315	39,33%

Fonte: Elaborada pelo autor.

representação gBoED_Dist. Nesse cenário os resultados obtidos foram de 88,269% de acurácia, Medida-F1 de 88,269%, $\gamma = 10^{-1}$, grau de confiança de 60%, com 60 documentos sendo consultados e 38 documentos reclassificados. Outros destaques são para o mesmo algoritmo, porém enriquecidos com as representações gBoED_Freq e gBoED_Dist. Em ***Support Vector Machine* – SVM, kernel Polinomial e Linear, enriquecido pelas representações gBoED_-Freq, gBoED_Dist e gBoED_Syntax** obtiveram outros bons resultados, superando os resultados da BoW e Listas_Orig_100% e Listas_Orig_10%, ficando atrás dos resultados de Listas_Syntax_-100%. Em nenhum desses cenários foram superados os resultados obtidos pelo enriquecimento das representações construídas a partir das Listas_Syntax_100%.

Outro destaque é o modelo MNB, enriquecido com gBoED_Freq, gBoED_Dist e gBoED_Syntax, que obtiveram resultados que superam tanto a BoW quanto as representações formadas pelas Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. Esse modelo possui um maior grau de explicabilidade e obtiveram melhores resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT.

Tabela 75 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75404	0,74904 32-18	60%	0,76793 316-106	70%	0,74529 32-17	55%	0,74779 32-17	60%	0,77784 356-108	75%	0,74529 32-17	60%	0,76657 33-20	60%	0,73290 58-27	60%	0,76407 1-1	50%
C4.5-Gini	0,77528	0,76028 31-21	60%	0,78529 2-1	50%	0,76028 31-20	60%	0,76028 31-20	60%	0,78529 2-1	50%	0,76028 31-20	60%	0,71902 366-208	85%	0,78404 2-2	50%	0,68648 366-217	85%
KNN-Cosseno	0,81526	0,79778 40-30 $n = 25$	55%	0,82529 74-40 $n = 13$	55%	0,79528 40-30 $n = 25$	55%	0,79653 40-29 $n = 25$	55%	0,82281 74-36 $n = 13$	55%	0,79528 40-31 $n = 25$	55%	0,81271 85-56 $n = 13$	55%	0,79781 35-26 $n = 25$	55%	0,79778 40-31 $n = 25$	55%
KNN-Euclideana	0,81651	0,79903 39-29 $n = 25$	55%	0,82529 75-41 $n = 13$	55%	0,79653 39-29 $n = 25$	55%	0,79778 39-28 $n = 25$	55%	0,82281 75-37 $n = 13$	55%	0,79653 39-30 $n = 25$	55%	0,81271 86-57 $n = 13$	55%	0,79781 35-26 $n = 25$	55%	0,79778 39-30 $n = 25$	55%
MNB	0,86145	0,85145 24-19 $\alpha = 10^{-2}$	55%	0,86272 70-37 $\alpha = 10^{-1}$	60%	0,85020 24-18 $\alpha = 10^{-2}$	55%	0,85148 35-25 $\alpha = 10^{-1}$	55%	0,86145 70-33 $\alpha = 10^{-1}$	60%	0,85020 24-17 $\alpha = 10^{-2}$	55%	0,86146 63-44 $\alpha = 10^{-1}$	60%	0,84900 28-21 $\alpha = 10^{-1}$	55%	0,85770 60-44 $\alpha = 10^{-1}$	60%
SVM-Linear	0,86148	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,87145 35-24 $\gamma = 10^{-4}$ a 10^4	55%	0,86023 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,86398 61-48 $\gamma = 10^{-4}$ a 10^4	60%	0,87644 98-51 $\gamma = 10^{-4}$ a 10^4	65%	0,86148 61-49 $\gamma = 10^{-4}$ a 10^4	60%	0,85523 34-27 $\gamma = 10^{-4}$ a 10^4	55%	0,86395 35-25 $\gamma = 10^{-4}$ a 10^4	55%	0,85148 34-31 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,87270	0,87020 5-3 $\gamma = 1$	50%	0,87769 61-20 $\gamma = 1$	60%	0,87020 5-2 $\gamma = 1$	50%	0,87270 5-4 $\gamma = 1$	50%	0,88144 61-15 $\gamma = 1$	60%	0,87147 30-16 $\gamma = 1$	55%	0,86768 2-2 $\gamma = 10^{-2}$	50%	0,87269 6-4 $\gamma = 1$	50%	0,86768 2-2 $\gamma = 10^{-2}$	50%
SVM-RBF	0,87390	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87770 60-43 $\gamma = 10^{-1}$	60%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,88269 60-38 $\gamma = 10^{-1}$	60%	0,87142 2-2 $\gamma = 10^{-1}$	50%	0,87140 2-2 $\gamma = 10^{-1}$	50%	0,87147 1-1 $\gamma = 10^{-1}$	50%	0,87015 4-4 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

Tabela 76 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2015*.

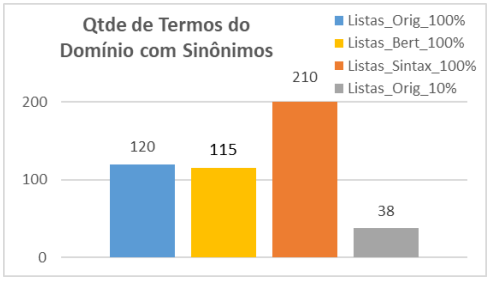
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75404	0,78773 373-97	95%	0,76793 316-106	70%	0,787713 72-91	95%	0,77784 356-108	75%	0,75279 32-21	60%	0,73290 58-27	60%
C4.5-Gini	0,77528	0,79898 373-92	95%	0,78529 2-1	50%	0,787713 373-92	95%	0,78529 2-1	50%	0,76902 31-18	60%	0,78404 2-2	50%
KNN-Cosseno	0,81526	0,81645 40-30 $n = 7$	60%	0,82529 74-40 $n = 13$	55%	0,82018 155-79 $n = 7$	60%	0,82281 74-36 $n = 13$	55%	0,81778 155-90 $n = 7$	60%	0,79781 35-26 $n = 25$	55%
KNN-Euclideana	0,81651	0,81770 155-79 $n = 7$	60%	0,82529 75-41 $n = 13$	55%	0,82143 155-79 $n = 7$	60%	0,82281 75-37 $n = 13$	55%	0,81902 155-90 $n = 7$	60%	0,79781 35-26 $n = 25$	55%
MNB	0,86145	0,85895 24-13 $\alpha = 10^{-2}$	55%	0,86272 70-37 $\alpha = 10^{-1}$	60%	0,86146 97-39 $\alpha = 10^{-1}$	65%	0,86145 70-33 $\alpha = 10^{-1}$	60%	0,86145 64-37 $\alpha = 10^{-1}$	60%	0,84900 28-21 $\alpha = 10^{-1}$	55%
SVM-Linear	0,86148	0,88146 160-85 $\gamma = 10^{-4}$ a 10^4	75%	0,87145 35-24 $\gamma = 10^{-4}$ a 10^4	55%	0,89020 160-78 $\gamma = 10^{-4}$ a 10^4	75%	0,87644 98-51 $\gamma = 10^{-4}$ a 10^4	65%	0,86020 30-23 $\gamma = 10^{-4}$ a 10^4	55%	0,86395 35-25 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,87270	0,88895 91-30 $\gamma = 1$	65%	0,87769 61-20 $\gamma = 1$	60%	0,89518 91-28 $\gamma = 1$	65%	0,88144 61-15 $\gamma = 1$	60%	0,87145 5-4 $\gamma = 1$	50%	0,87269 6-4 $\gamma = 1$	50%
SVM-RBF	0,87390	0,88516 129-32 $\gamma = 1$	70%	0,87770 60-43 $\gamma = 10^{-1}$	60%	0,89389 192-47 $\gamma = 1$	80%	0,88269 60-38 $\gamma = 10^{-1}$	60%	0,88013 68-44 $\gamma = 10^{-1}$	60%	0,87147 1-1 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

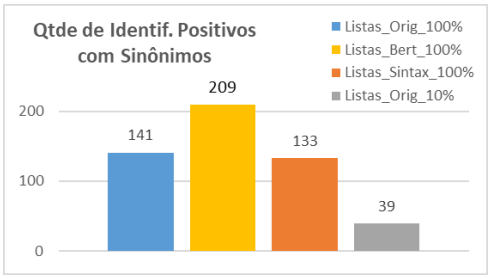
Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2015 Hotel*. Nos gráficos da Figura 56 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_-100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 56a pode-se observar as

quantidades de termos do domínio, na Figura 56b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 56c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 56a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT, atingiu uma quantidade bastante semelhante às Listas_Orig_100%. Os termos do domínio extraídos pelo método de extração baseado em análise morfossintática (Listas_Syntax_100%) ainda é bastante superior à quantidade de termos do domínio contida outras listas. Na Figura 56b e Figura 56c verifica-se que para os identificadores da classe positiva e identificadores da classe negativa respectivamente, a quantidade é semelhante às Listas_Orig_100% e bastante superior às Listas_Syntax_100% e Listas_Orig_10%.

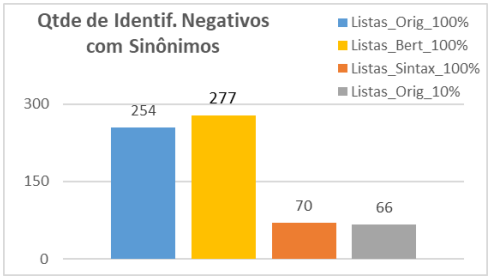
Figura 56 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Hotel*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Nas Tabelas 77 e 78 é apresentada a representatividade para o conjunto de documentos *SemEval 2015 Hotel*. Na Tabela 77 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 78 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram níveis altos de representatividade, comparando-se às versões geradas por Listas_Orig_100% e Listas_Syntax_100%. Para gBoED_Syntax formada a partir das Listas_Bert_100% foram obtidos níveis intermediários de representatividade, não superando Listas_Orig_100% e Listas_Syntax_100%. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 100% dos documentos representados, Listas_Orig_100% obteve 100%, Listas_Syntax_100%

obteve 100% e Listas_Orig_10% obteve 96,56%. Na representação gBoED_Syntax, Listas_-Bert_100% obteve 86,20% dos documentos representados, Listas_Orig_100% obteve 96,56%, Listas_Syntax_100% obteve 96,56% e Listas_Orig_10% obteve 75,86%.

Com relação à quantidade de acertos na predição, de maneira geral, as Listas_Bert_-100% não obtiveram bons resultados. Na representação gBoED_Dist, Listas_Bert_100% obteve 68,97%, Listas_Orig_100% obteve 82,77%, Listas_Syntax_100% obteve 86,22% e Listas_-Orig_10% obteve 79,32%. Na representação gBoED_Freq, Listas_Bert_100% obteve 62,07%, Listas_Orig_100% obteve 82,77%, Listas_Syntax_100% obteve 82,77% e Listas_Orig_10% obteve 75,87%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 62,07%, Listas_-Orig_100% obteve 79,32%, Listas_Syntax_100% obteve 79,32% e Listas_Orig_10% obteve 58,63%.

Tabela 77 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Hotel*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	29	100%	29	100%	28	96,56%	29	100%	29	100%	28	96,56%	28	96,56%	25	86,20%	22	75,86%
Qtde de documentos sem representação	0	0%	0	0%	1	3,44%	0	0%	0	0%	1	3,44%	1	3,44%	4	13,80%	7	24,14%
Número de ACERTOS na predição	24	82,77%	18	62,07%	22	75,87%	24	82,77%	20	68,97%	23	79,32%	23	79,32%	18	62,07%	17	58,63%
Número de ERROS na predição	4	13,79%	10	34,49%	5	17,24%	4	13,79%	8	27,59%	4	13,79%	3	10,34%	4	13,79%	4	13,79%
Número de NEUTROS na predição	1	3,44%	1	3,44%	2	6,89%	1	3,44%	1	3,44%	2	6,89%	3	10,34%	7	24,14%	8	27,58%

Fonte: Elaborada pelo autor.

Tabela 78 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2015 Hotel*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	29	100%	29	100%	29	100%	29	100%	28	96,56%	25	86,20%
Qtde de documentos sem representação	0	0%	0	0%	0	0%	0	0%	1	3,44%	4	13,80%
Número de ACERTOS na predição	24	82,77%	18	62,07%	25	86,22%	20	68,97%	23	79,32%	18	62,07%
Número de ERROS na predição	3	10,34%	10	34,49%	2	6,89%	8	27,59%	2	6,89%	4	13,79%
Número de NEUTROS na predição	2	6,89%	1	3,44%	2	6,89%	1	3,44%	4	13,79%	7	24,14%

Fonte: Elaborada pelo autor.

Nas Tabelas 79 e 80 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Hotel*. Para essa coleção de documentos, os modelos enriquecidos pelas representações construídas a partir das Listas_Bert_100% não

superaram Listas_Orig_100% e Listas_Syntax_100%, portanto não serão marcados pela linha cinza. Os melhores resultados para Listas_Bert_100% estão relacionados com ***Multinomial Naïve Bayes* – MNB, enriquecido pela representação gBoED_Freq e gBoED_Dist**. Nesse cenário os resultados obtidos foram de 83,333% de acurácia, Medida-F1 de 83,333%, $\gamma = 10^{-1}$, grau de confiança de 85%, com 11 documentos sendo consultados e 4 documentos reclassificados. Com o mesmo nível de importância, outros cenários de destaque que obtiveram resultados semelhantes foram com relação aos modelos C4.5-Entropia, C4.5-Gini, KNN-Cosseno e KNN-Euclidiana, todos eles a partir das representações gBoED_Freq e gBoED_Dist e pertencentes a maiores níveis de explicabilidade.

Tabela 79 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Hotel*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,83333	0,83333 29-8	100%	0,83333 29-13	100%	0,61667 29-10	100%	0,83333 29-8	100%	0,68333 29-11	100%	0,80000 29-9	100%	0,80000 29-9	100%	0,61667 29-13	100%	0,60000 29-14	100%
C4.5-Gini	0,83333	0,83333 29-8	100%	0,83333 29-13	100%	0,61667 29-10	100%	0,83333 29-8	100%	0,68333 29-11	100%	0,80000 29-9	100%	0,80000 29-9	100%	0,61667 29-13	100%	0,60000 29-14	100%
KNN-Cosseno	0,80000	0,86667 15-5 $n = 7$	90%	0,83333 9-3 $n = 9$	80%	0,80000 11-4 $n = 3$	70%	0,86667 15-4 $n = 7$	90%	0,83333 9-3 $n = 9$	80%	0,80000 11-4 $n = 3$	70%	0,86667 16-6 $n = 7$	90%	0,73333 3-1 $n = 7$	60%	0,73333 1-1 $n = 11$	55%
KNN-Euclideana	0,80000	0,86667 15-5 $n = 7$	90%	0,83333 9-3 $n = 9$	80%	0,80000 11-4 $n = 3$	70%	0,86667 15-4 $n = 7$	90%	0,83333 9-3 $n = 9$	80%	0,80000 11-4 $n = 3$	70%	0,86667 16-6 $n = 7$	90%	0,73333 3-1 $n = 5$	60%	0,73333 2-1 $n = 7$	60%
MNB	0,73333	0,86667 14-5 $\alpha = 1$	85%	0,83333 11-4 $\alpha = 10^{-1}$	85%	0,80000 8-3 $\alpha = 1$	80%	0,86667 14-4 $\alpha = 1$	85%	0,83333 11-4 $\alpha = 10^{-1}$	85%	0,80000 8-3 $\alpha = 1$	80%	0,86667 14-6 $\alpha = 10^{-2}$ a 1	85%	0,76667 1-0 $\alpha = 10^{-1}$	60%	0,73333 2-2 $\alpha = 10^{-1}$	65%
SVM-Linear	0,73333	0,83333 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,76667 4-3 $\gamma = 10^{-4}$ a 10^{4}	55%	0,80000 8-2 $\gamma = 10^{-4}$ a 10^{4}	70%	0,83333 8-4 $\gamma = 10^{-4}$ a 10^{4}	70%	0,80000 4-3 $\gamma = 10^{-4}$ a 10^{4}	55%	0,80000 8-2 $\gamma = 10^{-4}$ a 10^{4}	70%	0,83333 9-5 $\gamma = 10^{-4}$ a 10^{4}	75%	0,73333 4-2 $\gamma = 10^{-4}$ a 10^{4}	55%	0,73333 2-1 $\gamma = 10^{-4}$ a 10^{4}	60%
SVM-Polinomial	0,73333	0,83333 6-3 $\gamma = 10^{-1}$	70%	0,76667 1-0 $\gamma = 10^{-1}$	50%	0,80000 6-2 $\gamma = 10^{-1}$	70%	0,83333 6-3 $\gamma = 10^{-1}$	70%	0,80000 6-3 $\gamma = 10^{-1}$	60%	0,80000 6-2 $\gamma = 10^{-1}$	70%	0,86667 15-6 $\gamma = 1$	80%	0,76667 1-0 $\gamma = 10^{-1}$	50%	0,73333 1-0 $\gamma = 1$	55%
SVM-RBF	0,73333	0,83333 21-3 $\gamma = 1$	75%	0,73333 1-0 $\gamma = 1$	55%	0,76667 21-2 $\gamma = 1$	60%	0,87142 2-2 $\gamma = 10^{-1}$	75%	0,73333 1-0 $\gamma = 1$	55%	0,76667 21-2 $\gamma = 1$	60%	0,80000 29-8 $\gamma = 10^{-4}$ a 10^{4}	85%	0,73333 1-0 $\gamma = 1$	55%	0,70000 3-1 $\gamma = 1$	60%

Fonte: Elaborada pelo autor.

Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2015 Laptop*. Nos gráficos da Figura 57 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_-100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 57a pode-se observar as quantidades de termos do domínio, na Figura 57b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 57c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 57a que a quantidade de termos do domínio extraída pelo método de extração de termos baseado em modelos de linguagem BERT é bastante superior a todas as outras listas. Na Figura 57b e Figura 57c verifica-se que para os identificadores da classe positiva e negativa, respectivamente, a quantidade é levemente superior às Listas_Orig_100% e bastante superior às Listas_Syntax_100% e Listas_Orig_10%.

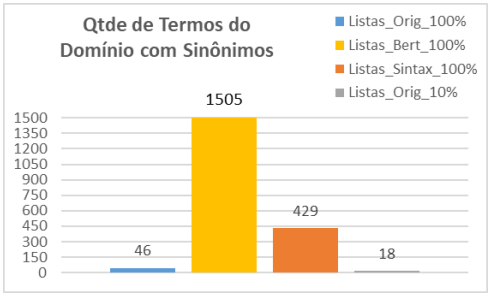
Nas Tabelas 81 e 82 é apresentada a representatividade para o conjunto de documentos *SemEval 2015*. Na Tabela 81 é feita a comparação entre a representatividade das gBoEDs cons-

Tabela 80 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2015 Hotel*.

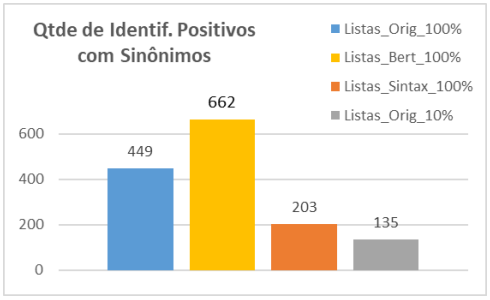
Algoritmos	**BoW**	**gBoED_Freq**				**gBoED_Dist**				**gBoED_Syntax**			
		Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,83333	0,83333 29-6	100%	0,83333 29-13	100%	0,86667 29-5	100%	0,68333 29-11	100%	0,80000 29-8	100%	0,61667 29-13	100%
C4.5-Gini	0,83333	0,83333 29-6	100%	0,83333 29-13	100%	0,86667 29-5	100%	0,68333 29-11	100%	0,80000 29-8	100%	0,61667 29-13	100%
KNN-Cosseno	0,80000	0,83333 11-5 $n = 15$	80%	0,83333 9-3 $n = 9$	80%	0,90000 11-5 $n = 15$	80%	0,83333 9-3 $n = 9$	80%	0,83333 11-4 $n = 3$	70%	0,73333 3-1 $n = 7$	60%
KNN-Euclideana	0,80000	0,83333 11-5 $n = 15$	80%	0,83333 9-3 $n = 9$	80%	0,90000 11-5 $n = 15$	80%	0,83333 9-3 $n = 9$	80%	0,83333 11-4 $n = 3$	70%	0,73333 3-1 $n = 5$	60%
MNB	0,73333	0,83333 14-5 $\alpha = 1$	85%	0,83333 11-4 $\alpha = 10^{-1}$	85%	0,90000 14-5 $\alpha = 1$	85%	0,83333 11-4 $\alpha = 10^{-1}$	85%	0,83333 14-6 $\alpha = 1$	85%	0,76667 1-0 $\alpha = 10^{-1}$	60%
SVM-Linear	0,73333	0,83333 8-4 $\gamma = 10^{-4}$ a 10^4	70%	0,76667 4-3 $\gamma = 10^{-4}$ a 10^4	55%	0,86667 8-4 $\gamma = 10^{-4}$ a 10^4	70%	0,80000 4-3 $\gamma = 10^{-4}$ a 10^4	55%	0,83333 17-6 $\gamma = 10^{-4}$ a 10^4	90%	0,73333 4-2 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,73333	0,83333 10-4 $\gamma = 10^{-1}$	75%	0,76667 1-0 $\gamma = 10^{-1}$	50%	0,86667 10-4 $\gamma = 10^{-1}$	75%	0,80000 6-3 $\gamma = 10^{-1}$	60%	0,83333 15-6 $\gamma = 1$	80%	0,76667 1-0 $\gamma = 10^{-1}$	50%
SVM-RBF	0,73333	0,83333 29-5 $\gamma = 10^{-1}$ a 10^4	85%	0,73333 1-0 $\gamma = 1$	55%	0,86667 29-6 $\gamma = 10^{-1}$ a 10^4	85%	0,73333 1-0 $\gamma = 1$	55%	0,80000 29-7 $\gamma = 10^{-1}$ a 10^4	85%	0,73333 1-0 $\gamma = 1$	55%

Fonte: Elaborada pelo autor.

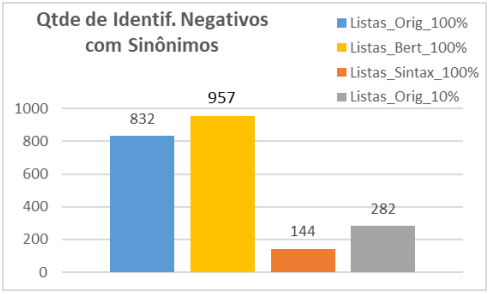
Figura 57 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Laptop*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

truídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 82 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax, formadas a partir das Listas_Syntax_100% e Listas_Bert_100% obtiveram níveis bem mais altos de representatividade, quando comparados às Listas_Orig_100% e Listas_Orig_10%. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 63,08% dos documentos representados, Listas_Orig_100% obteve 21,03%, Listas_Syntax_100% obteve 90,66% e Listas_Orig_10% obteve 15,66%, uma diferença bastante significativa. Na representação gBoED_Syntax, Listas_-Bert_100% obteve 69,86% dos documentos representados, Listas_Orig_100% obteve 63,32%, Listas_Syntax_100% obteve 76,64% e Listas_Orig_10% obteve 54,21%.

Com relação à quantidade de acertos na predição, na representação gBoED_Dist, Listas_-Bert_100% obteve 43%, Listas_Orig_100% obteve 14,96%, Listas_Syntax_100% obteve 75,95% e Listas_Orig_10% obteve 10,98%. Na representação gBoED_Freq, Listas_Bert_100% obteve 39,72%, Listas_Orig_100% obteve 13,56%, Listas_Syntax_100% obteve 72,68% e Listas_-Orig_10% obteve 10,29%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 42,05%, Listas_Orig_100% obteve 48,38%, Listas_Syntax_100% obteve 63,78% e Listas_Orig_10% obteve 40,68%.

Tabela 81 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Laptop*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	90	21,03%	270	63,08%	67	15,66%	90	21,03%	270	63,08%	67	15,66%	271	63,32%	299	69,86%	232	54,21%
Qtde de documentos sem representação	338	78,97%	158	36,92%	361	84,34%	338	78,97%	158	36,92%	361	84,34%	157	36,68%	129	30,14%	196	45,79%
Número de ACERTOS na predição	58	13,56%	170	39,72%	44	10,29%	64	14,96%	184	43%	47	10,98%	207	48,38%	180	42,05%	174	40,68%
Número de ERROS na predição	23	5,37%	87	20,33%	19	4,43%	17	3,97%	73	17,05%	16	3,74%	32	7,47%	80	18,70%	30	7%
Número de NEUTROS na predição	347	81,07%	171	39,95%	365	85,28%	347	81,07%	171	39,95%	365	85,28%	189	44,15%	168	39,25%	224	52,33%

Fonte: Elaborada pelo autor.

Nas Tabelas 83 e 84 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Laptop*. Em ambas, os resultados de maior destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel RBF, enriquecido pela representação gBoED_Dist**. Nesse cenário os resultados obtidos foram de 88,577% de acurácia, Medida-F1 de 88,575%, $\gamma = 10^{-1}$, grau de confiança de 50%, com 1 documentos sendo consultados e 1 documentos reclassificados. Outros destaques são para o mesmo algoritmo, porém enriquecidos com as representações gBoED_Freq e gBoED_Freq e gBoED_Syntax.

Outro destaque é o modelo C4.5-Entropia (enriquecido com gBoED_Freq, gBoED_Dist e gBoED_Syntax), que obtiveram resultados que superam tanto a BoW quanto as representações

Tabela 82 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_Syntax_100%, no conjunto de dados *SemEval 2015 Laptop*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	388	90,66%	270	63,08%	388	90,66%	270	63,08%	328	76,64%	299	69,86%
Qtde de documentos sem representação	40	9,34%	158	36,92%	40	9,34%	158	36,92%	100	23,36%	129	30,14%
Número de ACERTOS na predição	311	72,68%	170	39,72%	325	75,95%	184	43%	273	63,78%	180	42,05%
Número de ERROS na predição	63	14,71%	87	20,33%	49	11,44%	73	17,05%	28	6,55%	80	18,70%
Número de NEUTROS na predição	54	12,61%	171	39,95%	54	12,61%	171	39,95%	127	29,67%	168	39,25%

Fonte: Elaborada pelo autor.

formadas pelas Listas_Orig_100% e Listas_Orig_10%. Esse modelo possui um maior grau de explicabilidade e obtiveram bons resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT.

Tabela 83 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Laptop*.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**						**gBoED_Syntax**					
		Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,75238	0,71285 41-37	55%	0,76417 27-16	60%	0,71052 41-38	55%	0,71285 41-37	55%	0,76417 27-15	60%	0,71052 41-38	55%	0,73837 27-20	55%	0,75487 27-16	60%	0,71052 41-38	55%
C4.5-Gini	0,75476	0,75011 12-11	55%	0,75421 20-12	65%	0,74779 12-11	55%	0,75011 12-11	55%	0,75421 20-12	65%	0,74779 12-11	55%	0,76871 12-6	55%	0,75659 20-10	65%	0,74779 12-11	55%
KNN-Cosseno	0,85753	0,81085 31-26 $n = 17$	55%	0,82027 37-24 $n = 15$	55%	0,80853 31-26 $n = 17$	55%	0,81085 31-26 $n = 17$	55%	0,82492 37-25 $n = 15$	55%	0,80853 31-26 $n = 17$	55%	0,83892 44-30 $n = 11$	55%	0,81800 37-24 $n = 15$	55%	0,83194 31-23 $n = 17$	55%
KNN-Euclideana	0,85753	0,80847 32-27 $n = 17$	55%	0,82259 37-24 $n = 15$	55%	0,80615 32-27 $n = 17$	55%	0,80847 32-27 $n = 17$	55%	0,82724 37-25 $n = 15$	55%	0,80615 32-27 $n = 17$	55%	0,83654 44-30 $n = 11$	55%	0,82032 37-24 $n = 15$	55%	0,82956 32-24 $n = 17$	55%
MNB	0,87625	0,86927 15-14 $\alpha = 10^{-1}$	55%	0,86927 12-8 $\alpha = 10^{-1}$	55%	0,85997 15-14 $\alpha = 10^{-1}$	55%	0,86229 15-14 $\alpha = 10^{-1}$	55%	0,86927 12-8 $\alpha = 10^{-1}$	55%	0,86229 15-14 $\alpha = 10^{-1}$	55%	0,87392 13-10 $\alpha = 10^{-1}$	55%	0,87159 12-11 $\alpha = 10^{-1}$	55%	0,86926 13-12 $\alpha = 10^{-1}$	55%
SVM-Linear	0,86916	0,86683 2-1 $\gamma = 10^{-4}$ a 10^4	50%	0,86932 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^4	50%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^4	50%	0,86932 3-3 $\gamma = 10^{-4}$ a 10^4	50%	0,86451 2-2 $\gamma = 10^{-4}$ a 10^4	70%	0,86218 2-2 $\gamma = 10^{-4}$ a 10^4	50%	0,87165 3-2 $\gamma = 10^{-4}$ a 10^4	50%	0,86218 2-2 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,89020	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88344 3-0 $\gamma = 10^{-2}$	50%	0,88555 2-2 $\gamma = 10^{-2}$	50%	0,88555 3-2 $\gamma = 10^{-2}$	50%	0,88344 3-0 $\gamma = 10^{-2}$	50%	0,88555 3-2 $\gamma = 10^{-2}$	55%	0,88322 3-2 $\gamma = 10^{-2}$	50%	0,88560 3-1 $\gamma = 1$	50%	0,88322 3-2 $\gamma = 10^{-2}$	50%
SVM-RBF	0,88311	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88577 1-1 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88577 1-1 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88577 1-1 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

Em seguida são analisados os resultados obtidos pela coleção de documentos *SemEval 2015 Restaurant*. Nos gráficos da Figura 58 é apresentada uma comparação entre a quantidade de termos (ou expressões) e seus sinônimos presentes nas Listas_Orig_100%, Listas_Bert_100%, Listas_Syntax_100% e Listas_Orig_10%. No gráfico da Figura 58a pode-se observar as quantidades de termos do domínio, na Figura 58b observa-se as quantidades para identificadores da classe "*Positive*" e na Figura 58c observa-se as quantidades para identificadores da classe "*Negative*". É possível verificar na Figura 58a que a quantidade de termos do domínio extraída

Tabela 84 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2015 Laptop*.

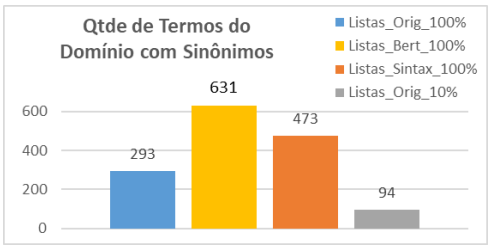
Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,75238	0,76644 146-60	90%	0,76417 27-16	60%	0,77569 146-58	90%	0,76417 27-15	60%	0,74075 41-29	55%	0,75487 27-16	60%
C4.5-Gini	0,75476	0,76173 108-40	75%	0,75421 20-12	65%	0,76185 151-59	90%	0,75421 20-12	65%	0,75011 12-8	55%	0,75659 20-10	65%
KNN-Cosseno	0,85753	0,84352 62-36 $n = 9$	60%	0,82027 37-24 $n = 15$	55%	0,85055 79-34 $n = 7$	60%	0,82492 37-25 $n = 15$	55%	0,85049 43-23 $n = 11$	55%	0,81800 37-24 $n = 15$	55%
KNN-Euclideana	0,85753	0,84114 63-37 $n = 9$	60%	0,82259 37-24 $n = 15$	55%	0,84817 80-35 $n = 7$	60%	0,82724 37-25 $n = 15$	55%	0,84811 44-24 $n = 11$	55%	0,82032 37-24 $n = 15$	55%
MNB	0,87625	0,87392 15-6 $\alpha = 10^{-1}$	55%	0,86927 12-8 $\alpha = 10^{-1}$	55%	0,87857 15-8 $\alpha = 10^{-1}$	55%	0,86927 12-8 $\alpha = 10^{-1}$	55%	0,87857 15-9 $\alpha = 10^{-1}$	55%	0,87159 12-11 $\alpha = 10^{-1}$	55%
SVM-Linear	0,86916	0,86910 25-16 $\gamma = 10^{-4}$ a 10^{4}	60%	0,86932 3-3 $\gamma = 10^{-4}$ a 10^{4}	50%	0,87602 122-55 $\gamma = 10^{-4}$ a 10^{4}	85%	0,86932 3-3 $\gamma = 10^{-4}$ a 10^{4}	50%	0,86683 2-1 $\gamma = 10^{-4}$ a 10^{4}	50%	0,87165 3-2 $\gamma = 10^{-4}$ a 10^{4}	50%
SVM-Polinomial	0,89020	0,88316 3-3 $\gamma = 10^{-2}$	50%	0,88344 3-0 $\gamma = 10^{-2}$	50%	0,88554 3-2 $\gamma = 10^{-2}$	50%	0,88344 3-0 $\gamma = 10^{-2}$	50%	0,88554 3-2 $\gamma = 10^{-2}$	50%	0,88560 3-1 $\gamma = 1$	50%
SVM-RBF	0,88311	0,88078 26-10 $\gamma = 1$	60%	0,88577 1-1 $\gamma = 10^{-1}$	50%	0,88311 26-9 $\gamma = 1$	60%	0,88577 1-1 $\gamma = 10^{-1}$	50%	0,88073 2-2 $\gamma = 10^{-1}$	50%	0,88577 1-1 $\gamma = 10^{-1}$	50%

Fonte: Elaborada pelo autor.

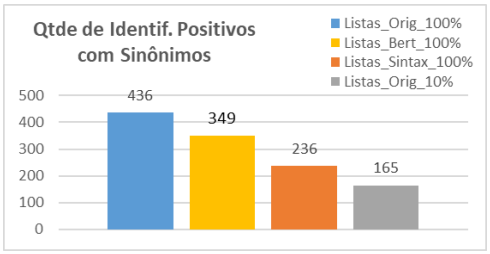
pelo método de extração de termos baseado em modelos de linguagem BERT é superior a todas as outras listas. Na Figura 58b verifica-se que a quantidade de identificadores da classe positiva é maior do que Listas_Syntax_100% e Listas_Orig_10% e, maior do que Listas_Orig_100%. Na Figura 58c verifica-se que a quantidade de identificadores da classe negativa é levemente maior do que Listas_Syntax_100% e Listas_Orig_10% e, menor do que Listas_Orig_100%.

Nas Tabelas 85 e 86 é apresentada a representatividade para o conjunto de documentos *SemEval 2015 Restaurant*. Na Tabela 85 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Orig_100%, Listas_Bert_100% e Listas_Orig_10%. Na Tabela 86 é feita a comparação entre a representatividade das gBoEDs construídas a partir da Listas_Syntax_100% e Listas_Bert_100%. Em ambas as tabelas é possível observar que, para essa coleção de documentos as representações gBoED_Freq e gBoED_Dist formadas a partir das Listas_Bert_100% obtiveram níveis altos de representatividade, aproximando-se das versões geradas por Listas_Syntax_100%, que obteve a maior representatividade. Para gBoED_Syntax formada a partir das Listas_Bert_100% foram obtidos níveis intermediários de representatividade. Na representação gBoED_Freq e gBoED_Dist, Listas_Bert_100% obteve 95,93% dos documentos representados, Listas_Orig_100% obteve 95,35%, Listas_Syntax_100% obteve 96,81% e Listas_Orig_10% obteve 90,70%. Na representação gBoED_Syntax, Listas_-Bert_100% obteve 72,67% dos documentos representados, Listas_Orig_100% obteve 86,63%, Listas_Syntax_100% obteve 76,64% e Listas_Orig_10% obteve 77,33%.

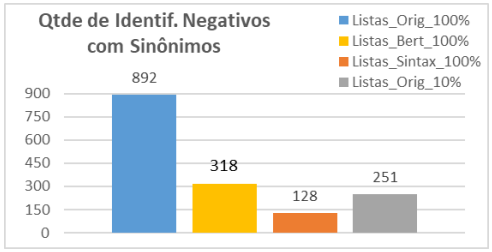
Figura 58 – Gráficos de quantidade de termos por tipo de lista - Coleção *SemEval 2015 Restaurant*.



(a) Termos do Domínio



(b) Identificadores de Classe "*Positive*"



(c) Identificadores de Classe "*Negative*"

Fonte: Elaborada pelo autor.

Com relação à quantidade de acertos na predição verifica-se de Listas_Bert_100% não obteve boa taxa de acertos. Na representação gBoED_Dist, Listas_Bert_100% obteve 70,36%, Listas_Orig_100% obteve 78,20%, Listas_Syntax_100% obteve 81,70% e Listas_Orig_10% obteve 74,13%. Na representação gBoED_Freq, Listas_Bert_100% obteve 69,77%, Listas_Orig_-100% obteve 78,20%, Listas_Syntax_100% obteve 79,08% e Listas_Orig_10% obteve 71,81%. Na representação gBoED_Syntax, Listas_Bert_100% obteve 51,45%, Listas_Orig_100% obteve 67,45%, Listas_Syntax_100% obteve 63,78% e Listas_Orig_10% obteve 61,06%. Portanto, as melhores taxas de acerto foram obtidas em cada representação por gBoED_Freq e gBoED_Dist formadas pelas Listas_Syntax_100% e por gBoED_Syntax formada pela Listas_Orig_100%.

Tabela 85 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Restaurant*.

	gBoED_Freq						**gBoED_Dist**						**gBoED_Syntax**					
	Listas_Orig 100%		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**		**Listas_Orig 100%**		**Listas_Bert 100%**		**Listas_Orig 10%**	
Qtde de documentos representados	328	95,35%	330	95,93%	312	90,70%	328	95,35%	330	95,93%	312	90,70%	298	86,63%	250	72,67%	266	77,33%
Qtde de documentos sem representação	16	4,65%	14	4,07%	32	9,30%	16	4,65%	14	4,07%	32	9,30%	46	13,37%	94	27,33%	78	22,67%
Número de ACERTOS na predição	269	78,20%	240	69,77%	247	71,81%	269	78,20%	242	70,36%	255	74,13%	232	67,45%	177	51,45%	210	61,06%
Número de ERROS na predição	45	13,08%	82	23,84%	51	14,82%	45	13,08%	80	23,25%	43	12,50%	34	9,88%	52	15,11%	33	9,59%
Número de NEUTROS na predição	30	8,72%	22	6,39%	46	13,37%	30	8,72%	22	6,39%	46	13,37%	78	22,67%	115	33,44%	101	29,36%

Fonte: Elaborada pelo autor.

Nas Tabelas 87 e 88 são apresentadas as melhores acurácias obtidas em cada algoritmo aplicado à coleção de documentos *SemEval 2015 Restaurant*. Em ambas, os resultados de maior

Tabela 86 – Representatividade das gBoEDs usando Listas_Bert_100% em comparação com Listas_-Syntax_100%, no conjunto de dados *SemEval 2015 Restaurant*.

	gBoED_Freq				**gBoED_Dist**				**gBoED_Syntax**			
	Listas_Syntax 100%		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**		**Listas_Syntax 100%**		**Listas_Bert 100%**	
Qtde de documentos representados	333	96,81%	330	95,93%	333	96,80%	330	95,93%	328	76,64%	250	72,67%
Qtde de documentos sem representação	11	3,19%	14	4,07%	11	3,19%	14	4,07%	100	23,36%	94	27,33%
Número de ACERTOS na predição	272	79,08%	240	69,77%	281	81,70%	242	70,36%	273	63,78%	177	51,45%
Número de ERROS na predição	51	14,82%	82	23,84%	42	12,20%	80	23,25%	28	6,55%	52	15,11%
Número de NEUTROS na predição	21	6,10%	22	6,39%	21	6,10%	22	6,39%	127	29,67%	115	33,44%

Fonte: Elaborada pelo autor.

destaque obtidos pelas representações construídas a partir das Listas_Bert_100% estão indicados linha cinza e se dá pelo algoritmo ***Support Vector Machine* – SVM, kernel Linear, enriquecido pela representação gBoED_Freq, gBoED_Dist e gBoED_Syntax**. Nesse cenário os resultados obtidos foram de 88,577% de acurácia, Medida-F1 de 88,655%, $\gamma = 10^{-4}$ a 10^{4}, grau de confiança de 50%, com 1 documentos sendo consultados e 0 documentos reclassificados. Para esse algoritmo, gBoED_Freq com Listas_Orig_100% obtiveram os melhores resultados, seguidos Listas_Orig_10%, Listas_Bert_100% e Listas_Syntax_100%. Em gBoED_Dist com Listas_-Orig_100% obtiveram os melhores resultados, seguidos Listas_Syntax_100%, Listas_Bert_100% e Listas_Orig_10%. Em gBoED_Syntax, Listas_Bert_100% obtive o melhor resultado, seguido de Listas_Syntax_100%, Listas_Orig_100% e Listas_Orig_10%. Outro destaque de Listas_-Bert_100% se dá em C4.5-Gini (enriquecido com gBoED_Freq, gBoED_Dist e gBoED_Syntax), obtendo resultado intermediário com acurácia maior do que a BoW e porém menor do que Listas_Orig_100%. Esse modelo possui um maior grau de explicabilidade e obtiveram bons resultados aos serem enriquecidos pelas representações formadas pelas listas extraídas usando o método semiautomático baseado em BERT.

5.4 Considerações finais

Neste capítulo foi proposto e aplicado um método para extração de termos baseado em modelos de linguagem BERT composto por 4 etapas principais. Assim como no modelo proposto no Capítulo 4, o principal objetivo do modelo de extração de termos baseado em BERT é o processo de extração mais automatizado auxiliando o trabalho dos especialistas de domínio, responsáveis por gerar a base de conhecimento necessária para a construção das representações semanticamente enriquecidas por expressões do domínio.

A primeira etapa do método é composta pela rotulação e entrada de listas de termos do domínio e identificadores de classe de forma manual, como base para o treinamento do modelo

Tabela 87 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Orig_100% e Listas_Orig_10%, no conjunto de dados *SemEval 2015 Restaurant*.

Algoritmos	BoW	gBoED_Freq						gBoED_Dist						gBoED_Syntax					
		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%		Listas_Orig 100%		Listas_Bert 100%		Listas_Orig 10%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,81723	0,81403 30-8	85%	0,81042 3-3	50%	0,81118 30-7	55%	0,81983 30-7	85%	0,81042 3-3	50%	0,81697 30-6	60%	0,81117 30-11	85%	0,80748 3-3	50%	0,78789 1-1	85%
C4.5-Gini	0,77328	0,78168 344-96	100%	0,77555 1-1	75%	0,72672 272-46	95%	0,81983 30-7	85%	0,77555 1-0	75%	0,74126 272-46	100%	0,78789 1-1	85%	0,77555 1-1	75%	0,81403 1-1	75%
KNN-Cosseno	0,81378	0,84303 81-29 $n = 11$	65%	0,79378 39-16 $n = 9$	60%	0,81378 28-16 $n = 13$	55%	0,78176 344-97 $n = 9$	100%	0,79378 39-16 $n = 9$	60%	0,82563 96-38 $n = 5$	60%	0,83411 79-42 $n = 11$	65%	0,79361 15-8 $n = 17$	55%	0,81092 39-22 $n = 11$	55%
KNN-Euclideana	0,81378	0,84303 81-29 $n = 11$	65%	0,79378 39-16 $n = 9$	60%	0,81378 28-16 $n = 13$	55%	0,85168 81-30 $n = 11$	65%	0,79378 39-16 $n = 9$	60%	0,82563 96-38 $n = 5$	55%	0,83411 79-42 $n = 11$	65%	0,79361 15-8 $n = 17$	55%	0,81092 39-22 $n = 11$	55%
MNB	0,86908	0,88084 34-12 $\alpha = 10^{-2}$	65%	0,86345 11-6 $\alpha = 10^{-2}$	55%	0,87773 11-5 $\alpha = 10^{-2}$	55%	0,88076 34-10 $\alpha = 10^{-2}$	65%	0,86345 11-6 $\alpha = 10^{-2}$	55%	0,87773 11-5 $\alpha = 10^{-2}$	55%	0,88067 11-7 $\alpha = 10^{-2}$	55%	0,86059 11-9 $\alpha = 10^{-2}$	55%	0,87478 11-7 $\alpha = 10^{-2}$	55%
SVM-Linear	0,87765	0,89513 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%	0,88933 50-15 $\gamma = 10^{-4}$ a 10^4	70%	0,89513 50-14 $\gamma = 10^{-4}$ a 10^4	70%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%	0,86398 50-14 $\gamma = 10^{-4}$ a 10^4	70%	0,85453 92-50 $\gamma = 10^{-4}$ a 10^4	80%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%	0,85436 15-10 $\gamma = 10^{-4}$ a 10^4	55%
SVM-Polinomial	0,88059	0,89504 50-13 $\gamma = 10^{-1}$	70%	0,88059 2-0 $\gamma = 10^{-1}$	50%	0,88639 50-13 $\gamma = 10^{-1}$	70%	0,89218 50-10 $\gamma = 10^{-1}$	70%	0,88059 2-0 $\gamma = 10^{-1}$	55%	0,89210 50-10 $\gamma = 10^{-1}$	55%	0,88369 2-1 $\gamma = 1$	50%	0,87773 2-1 $\gamma = 10^{-1}$	50%	0,88084 2-2 $\gamma = 1$	50%
SVM-RBF	0,82261	0,83983 139-29 $\gamma = 1$	90%	0,81681 1-1 $\gamma = 1$	50%	0,82563 27-3 $\gamma = 1$	60%	0,84588 166-30 $\gamma = 1$	95%	0,81681 1-1 $\gamma = 1$	50%	0,82857 27-4 $\gamma = 1$	60%	0,82277 16-8 $\gamma = 1$	55%	0,81395 1-0 $\gamma = 1$	50%	0,82268 1-1 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

Tabela 88 – Melhores acurácias dos classificadores gerados pelas gBoEDs usando Listas_Bert_100% comparadas com Listas_Syntax_100%, no conjunto de dados *SemEval 2015 Restaurant*.

Algoritmos	BoW	gBoED_Freq				gBoED_Dist				gBoED_Syntax			
		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%		Listas_Syntax 100%		Listas_Bert 100%	
	Acc	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf	Acc	Conf
C4.5-Entropia	0,81723	0,81714 30-3	85%	0,81042 3-3	50%	0,82865 158-19	90%	0,81042 3-3	50%	0,81117 30-11	85%	0,80748 3-3	50%
C4.5-Gini	0,77328	0,79050 344-81	100%	0,77555 1-1	75%	0,81672 344-74	100%	0,77555 1-0	75%	0,81403 1-1	75%	0,77555 1-1	75%
KNN-Cosseno	0,81378	0,82563 46-16 $n = 9$	60%	0,79378 39-16 $n = 9$	60%	0,82571 126-38 $n = 55$	75%	0,79378 39-16 $n = 9$	60%	0,83403 28-16 $n = 13$	55%	0,79361 15-8 $n = 17$	55%
KNN-Euclideana	0,81378	0,82563 46-16 $n = 9$	60%	0,79378 39-16 $n = 9$	60%	0,83151 96-30 $n = 5$	60%	0,79378 39-16 $n = 9$	60%	0,83403 28-16 $n = 13$	55%	0,79361 15-8 $n = 17$	55%
MNB	0,86908	0,87210 34-13 $\alpha = 10^{-2}$	65%	0,86345 11-6 $\alpha = 10^{-2}$	55%	0,87504 34-11 $\alpha = 10^{-2}$	65%	0,86345 11-6 $\alpha = 10^{-2}$	55%	0,88369 34-14 $\alpha = 10^{-2}$	65%	0,86059 11-9 $\alpha = 10^{-2}$	55%
SVM-Linear	0,87765	0,88647 42-8 $\gamma = 10^{-4}$ a 10^4	65%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%	0,89226 42-9 $\gamma = 10^{-4}$ a 10^4	70%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%	0,88050 20-9 $\gamma = 10^{-4}$ a 10^4	55%	0,88655 1-0 $\gamma = 10^{-4}$ a 10^4	50%
SVM-Polinomial	0,88059	0,88655 26-4 $\gamma = 1$	60%	0,88059 2-0 $\gamma = 10^{-1}$	50%	0,89226 20-4 $\gamma = 10^{-1}$	55%	0,88059 2-0 $\gamma = 10^{-1}$	55%	0,88344 17-11 $\gamma = 10^{-1}$	55%	0,87773 2-1 $\gamma = 10^{-1}$	50%
SVM-RBF	0,82261	0,83689 140-20 $\gamma = 1$	90%	0,81681 1-1 $\gamma = 1$	50%	0,84882 172-26 $\gamma = 1$	95%	0,81681 1-1 $\gamma = 1$	50%	0,82546 3-2 $\gamma = 1$	50%	0,81395 1-0 $\gamma = 1$	50%

Fonte: Elaborada pelo autor.

de extração. A segunda etapa é composta por uma etapa de pré-processamento e treinamento do modelos de extração de termos. Nesta etapa, cada tipo de lista exigem um modelo diferente. A terceira etapa corresponde ao próprio modelo de extração gerado, a entrada de novos documentos e a extração de novos conjuntos de termos. Por último, os especialista realizam uma limpeza e organização das listas de termos finais.

O método foi avaliado em 10 coleções de documentos que se diferem em domínio, idioma e tipo de classificação. Como o objetivo do método é tornar o processo de extração mais automatizado, auxiliando o trabalho dos especialistas, o treinamento do modelo foi realizado a partir de listas anotado por apenas 10% dos documentos e aplicado em 100% dos documentos contidos em cada coleção. Esse cenário foi identificado como Listas_Bert_100%. A partir desse conjunto de listas, foram geradas as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax. As representações foram aplicadas ao Método de Classificação Semanticamente Enriquecido por Expressões do Domínio. Foram verificas as quantidades de termos extraídas, a representatividade de cada representação usando as Listas_Bert_100% e as melhores acurácias dos modelos obtidos no método de classificação.

Com relação à extração de termos é possível avaliar que, de modo geral, as listas de termos geradas usando o método de extração baseado em modelos de linguagem BERT obtiveram uma quantidade de termos bastante superior às Listas_Syntax_100%, construídas de pelo método de extração baseado em regras morfossintáticas. Em alguns caso como *HuLiu 2004* (Termos do domínio), *SemEval 2014* (Termos do domínio), *SemEval 2014 Laptop*, *SemEval 2014 Restaurant* (Termos do domínio), *SemEval 2015 Hotel*, *SemEval 2015 Laptop*e *SemEval 2015 Restaurant* as listas atingiram quantidade de termos muito próximas ou superiores às Listas_Orig_100%.

Ao analisar a representatividade das listas e das representações verifica-se que o uso das Listas_Bert_100% obteve boa representação do conjunto de dados para construir as representações gBoED_Freq e gBoED_Dist. Com relação à quantidade de documentos representados, os destaques são para *SemEval 2014 Laptop*, *SemEval 2015*, *SemEval 2015 Hotel* e *SemEval 2015 Restaurant*. Para gBoED_Syntax a representatividade não foi satisfatória. Com relação à quantidade de acertos na predição, de maneira geral, as representações obtiveram níveis intermediários de acertos. Importante destacar também que, em *SemEval 2015 Laptop* as representações formadas pelas listas originais obtiveram valores extremamente baixos na quantidade de documentos representados e na taxa de acertos. As representações formadas pelas listas geradas pelos dois métodos semiautomáticos permitiram uma melhora significativa na representatividade desta coleção.

Nos experimentos utilizando o Método de Classificação Semanticamente Enriquecido por Expressões do Domínio foram avaliados diferentes cenários com as representações gBoED_Freq, gBoED_Dist e gBoED_Syntax, construídas a partir das Listas_Orig_100%, Listas_Syntax_-100%, Listas_Bert_100% e Listas_Orig_10%. Os resultados foram avaliados comparando as melhores acurácias obtidas nos diferentes cenário por modelos construídos por 4 algoritmos principais. Assim como nos experimentos dos Capítulos 3 e 4, os melhores resultados utilizando enriquecimento por gBoED_Bert ocorreram nos modelos gerados pelo algoritmo SVM. De modo semelhante ao nível de representatividade, na maioria dos casos os modelos enriquecidos pelas representações gBoED_Freq, gBoED_Dist e gBoED_Syntax construídas a partir das Listas_Bert_100% apresentaram resultados intermediários às representações construídas a partir

das Listas_Orig_100%, Listas_Syntax_100% e Listas_Orig_10%. Os principais destaques são os resultados obtidos nas coleções *B2W Reviews 2019 Info*, *SemEval 2014 Laptop*, *SemEval 2014 Restaurant* (gBoED_Syntax), *SemEval 2015* e *SemEval 2015 Laptop* com enriquecimento pelas representações construídas a partir de Listas_Bert_100%. Nesses casos, além dos resultados obtidos pelo algoritmo SVM, diversos outros algoritmos, considerados mais explicáveis, obtiveram resultados significativos com o enriquecimento pelas representações construídas a partir de Listas_Bert_100%.

Nesse capítulo, foi apresentado um método semiautomático para extração de termos do domínio e identificadores de classe usando modelo de linguagem BERT, bem como a análise de impacto do uso das listas extraídas pelo método semiautomático no Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. Esse resultado está relacionado às questões de pesquisa **Q1** e **Q2** e aos objetivos específicos **O2** e **O3**.

No Capítulo 6 é apresentado um estudo de caso de aplicação Método de Classificação Semanticamente Enriquecido por Expressões do Domínio, bem como as representações enriquecidas em uma processo de tomada de decisão em pedidos de acesso à informação de origem da Controladoria-Geral da União (CGU).

CAPÍTULO

6

ESTUDO DE CASO: CLASSIFICAÇÃO SEMÂNTICA EM PEDIDOS DE INFORMAÇÃO

6.1 Considerações iniciais

Nesse capítulo é abordado o estudo de caso do uso do método de classificação semântica aplicado em apoio ao processo decisório de pedidos de acesso à informação no âmbito da Controladoria-Geral da União (CGU) realizado em parceria com a aluna do Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI) de 2020 e especialista na CGU, Flávia Lemos Santos Xavier. Além do apoio dado ao órgão governamental, com a automatização do processo e redução de custos operacionais para a administração pública, o principal objetivo desse estudo foi verificar a viabilidade e o desempenho dos métodos de classificação semântica apresentados nesta tese em uma aplicação real e apoiado por especialistas da área.

6.2 Estudo de Caso em Pedidos de Informação

De acordo com Mendel (2009), em seu estudo do direito comparado sobre a liberdade de acesso à informação no mundo, o direito à informação ou o direito ao saber é um tema cada vez mais constante no discurso dos especialistas em desenvolvimento, da sociedade civil, dos acadêmicos, da mídia e dos governos. Nos últimos anos, houve uma verdadeira revolução no direito à informação, que é comumente compreendido como o direito de acesso à informação mantida por órgãos públicos. Segundo o estudo de Mendel (2009), em 1990, nenhuma organização intergovernamental reconhecia o direito à informação. Atualmente, mais de 70 países e todos os bancos multilaterais de desenvolvimento, além de uma série de outras instituições

financeiras internacionais adotaram políticas de divulgação de informações, visando melhorar a prestação de contas de gestores e o controle social. E para além de ser uma medida de governança administrativa, este direito é hoje considerado cada vez mais como um direito humano fundamental.

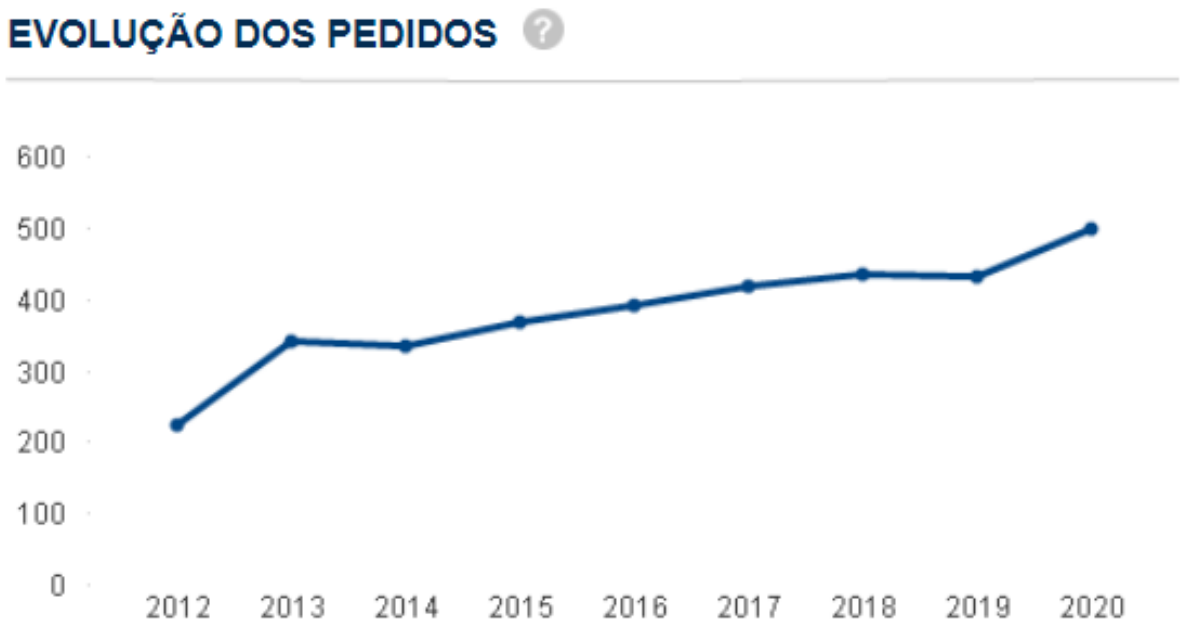
No Brasil, a Constinuição de 1988, chamada Constituição Cidadã, garante em seu artigo 5º, XXXIII, o direito básico de acesso à informação. A Lei de Acesso à Informação (LAI) (Lei nº 12.527/2011) (BRASIL, 2020), regulamenta este direito constitucional, para estabelecer que os órgãos e entidades públicas devam garantir um processo transparente de gestão da informação, por amplo acesso e divulgação; disponibilidade, autenticidade e integridade; proteção de informações confidenciais e informações pessoais e, eventualmente, restrição de acesso à informação, nos casos em que a publicidade de tais informações pode colocar em risco a segurança da sociedade ou do Estado.

A Lei de Transparência e Acesso à Informação Pública foi aprovada no final de 2011 e entrou em vigor em 2012 no Brasil. De acordo com essa lei, o princípio da máxima transparência é a regra e a prestação de informações é gratuita. Os pedidos de informação dos cidadãos à administração pública devem ser respondidos no prazo de 20 dias, prorrogáveis por mais 10 dias, mediante justificativa expressa nesse caso. Qualquer negativa de acesso à informação é exceção e deve ser justificada, com direito à interposição de recursos no órgão público e, em terceira instância recursal, à Controladoria-Geral da União, que é o órgão garantidor desse direito no âmbito do Poder Executivo Federal.

O potencial de aplicação desta pesquisa evidencia-se ao apresentar os dados da crescente demanda da implementação efetiva do direito de acesso à informação no Brasil. De acordo com o Painel da Lei de Acesso à Informação, desenvolvido pela CGU (2020), somente os órgãos e entidades federais receberam um total de 974.078 pedidos de acesso à informação desde maio de 2012, início da vigência da Lei de Acesso à informação no Brasil. No primeiro ano da implementação, foram registrados 55.212 pedidos e, até o dia 15 de dezembro de 2020, foram registrados um total de 150.036 pedidos, o que corresponde a mais de 170% de crescimento de demanda nos últimos 8 anos. Esses dados significam que a demanda diária de pedidos de acesso à informação ao Poder Executivo Federal que foi de 224 pedidos em 2012, em 2020 chegou a 500 pedidos, conforme apresentado na Figura 59.

Este estudo de caso possui dois objetivos principais. O primeiro deles é validar, em um caso real, o método de classificação semântica desenvolvido nesta tese. O segundo é contribuir com um dos importantes mecanismos democráticos de participação social no Brasil: o exercício do direito de acesso à informação, introduzindo a Inteligência Artificial como parte do processo. Como dito anteriormente, o estudo de caso se dedica ao desenvolvimento de um modelo que irá apoiar a classificação dos novos pedidos de acesso à informação, interpostos à Controladoria-Geral da União (CGU), conforme a probabilidade de eles serem concedidos ou de serem negados ao cidadão.

Figura 59 – Evolução da média diária dos pedidos de acesso à informação no Poder Executivo Federal.



Fonte: Extraída de (CGU, 2020)

O classificador visa contribuir com a automação e com o alcance de maior eficiência da fase inicial do processo de resposta aos pedidos de acesso à informação à CGU. Esta fase consiste em classificar e realizar a triagem dos pedidos, com base em pesquisas dos precedentes jurisprudenciais que contenham semelhanças temáticas e circunstanciais aos novos pedidos de acesso. Esta fase do trabalho precede a elaboração de pareceres técnicos realizados pelos servidores da equipe da CGU, que visam fundamentar a decisão do Ouvidor-Geral da União Adjunto, a quem cabe decidir em última instância recursal no órgão pelo tipo de resposta ao pedido de acesso à informação.

Portanto, neste estudo de caso foi desenvolvido um modelo preditivo, transparente e replicável que possibilite ao servidor da CGU contar com a classificação automatizada dos novos pedidos de acesso à informação, no processo de triagem, com base no banco de precedentes dos pedidos à CGU. O sucesso deste resultado possibilita caracterizar o método como um incremento aos métodos tradicionais de classificação e validado como um agregador de conhecimento à tomada de decisão no processo de triagem de pedidos de informação da CGU. A capacidade de generalização do método também permite que o processo possa ser adaptado a diversos outros casos, bem como às necessidades de outros órgãos do Poder Executivo Federal e até mesmo de todo o País.

6.3 Classificação de pedidos de acesso à informação

Nessa seção apresenta-se o processo de classificação e validação realizado no estudo de caso de identificação de pedidos de acesso à informação da CGU. Inicia-se com a apresentação da base de dados dos pedidos e, na sequência, é apresentado o precesso de desenvolvimento do modelo e sua validação.

Tendo em vista que se trata de um processo de apoio à tomada de decisão referente à concessão de um direito fundamental, os critérios a serem considerados devem ser claros para todos os envolvidos: a administração pública e os administrados. E é por isso que quanto maior o grau de transparência do classificador mais interessante a solução será para o processo.

6.3.1 A base de dados

A base de dados de origem a base dados era composta por 3.617 registros de pedidos de acesso à informação realizados à CGU no período de 2016 a 2020, em formato semiestruturado do "Relatório de pedidos de acesso à informação e solicitantes" no âmbito do Poder Executivo Federal que estão disponíveis na plataforma FalaBr (2020), com atualização dinâmica no Sistema Eletrônico de Serviço de Informação ao Cidadão (e-SIC) (eSIC, 2020) e atualização periódica no Portal Brasileiro de Dados Abertos[1]. Esse recorte temporal foi escolhido considerando a disponibilidade dos dados que continham o campo 'DetalhamentoSolicitacao' até o ano que antecedeu a realização do estudo. Os 3 principais campos considerados nesta base de dados foram:

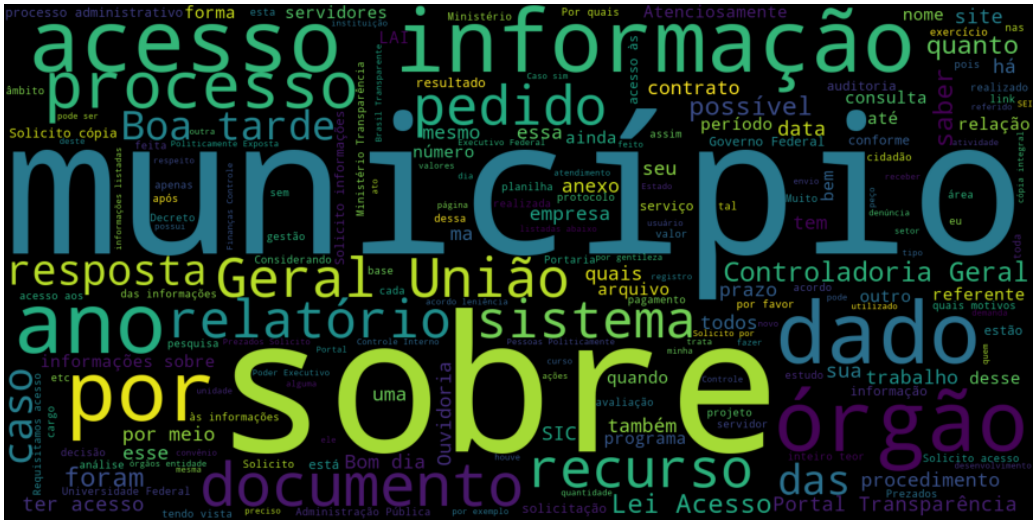
- IdPedido: é o identificador único do pedido;
- DetalhamentoSolicitacao: descrição textual da íntegra do pedido;
- TipoResposta: campo que possui a resposta dada ao pedido pela equipe de análise da CGU. É o atributo utilizado como rótulo para o treinamento do classificador.

Como primeira exploração dos dados, uma nuvem de palavras referente às palavras contidas na coluna de detalhamento dos pedidos à CGU, na qual o tamanho de cada palavra está relacionado à frequência em que parecem no conjunto de textos. A nuvem de palavras é aparesentada na Figura 60.

É possível identificar que as palavras mais frequentes são palavras como "município", "acesso", "informação", "boa tarde" ou "sobre", palavras essas que não permitem ainda extrair muita informação para a construção de um classificador para a triagem dos pedidos. Assim, uma palavra que aparece muito em uma frase poderia parecer importante, mas ao aparecer em 100% dos textos analisados, se torna uma informação irrelevante para esta pesquisa. Por essa razão,

[1] Portal Brasileiro de Dados Abertos: https://dados.gov.br

Figura 60 – Nuvem de Palavras do Detalhamento dos Pedidos.



Fonte: Xavier (2021).

nesse estudo de caso, a etapa de pré-processamento é umas das mais importantes do processo de mineração.

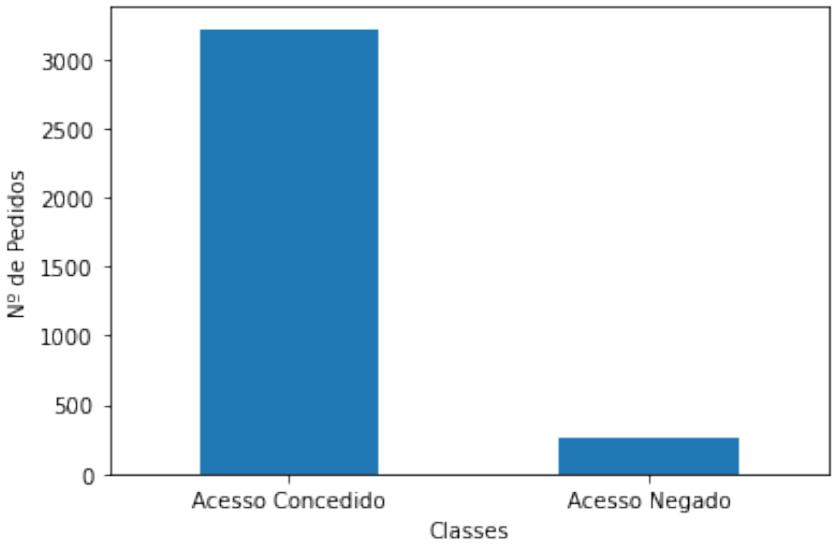
6.3.2 Pré-processamento

Originalmente, a base dados era composta por 3.617 registros de pedidos de acesso à informação. Na etapa de pré-processamento, inicialmente os dados passaram por um processo de limpeza, nos quais removidos aqueles cujos tipos de respostas eram *"Informação inexistente"*, *"Não se trata de solicitação de informação"*, *"Órgão não tem competência para responder sobre o assunto"*, *"Pedido duplicado ou repetido"* e *"Acesso Parcialmente Concedido"*. Também foram removidos pedidos que continham dados ruidosos, gerados devido a falhas na coleta de dados ou dados duplicados. Por último, foram removidos pedidos negados cuja causa é muito difícil de prever pois não estão relacionados com sigilos ou restrições, como é o caso dos pedidos genéricos, incompreensíveis ou que demandam tratamento adicional às competências da CGU. Como resultado do processo de limpeza e remoção de pedidos a base passou a contar com 3.473 pedidos. Nesses pedidos, as classes resultantes e consideradas no processo de treinamento dos classificadores foram *"Acesso Concedido"* e *"Acesso Negado"*.

Após o processo de limpeza dos dados dos pedidos, detectou-se um esperado desbalanceamento das classes. Dos 3.473 pedidos resultantes da base limpa, 3.217 pedidos tiveram como resposta o acesso concedido (93% do total) e 256 tiveram como resposta o acesso negado pela CGU (7% do total). O gráfico da Figura 61 ilustra o desbalanceamento do conjunto de dados.

Ignorar esse desbalanceamento evidente e treinar um algoritmo sem o pré-processamento desses dados, poderia impactar diretamente no enviesamento do classificador. A consequência desse desequilíbrio é que o modelo tende a predizer muitos falsos negativos em seus resultados.

Figura 61 – Tipos e Quantidade de Respostas com a base desbalanceada.



Fonte: Xavier (2021).

Ou seja, na prática ele pode classificar muito bem entradas para a classe majoritária ("Acessos Concedidos"), mas tende a ter um desempenho bastante inferior para calcular quando um novo pedido for da classe minoritária ("Acesso Negado").

Foram realizados experimentos preliminares de treinamento de classificadores usando BoW a partir da base desbalanceada que não resultaram em modelos com classificação precisa para a classe minoritária. Um exemplo disso é o treinamento do modelo de classificação gerado a partir do algoritmo KNN[2]. Neste teste, apesar dos resultados terem atingindo acurácia global de 94% e Medida-F1 global de 93%, a Medida-F1 para a classe majoritária foi de 97%, enquanto a Medida-F1 para a classe minoritária foi de apenas 40%. Embora a acurácia e a Medida-F1 globais tenham sido altos, os resultados de classificação da classe minoritária nos experimentos preliminares não permitiram prosseguir com a base desbalanceada tornando os resultados insatisfatórios para o estudo, como é possível observar na Matriz de confusão da Tabela 89.

Como próximo passo da etapa de pré-precessamento, foi adotada a abordagem de *under-sampling* para realizar o balanceamento das classes relacionadas aos pedidos de acesso à informação. A técnica de *under-sampling* torna-se mais adequada para dados com desbalanceamento muito grande, pois visa preservar ao máximo as características da classe minoritária e descarta informações da classe majoritária. Neste estudo, os exemplos da classe majoritária foram descartados de forma aleatória. Após o rebalanceamento das classes o conjunto de pedidos de informação passou a contar com 256 pedidos de cada classe, totalizando 512 pedidos.

Como último passo da etapa de pré-processamento, para a construção da representação textual no modelo *Bag of Words*(BoW), algumas técnicas foram aplicadas ao conjunto de textos

[2] Foram utilizados os seguintes parâmetros: *euclidean* e *cossine*, *n_neighbors*: entre 1 e 55.

Tabela 89 – Resultados do Modelo de Classificação KNN com a base desbalanceada.

Relatório de classificação:				
	Precisão	Revocação	Medida-F1	Suporte
Acesso Concedido	0.95	0.98	0.97	645
Acesso Negado	0.60	0.30	0.40	50
Acurácia do modelo	0.94			
Medida-F1	0.93			
Revocação	0.94			

Fonte: Adaptada de Xavier (2021).

de modo. São elas:

1. **Limpeza e padronização do texto:** o objetivo da limpeza dos textos dos pedidos é a remoção de partes identificadas com irrelevantes e remoção de dados ruidosos. Inicialmente foram removidos valores numéricos como números de protocolos, prazos dos atendimentos dos pedidos, informações sobre prorrogação das respostas. Realizou-se, também, a remoção de caracteres especiais, como *hashtags*, pontuação, e acentos. Em seguida, realizou-se a padronização de caixa, convertendo o texto para caracteres minúsculos. A saída é mostrada na Figura 62.

Figura 62 – Base de dados após a limpeza e padronização.

	IdPedido	DetalhamentoSolicitacao	TipoResposta	texto_limpo
0	1679692	O download de dados disponibilizado pela CGU (...	Acesso Concedido	o download de dados disponibilizado pela cgu h...
1	1680611	Dei entrada em 09/11/2015, na Regional da CGU ...	Acesso Concedido	dei entrada em na regional da cgu no estado d...
2	1681387	Solicito a inclusão da Portaria nº 50.253 de 1...	Acesso Concedido	solicito a inclusao da portaria n de de deze...
3	1682160	Solicito informações quanto ao número total de...	Acesso Concedido	solicito informacoes quanto ao numero total de...
4	1682294	Solicito copia de todos os documentos preparat...	Acesso Concedido	solicito copia de todos os documentos preparat...
...	...	...	...	...
507	2311299	Solicito arquivo digital em PDF Nota Técnica n...	Acesso Negado	solicito arquivo digital em pdf nota tecnica n...
508	2344555	Gostaria de confirmar a informação, proferida ...	Acesso Negado	gostaria de confirmar a informacao proferida n...
509	2379538	Olá, estou trabalhando em uma empresa de colet...	Acesso Negado	ola estou trabalhando em uma empresa de coleta...
510	2389979	Hoje (10/09/2020) recebi a DECISÃO (vide anexo...	Acesso Negado	hoje recebi a decisao vide anexo da cgu ao me...
511	2466445	Venho por meio desta, com fundamento na Lei 13...	Acesso Negado	venho por meio desta com fundamento na lei no...

Fonte: Xavier (2021).

2. **Remoção de *Stopwords* em português e de palavras com tamanho atípico:** foi realizada a remoção de *Stopwords* e palavras que por terem sido coletadas ou migradas incorretamente introduzem ruídos desnecessários, por exemplo, palavras que contêm

menos de 2 letras ou mais de 10 letras. O resultado dessa fase do pré-processamento é apresentado na Figura 63.

Figura 63 – Base de dados após remoção de *Stopwords*.

	IdPedido	DetalhamentoSolicitacao	TipoResposta	texto_limpo
0	1679692	O download de dados disponibilizado pela CGU (...	Acesso Concedido	download dados permite apenas formatos xml csv...
1	1680611	Dei entrada em 09/11/2015, na Regional da CGU ...	Acesso Concedido	dei entrada regional estado ceara pedido relat...
2	1681387	Solicito a inclusão da Portaria nº 50.253 de 1...	Acesso Concedido	inclusao portaria dezembro site
3	1682160	Solicito informações quanto ao número total de...	Acesso Concedido	quanto numero total cargos analista financas c...
4	1682294	Solicito copia de todos os documentos preparat...	Acesso Concedido	copia todos entendese lei todos documento toma...
...	...	...	...	...
507	2311299	Solicito arquivo digital em PDF Nota Técnica n...	Acesso Negado	arquivo digital pdf nota tecnica processo pois...
508	2344555	Gostaria de confirmar a informação, proferida ...	Acesso Negado	confirmar proferida dia hoje ago jair bolsonar...
509	2379538	Olá, estou trabalhando em uma empresa de colet...	Acesso Negado	ola empresa coleta dados dados sobre operacao ...
510	2389979	Hoje (10/09/2020) recebi a DECISÃO (vide anexo...	Acesso Negado	hoje recebi decisao vide anexo recurso instanc...
511	2466445	Venho por meio desta, com fundamento na Lei 13...	Acesso Negado	venho meio desta lei art lei art inciso federa...

Fonte: Xavier (2021).

3. **Radicalização:** como último passo, foi aplicada a técnica de "tokenização" e Radicalização (*stemming*) nos textos do conjunto de pedidos de modo a reduzir as palavras à sua forma raiz e, com isso remover variações de uma mesma palavras. O objetivo é contribuir para a reduzir o ruído e a dimensionalidade dos dados. Verifica-se na Figura 64 as diferentes saídas nas novas colunas chamadas "texto_tokens" e "texto_tokens_stem".

Figura 64 – Base de dados "Tokenizada" e Radicalizada.

	IdPedido	DetalhamentoSolicitacao	TipoResposta	texto_limpo	texto_tokens	texto_tokens_stem
0	1679692	O download de dados disponibilizado pela CGU (...	Acesso Concedido	download dados permite apenas formatos xml csv...	[download, dados, permite, apenas, formatos, x...	[download, dad, permit, apen, format, xml, csv...
1	1680611	Dei entrada em 09/11/2015, na Regional da CGU ...	Acesso Concedido	dei entrada regional estado ceara pedido relat...	[dei, entrada, regional, estado, ceara, pedido...	[dei, entr, region, est, ce, ped, relat, port,...
2	1681387	Solicito a inclusão da Portaria nº 50.253 de 1...	Acesso Concedido	inclusao portaria dezembro site	[inclusao, portaria, dezembro, site]	[inclusa, port, dezembr, sit]
3	1682160	Solicito informações quanto ao número total de...	Acesso Concedido	quanto numero total cargos analista financas c...	[quanto, numero, total, cargos, analista, fina...	[quant, numer, total, carg, anal, financ, cont...
4	1682294	Solicito copia de todos os documentos preparat...	Acesso Concedido	copia todos entendese lei todos documento toma...	[copia, todos, entendese, lei, todos, document...	[cop, tod, entendes, lei, tod, document, tom, ...
...	...	...	...	...	...	...
507	2311299	Solicito arquivo digital em PDF Nota Técnica n...	Acesso Negado	arquivo digital pdf nota tecnica processo pois...	[arquivo, digital, pdf, nota, tecnica, process...	[arqu, digit, pdf, not, tecn, process, poi, na...
508	2344555	Gostaria de confirmar a informação, proferida ...	Acesso Negado	confirmar proferida dia hoje ago jair bolsonar...	[confirmar, proferida, dia, hoje, ago, jair, b...	[confirm, profer, dia, hoj, ago, jair, bolsona...
509	2379538	Olá, estou trabalhando em uma empresa de colet...	Acesso Negado	ola empresa coleta dados dados sobre operacao ...	[ola, empresa, coleta, dados, dados, sobre, op...	[ola, empr, colet, dad, dad, sobr, operaca, bo...
510	2389979	Hoje (10/09/2020) recebi a DECISÃO (vide anexo...	Acesso Negado	hoje recebi decisao vide anexo recurso instanc...	[hoje, recebi, decisao, vide, anexo, recurso, ...	[hoj, receb, decisa, vid, anex, recurs, instan...
511	2466445	Venho por meio desta, com fundamento na Lei 13...	Acesso Negado	venho meio desta lei art lei art inciso federa...	[venho, meio, desta, lei, art, lei, art, incis...	[venh, mei, dest, lei, art, lei, art, incis, f...

Fonte: Xavier (2021).

Concluída a etapa de pré-processamento, foi gerada uma BoW com um total de 1962 atributos. Na Figura 65 observa-se um recorte da representação gerada.

Figura 65 – *Bag of Words* com Frequência das Palavras por Pedido.

	ab	abaix	abaixon	abert	abon	abr	abrang	abril	aca	academ	...	willy	xii	xingu	xingup	xl	xlsx	xml	yur	zel	zip
1679692	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	1	0	0	1
1680611	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1681387	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1682160	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1682294	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Fonte: Xavier (2021).

Com base na representação BoW gerada, observou-se que algumas das palavras mais frequentes na base de dados são: anexo, Brasil, prazo, gasto, auditoria, acordo etc. Em seguida, é apresentada a etapa de treinamento do modelo de classificação dos pedidos de acesso à informação.

6.3.3 Primeiros modelos de classificação de pedidos de informação

Como apresentado na Subseção 6.3.2, foi gerada uma representação BoW com o objetivo de treinar modelos de classificação para pedidos de acesso à informação na Controladoria Geral da União (CGU). Por se tratar de uma solução para apoio à tomada de decisão vale, ressaltar a importância da utilização de algoritmos que permitam ao ser humano, seja ele parte da equipe da CGU, requerente da informação ou qualquer outro interessado, compreender o conteúdo das bases de dados, das ações e dos critérios adotados pelo classificador. No domínio de aplicação deste estudo, existe uma necessidade de transparência dos algoritmos, de modo a minimizar os riscos de distorção, de imparcialidades embutidas ou de erros dos resultados que porventura possam prejudicar um indivíduo em detrimento de outro.

No entanto, para fins de comparação de resultados quantitativos também foram considerados algoritmos do tipo "caixa-preta". Essa análise comparativa do desempenho dos diferentes tipos de classificadores é importante visto que interpretabilidade ou explicabilidade versus desempenho dos modelos é um *trade-off* comum no aprendizado. Em geral, modelos mais complexos (como SVM e redes neurais) tendem a ter desempenho melhor do que os modelos mais transparentes (como as árvores de decisão – C4.5), no entanto o maior desempenho vem acompanhado de menor explicabilidade (GUNNING, 2017; LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020).

O estudo priorizou os algoritmos que, neste caso, possuem um maior nível de explicabilidade do ponto de vista de sua justificativa, para o suporte à decisão da autoridade da CGU.

Portanto, para a fase de Extração de Padrões os algoritmos utilizados foram: (i) C4.5 (ii) *K-nearest neighbor (KNN)* (iii) *Multinomial Naïve Bayes* (MNB) (iv) *Support Vector Machine* (SVM) , nas mesmas configurações daquelas utilizadas na Subseção 3.4.2. Foi utilizado também a técnica de validação *K-fold cross validation* com $K = 10$.

Para fins de comparação de resultados quantitativos foi realizado também o treinamento de classificadores utilizando Redes Neurais Profundas, como as Redes Neurais Convolucionais – RNCs – que contêm cinco tipos de camadas: camada de entrada, de convolução, de agrupamento, as completamente conectadas e as de saída - e algoritmos de Redes Recorrentes – GRU e LSTM. Para o treinamento dos modelos gerados pelos algoritmos baseados em Redes Neurais, foi considerada a utilização do método de representação *Word Embedding/ GloVe*: *Global Vectors for Word Representation*.

Inicialmente avaliou-se os modelos pela acurácia, para obter-se um melhor indicativo de investigação e, na sequência, pela Medida-F1. Esse nível de avaliação é importante pois o prejuízo de classificar incorretamente um pedido como "Acesso Concedido" ou como "Acesso Negado", em ambas situações poderia implicar prejuízos para a administração pública (custos desnecessários) e para o solicitante (frustração e perda da credibilidade institucional) e, em última instância, poderia inclusive implicar infrações administrativas e responsabilização administrativa e penal para a administração e para os servidores responsáveis.

Na Tabela 90, verifica-se que o melhor desempenho obtido pelo modelo baseado em BoW, foi obtido pelo algoritmo ***Support Vector Machine*, kernel Polinomial, com acurácia e Medida-F1 globais de 77%** . Verica-se também, que o resultado obtido pelo algoritmo SVM treinado a partir de BoW, supera o resultado obtidido pelas redes neurais profundas, cujo melhor resultado foi obtido pela Rede Neural Convolucional (RNC), com acurácia de 70,5% e Medida-F1 de 73%.

Tabela 90 – Comparação de Resultados de Algoritmos Transparentes e de Redes Neurais.

Algoritmos	**Parâmentros**	**Acurácia**	**F1**
C4.5	Gini e Euclideana	0.686	0.685
KNN	Euclidiana e Cosseno	0.768	0.767
MNB	$\alpha = 10^{-1}$	0.754	0.753
SVM	**Polinomial**	**0.770**	**0.770**
RNC	loss = "binary_crossentropy",	**0.705**	0.731
GRU	optimizer = "adam",	0.431	0.610
LSTM	batch_size = 16, epochs = 20	0.431	0.610

Fonte: Xavier (2021).

6.3.4 Aplicação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio em pedidos de acesso à informação

Modelos matemáticos são ferramentas poderosas para classificar textos em linguagem natural. No entanto, esses modelos acabam não conseguindo representar todos os detalhes da língua (WIDDOWS, 2004; SINOARA, 2018). Como apresentado na Seção 2.2, a tarefa de classificação podem ser dividida em diferentes níveis e a classificação de pedidos de acesso à informação pode ser considerada uma tarefa de nível semântico, ou seja, que requer mais do que o conjunto de palavras para gerar bons modelos de classificação. Por isso, torna-se necessária aplicação de abordagens alternativas à BoW de modo a elaborar uma representação dos textos com maior representatividade semântica.

Com base nas características do domínio, do idioma, e do nível de classificação que se deseja realizar, o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio foi escolhido como forma de incrementar os resultados obtidos nos experimentos da Subseção 6.3.3. Conforme explicado anteriormente no Capítulo 3, esse método de classificação tem como base o enriquecimento dos resultados de classificação pela representação gBoED formada por Expressões do Domínio. As expressões são formadas a partir do conhecimento do especialista e incorporadas à representação por meio de listas de termos de domínio e identificadores de classe.

Como base para a construção das representações gBoED_Freq e gBoED_Dist é necessária a identificação das palavras e expressões que irão compor as listas de termos, com base no tipo de classificação que se deseja realizar. Após a exploração e análise junto ao especialista na área de pedidos de acesso à informação do conjunto de textos da base, identificou-se que os Termos do Domínio são as palavras e expressões que indicam **perguntas** ou **pedidos** e que fazem referência ao **objeto das solicitações recebidas**, como por exemplo “solicito”, “requeiro”, “venho requerer”, “desejo saber”, “desejo conhecer”, etc. Sinônimos referentes a um conjunto de termos também foram adionados às listas.

Com relação aos indicadores da classe “Acesso Concedido”, toma-se como exemplo primeiramente termos e expressões que fazem referência a informações que podem ser concedidas ao público, por exemplo “salário”, “salários” e “remuneração” dos servidores públicos. A divulgação nominal da remuneração do servidor público foi outrora objeto de grande polêmica. A partir da edição da BRASIL (2020), entende-se que é legítima a publicação das remunerações dos servidores públicos (MARTINS; LOPES; CADEMARTORI, 2017). Por isso, tais indicadores e seus sinônimos foram incluídos nessa listagem.

No caso dos indicadores relacionados à classe “Acesso Negado”, os termos fazem referência às informações que não são permitidas de serem disponibilizadas ao público, como exemplo os pedidos recorrentes de disponibilização do código fonte da ferramenta “Análise de Licitações

e Editais" (ALICE) ou informações referentes a processos judiciais com sigilo. A ferramenta ALICE representa o procedimento executado durante o planejamento de ações de controle da CGU e do Tribunal de Contas da União (TCU), contendo toda a estratégia do trabalho a ser realizado e as evidências que servem de comprovação dos fatos durante o trabalho de campo das equipes. Esse tipo de pedido é considerado negado. Nessa mesma linha, é possível fundamentar a inclusão dos seguintes termos identificadores da classe "Acesso Negado": "relatório de auditoria"; "relatório final de auditoria"; "íntegra de relatório de auditoria"; "documentos de fiscalização'; "papéis de trabalho"; "papel de trabalho"; "nota técnica"; "notas técnicas"; "planejamento geral de auditoria", visto que essas informações são, em regra, também relacionadas a essas hipóteses legais de sigilo.

No Quadro 9 é possível verificar um exemplo mais completo de parte dos termos que compõem cada uma das listas.

Quadro 9 – Exemplos de termos do domínio e identificadores de classe em pedidos de acesso à informação.

Termos de Domínio	**Identificadores da Classe Positivo**	**Identificadores da Classe Negativo**
preciso de ajuda; preciso; solicito	programa de formação continuada em ouvidoria; PROFOCO	PAD; PADs; processo administrativo disciplinar; sindicância; investigação; em curso; em andamento
agradeceria que; agradeceria se	salário; salários; remuneração	relatório de auditoria; relatório final de auditoria;íntegra de relatório de auditoria; documentos de fiscalização; papéis de trabalho; papel de trabalho
disponibilizar	portal de transparência; site da transparência; grau de transparência; programa brasil transparente; escala brasil transparente; escala; ranking	extratos bancários
não encontrei	contrato; edital; licitação; gastos; orçamento	arquitetura detalhada do ALICE; código fonte; especificações produzidas para o software
desejo saber; desejo; desejo conhecer; desejo receber	programa de fortalecimento das ouvidorias; PROFORT	nome do denunciante; identidade do denunciante; cpf
informar; gentileza informar; vistas; dei entrada; informa	diário oficial da união; D.O.U.; DOU; boletim interno	procedimentos de caráter preparatório; documentos preparatórios; preparatório; preliminares

Fonte: Xavier (2021).

Nota – Para melhor entendimento, os termos utilizados nos exemplos deste quadro são apresentadas sem a aplicação de técnicas de pré-processamento definidas na Subseção 6.3.2.

Após a preparação das listas de termos, foi aplicado ao conjunto de pedidos de acesso à informação o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, apresentado no Capítulo 3. Na Subseção 6.3.5 são apresentados os resultados obtidos, bem como a utilização dos modelos gerados no âmbito do processo de triagem da CGU.

6.3.5 Resultados da aplicação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio em pedidos de acesso à informação

Nessa subseção são apresentados os resultados dos experimentos e modelos gerados utilizando o Método de Classificação Semanticamente Enriquecida por Expressões do Domínio aplicado ao conjunto de textos referentes aos pedidos de acesso à informação.

O primeiro resultado a ser observado está relacionado à representatividade e explicabilidade que as representações baseadas em expressões do domínio trazem ao conjunto de textos. Pode-se verificar no exemplo a seguir que as expressões do domínio podem ser consideradas como informações enriquecidas para a representação em relação ao texto original. O Quadro 10 apresenta um exemplo resumido das representações gBoED_Freq e gBoED_Dist para um trecho de um pedido de acesso à informação.

Tendo como base o texto original do pedido, podemos verificar que foi identificado o termo do domínio – “gostaria” – e o identificador de classe – “arquitetura_detalhada_do_alice”. A expressão formada foi “gostaria_1_arquitetura_detalhada_do_alice”. O valor “1” presente entre o termo do domínio e o identificador de classe indica que a expressão pertence à classe “Acesso Negado”. Pode-se observar que nesse exemplo, a expressão é um resumo das informações principais contidas no pedido. Como se trata de um texto curto, nesse caso, apenas uma expressão já contribui para uma boa explicabilidade e representatividade do conteúdo principal do pedido. Essas são as principais características de representações baseadas em termos ou expressões do domínio.

Outro resultado que é possível observar e que contribui para o desempenho do método de classificação é apresentado na Tabela 91. Esses resultados estão relacionados à completude da representação gBoED_Freq e gBoED_Dist, ou seja, o quanto as listas de termos e, consequentemente, as representações semanticamente enriquecidas realmente são capazes de representar a coleção de documentos como um todo e o qual o nível de acerto é possível atingir ao utilizá-las.

Verifica-se na Tabela 91 que no total de 512 pedidos existentes no conjunto de pedidos de informação da base balanceada, a representação gBoED usando Listas 100% conseguiu realizar a identificação de expressões do domínio e representar 467 pedidos (91,2% da base de dados). Apenas 45 pedidos (8,78%) não obtiveram representação. Usando Listas 10% representação gBoED conseguiu realizar a identificação de expressões do domínio e representar 464 pedidos (90,62% da base de dados). Quando comparadas as duas versões de listas é possível dizer que,

Quadro 10 – Exemplo de pedido de informação e suas representações no formato gBoED.

Pedido: "Gostaria de saber a arquitetura detalhada do ALICE (Análise de Licitações e Editais), de modo a entender quais seus componentes e formas de integração), e pedir o código fonte que foi usado para construir a ALICE, com instruções para replicação."

Neste pedido são considerados os seguintes termos extraídos para formação da BoW considerando o pré-processamento do texto: gost, sab, arquitet, detalh, alic, analis, licit, edit, mod, entend, component, form, integr, ped, codig, font, usad, constru, alic, instruc, replic.

Representações gBoED:

gBoED_Freq
gostaria_1_arquitetura_detalhada_do_alice
1

gBoED_Dist
gostaria_1_arquitetura_detalhada_do_alice
0,33

Sendo que o valor 1 existente no meio da expressão, indica que ela tende a pertencer à classe "negado". A classe "concedido" é representada pelo valor 0 (zero). Em gBoED_Dist a métrica 0,33 indica o inverso da distância em palavras entre "gostaria" e "arquitetura".

Fonte: Xavier (2021).

Tabela 91 – Representatividade da gBoED no conjunto de dados.

	gBoED_Freq				***gBoED_Dist***			
	Listas 100%		**Listas 10%**		**Listas 100%**		**Listas 10%**	
Qtde de documentos representados	467	91.21%	464	90.62%	467	91.21%	464	90.62%
Qtde de documentos sem representação	45	8.78%	48	9.37%	45	8.78%	48	9.37%
Qtde de ACERTOS na predição	310	60.54%	303	59.17%	342	66.79%	335	65.42%
Qtde de ERROS na predição	116	22.65%	127	24.80%	84	16.40%	95	18.55%
Qtde de NEUTROS na predição	86	16.79%	82	16.01%	86	16.79%	82	16.01%

Fonte: Elaborada pelo autor.

para esse conjunto de dados as diferenças de representação são bastante próximas. Acredita-se que a formação das listas por parte de um especialista do domínio contribui significativamente para uma melhor representatividades dos documentos.

Com relação a quantidade de acertos na predição ao consultar a representação, tem-se gBoED_Dist maior do que gBoED_Freq, totalizando 342 documentos (66,79%) e 310 documentos (60,54%), respectivamente. A taxa de erros na predição de ambas as representações atinge um total de 84 documentos (16,40%) e 116 documentos (22,65%) respectivamente. 86 documentos (16,79%) não obtiveram resultados na predição em ambas as representações.

A Tabela 92 apresenta as melhores acurácias obtidas pelos modelos gerados a partir da BoW, aqueles incrementados com os resultados da representação gBoED_Freq e pela gBoED_-Dist. Como tais modelos atendem aos requisitos de transparência e de explicabilidade do processo, podem ser utilizados para desenvolvimento do primeiro modelo de classificação aplicado ao domínio dos pedidos de acesso à informação, de competência da CGU. Nessa tabela também são apresentados os valores de comparação utilizando o experimento do tipo Oráculo (definido na Subseção 3.4.2), definindo o limite superior do método para este estudo, e o limite inferior definido pelo experimento que utiliza listas de termos mapeadas em apenas 10% da base. Nos resultados dos experimentos que considera o Oráculo são consideradas as acurácias para o mesmo nível de confiança das Listas 100% e, também, a acurácia obtida com grau de confiança de 75%. Nos resultados das duas gBoEDs, são apresentadas as acurácias referentes a cada versão e o grau de confiança das respostas utilizadas naqueles resultados.

Tabela 92 – Melhores acurácias dos classificadores.

Algoritmos	**BoW**	**gBoED_Freq**						**gBoED_Dist**					
		Oráculo		**Listas 100%**		**Listas 10%**		**Oráculo**		**Listas 100%**		**Listas 10%**	
	Acc	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**	**Acc**	**Conf**
C4.5-Entropia	0,66406	0,66595 0,73819	55% 75%	0,66595 27-19	55%	0,65618 10-6	65%	0,82839 0,73820	90% 75%	0,67198 274 - 131	90%	0,66214 52 - 36	80%
C4.5-Gini	0,68563	0,71497 0,70124	85% 75%	0,71308 58 - 34	85%	0,68379 13-8	70%	0,76007 0,70124	90% 75%	0,72474 123-56	90%	0,69544 123-56	90%
KNN-Cosseno	0,76757	0,85735 0,97266	55% 75%	0,77157 103-62 $n = 11$	55%	0,76376 103-62 $n = 13$	55%	0,84962 0,95897	55% 75%	0,78710 88-42 $n = 13$	55%	0,77534 88-42 $n = 15$	55%
KNN-Euclideana	0,76757	0,85735 0,97266	55% 75%	0,77157 103-62 $n = 11$	55%	0,76376 103-62 $n = 13$	55%	0,84962 0,95897	55% 75%	0,78710 88-42 $n = 13$	55%	0,77534 88-42 $n = 15$	55%
MNB	0,75419	0,80094 0,91610	55% 75%	0,76199 46-29 $\alpha = 10^{-1}$	55%	0,75807 46-29 $\alpha = 10^{-1}$	55%	0,80094 0,91610	55% 75%	0,76784 46-29 $\alpha = 10^{-1}$	55%	0,77176 59-33 $\alpha = 10^{-1}$	60%
SVM-Linear	0,76188	0,80094 0,91207	55% 75%	0,76384 36-25 $\gamma = 10^{-4}$ a 10^{4}	55%	0,77753 46-29 $\gamma = 10^{-4}$ a 10^{4}	55%	0,85750 0,91207	65% 75%	0,77353 127-68 $\gamma = 10^{-4}$ a 10^{4}	65%	0,77753 46-29 $\gamma = 10^{-4}$ a 10^{4}	55%
SVM-Polinomial	0,76995	0,80879 0,92768	55% 75%	0,77571 37-18 $\gamma = 10$	55%	0,77960 37-18 $\gamma = 10$	55%	0,84001 0,92768	60% 75%	0,79121 81-37 $\gamma = 10$	60%	0,78736 37-17 $\gamma = 10$	55%
SVM-RBF	0,76395	0,81075 0,92187	55% 75%	0,77756 31-15 $\gamma = 1$	55%	0,78345 31-15 $\gamma = 1$	55%	0,81075 0,92187	55% 75%	0,77945 31-11 $\gamma = 1$	55%	0,78345 31-15 $\gamma = 1$	55%

Fonte: Elaborada pelo autor.

Observando a Tabela 92, é possível verificar que a melhor acurácia, destacada pela linha cinza foi utilizando o algoritmo ***Support Vector Machine*, kernel Polinomial, $\gamma = 10$, enriquecido pela representação gBoED_Dist**. Nesse experimento, obteve-se **acurácia de 79,121% e Medida-F1 de 79,589%**, grau de confiança de 60%, com 81 documentos sendo consultados e 37 documentos reclassificados. É possível verificar também que os resultados enriquecidos tanto pela representação gBoED_Freq quanto gBoED_Dist, em Listas 100% e Listas 10%, obtiveram resultados superiores à BoW. O limite superior (Oráculo) definido pelo modelo da gBoED_Dist é de 84,001% em um grau de confiança de 60% e 92,768% em um grau de confiança de 75%. O limite inferior (Listas 10%) obteve acurácia de 78,73% com grau de confiança de 55%, 37 docu-

mentos enviados para reclassificação e 17 documentos reclassificados. Interessante notar que o limite inferior para o modelo gBoED_Dist obteve um resultado superior ao modelo gBoED_Freq com Listas 100%. Apesar de não serem os melhores resultados obtidos, é importante apresentar alguns destaques como o resultado do modelo gerado pelo algoritmo MNB, que obteve melhor acurácia em gBoED_Dist Listas 10% do que em Listas 100%. O mesmo ocorreu com SVM-rbf, tanto em gBoED_Freq quanto gBoED_Dist.

Em gBoED_Freq usando Listas 100%, a acurácia aumentou para 77,571% com Medida-F1 para 78,110%, com grau de confiaça de respostas de 55%, 37 documentos enviados para reclassificação e 18 reclassificados. No mesmo grau de confiança, no cenário Oráculo seria um total de 81,07%, atingindo valor máximo de 92,18% com confiança de 75%. No cenário Listas 10%, o resultado apresentou uma acurácia melhor do que em Listas 100%, 77,960% com 55% de confiança.

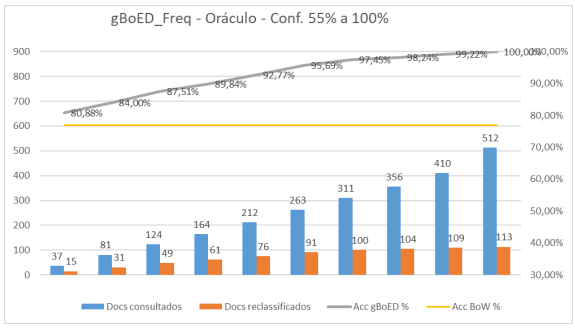
Outro destaque foram os resultados do modelo gerado pelo algoritmo C4.5-Gini, que oferece maior nível de explicabilidade e transparência. Nele a acurácia foi de 68,6% para 71,3%, usando gBoED_Freq, e chegou a 72,5%, usando o método de enriquecimento com representação gBoED_Dist, com Medida-F1 de 72,7%. Eles apresentaram valores de acurácia muito próximos àqueles atingidos pelo oráculo. Nesses experimentos, alguns dos melhores valores de acurácia aconteceram com reclassificação em níveis altos de confiança, em torno de 80 a 90%. Tais resultados possuem aplicabilidade em relação aos dados reais, o que traz grande contribuição à sociedade e o cumprimento dos direitos fundamentais relacionados ao acesso à informação.

A Figura 66 e a Figura 67 apresentam gráficos comparativos entre quantidade de documentos que necessitaram consulta à gBoED_Freq e gBoED_Dist, respectivamente, a quantidade de documentos que sofreram alteração em relação à classe predita na etapa 1 do método de classificação em cada um dos níveis de confiança testados e as acurácias obtidas em cada cenário de aplicação do modelo SVM, kernel Polinomial e $\gamma = 10$.

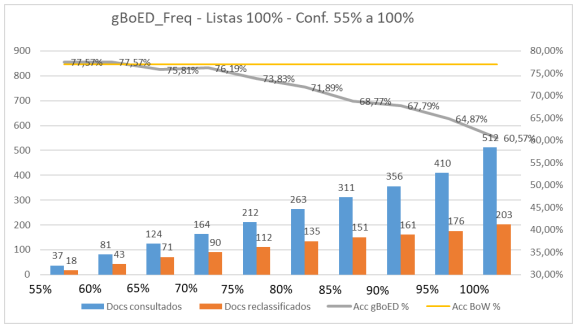
Nos gráficos da Figura 66 verifica-se o custo esperado de uso do método de classificação Semântica em Pedidos de Acesso à Informação quando utilizada a representação gBoED_Freq como informação enriquecida. Na Figura 66a observa-se o custo esperado entre a quantidade de documentos enviados para consulta e reclassificados. Nesta relação é possível verificar que quanto maior o grau de confiança utilizado como linha de corte, proporcionalmente maior é a quantidade de documentos consultados. Porém a quantidade de documentos reclassificados não segue a mesma proporção, bem como nos experimentos representados pelas Figuras 66b e 66c. Ainda na Figura 66a, observa-se que a acurácia esperada, quando aplicado cada grau de confiança, aumenta até chegar em 100%. Porém, o que se percebe na realidade dos experimentos das Figuras 66b e 66c é uma taxa decrescente na acurácia do método.

Nos gráficos da Figura 67 verificam-se características semelhantes à da Figura 66. Para o modelo que obteve o melhor valor de acurácia foi o SVM, kernel Polinomial e $\gamma = 10$, os valores ideais (Oráculo) apresentam uma proporcionalidade entre a quantidade de documentos

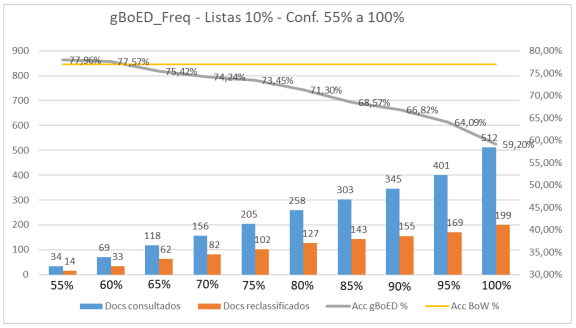
Figura 66 – Uso da gBoED_Freq nos cenários de SVM-Polinomial $\gamma = 10$.



(a) gBoED_Freq Oráculo



(b) gBoED_Freq 100%



(c) gBoED_Freq 10%

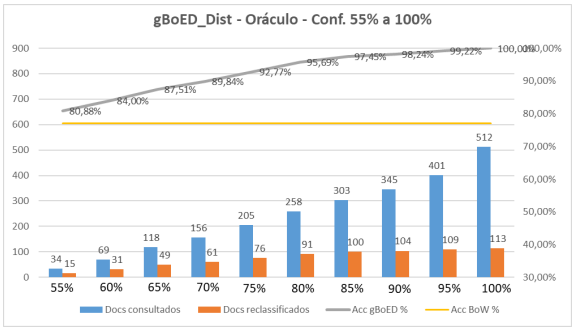
Fonte: Elaborada pelo autor.

consultados e a acurácia, conforme o valor de confiança aumenta. A quantidade de documentos reclassificados cresce em uma taxa menor. Para os modelos de Listas 100% e Listas 10%, conforme a confiança aumenta a acurácia diminui. O melhor valor de acurácia em Listas 100% ocorre com confiança igual a 60%.

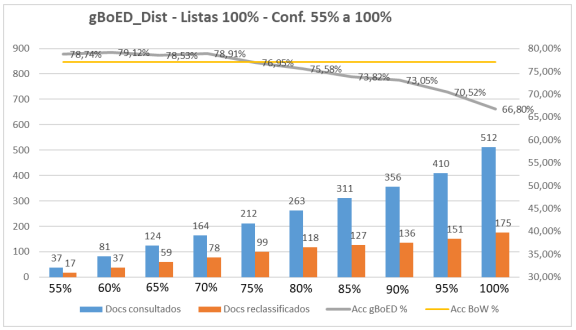
O conjunto de resultados apresentados mostra que os modelos gerados possuem representatividade e aplicabilidade em relação aos dados reais, o que traz grande contribuição à sociedade e o cumprimento dos direitos fundamentais relacionados ao acesso à informação. Este estudo de caso resultou no trabalho de conclusão do curso MBA em Ciências de Dados na Universidade de São Paulo, no ano de 2021 (XAVIER, 2021) e na publicação do artigo *"Right to information in the era of artificial intelligence in Brazil"*, realizado no *26th World Congress of Political Science (New Nationalisms in an Open World)* (XAVIER; REZENDE; SCHEICHER, 2021).

A equipe de gestão da Ouvidoria-Geral da União, órgão responsável pela tomada de decisão da CGU, em última instância recursal, está em processo de adequação dos modelos gerados aos sistemas da CGU para utilização no apoio à classificação dos pedidos de acesso à informação. Existe grande interesse em aplicar o método e adequá-lo à base dos recursos de primeira, segunda e terceira instâncias no mesmo órgão (desde que ajustados os parâmetros e o pré-processamento da base dos precedentes dos recursos interpostos).

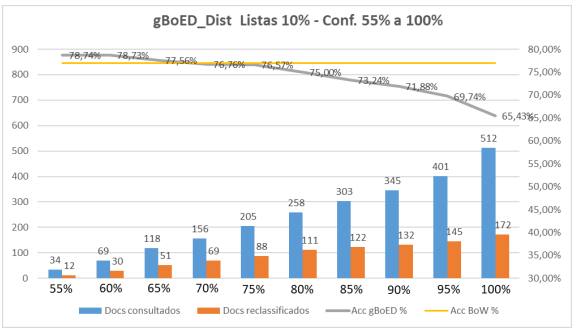
Figura 67 – Uso da gBoED_Dist nos cenários de SVM-Polinomial $\gamma = 10$.



(a) gBoED_Dist Oráculo



(b) gBoED_Dist 100%



(c) gBoED_Dist 10%

Fonte: Elaborada pelo autor.

6.4 Considerações finais

Em busca de uma solução para a crescente demanda de registros de pedidos de acesso à informação junto à Controladoria-Geral da União, este estudo de caso apresentou um conjunto de resultados que visam apoiar o processo de classificação dos pedidos de acesso à informação no âmbito da CGU e validar o Método de Classificação Semânticamente Enriquecido por Expressões do Domínio em uma caso real, de modo que ele se mostra um processo adequado para melhoria de resultados de classificação de nível semântico.

Na análise comparativa dos resultados de classificação dos pedidos de acesso à informação pelos algoritmos que oferecem maior transparência e explicabilidade para o processo de apoio à tomada de decisão na CGU, o algoritmo *Support Vector Machine* – SVM, kernel Polinomial e $\gamma = 10$, alcançou a melhor métrica de validação com acurácia de 79,1% e Medida-F1 de 79,5%, a partir do enriquecimento semântico oferecido pela representação de textos gBoED_Dist, inclusive com observância ao equilíbrio de classificação de cada classe.

O especialista avalia os resultados como tendo grande potencial quanto ao uso deste modelo para classificação dos pedidos e dos recursos de acesso à informação na CGU, bem como em toda a administração pública em todos os Poderes, Executivo, Legislativo e Judiciário, sendo necessário adaptá-lo às bases de dados e ao pré-processamento com a representação das expressões de novos domínios. O especialista ainda adiciona que pretende dar continuidade nesta

pesquisa em parceria com a equipe do Laboratório de Inteligência Computacional (LABIC) da USP, bem como em parceria com a Diretoria de Tecnologia da Informação da CGU possibilitando o desenvolvimento de melhorias no modelo de classificação a partir da indicação de precedentes mais próximos a novos pedidos e recursos, por algoritmos de associação. Será também possível desenvolver a incorporação de informações complementares, existentes em outras bases de dados do órgão, como por exemplo, o enriquecimento de informações contidas em sistemas de gestão de documentos e processos eletrônicos, em que se pode consultar se um processo é sigiloso ou se o processo apresenta alguma restrição de acesso. Será também possível integrar as informações da plataforma FalaBr. Essas informações poderão contribuir ainda mais para a melhoria do modelo, impulsionando uma implementação mais eficiente do direito de acesso à informação no Brasil.

CAPÍTULO

7

CONCLUSÕES

7.1 Contribuições científicas

As contribuições deste trabalho estão diretamente relacionadas às questões de pesquisa e aos objetivos específicos apresentados na introdução desta tese. As **questões de pesquisa** que direcionam este trabalho são apresentadas a seguir:

Q1 A utilização e a combinação de representações semanticamente enriquecidas com a *Bag of Words* pode levar a melhores resultados de classificação?

Q2 Qual o impacto da utilização de expressões do domínio geradas a partir de listas construídas de forma semiautomática, nas tarefas de classificação para resolverem problemas de diferentes níveis de complexidade semântica?

O **objetivo geral** deste trabalho foi desenvolver diferentes abordagens de representações semanticamente enriquecidas e um método de classificação que utilize as representações enriquecidades e que proporcione melhores resultados em cenários de classificação de nível semântico. Esse objetivo geral foi organizado nos seguintes objetivos específicos:

O1 Propor e desenvolver diferentes soluções de representações semanticamente enriquecidas por expressões do domínio que possam ser aplicadas em diferentes domínios e idiomas.

O2 Propor, desenvolver e analisar o impacto de um método de classificação que combine a representação BoW às representações semanticamente enriquecidas por expressões do domínio em problemas de diferentes níveis de complexidade semântica.

O3 Aplicar e analisar o impacto de soluções de extração de termos para a construção de listas termos do domínio e identificadores de classe, usando técnicas baseadas em análise

sintática e BERT, a fim de tornar o processo de construção das diferentes versões de representações enriquecidas por expressões do domínio mais automatizado.

As principais **contribuições científicas** são sintetizadas nessa seção (indicadas pelos tópicos $C_x, x = 1, \ldots, 6$), apresentando as técnicas propostas e desenvolvidas, além dos resultados obtidos e relacionando-os às questões e objetivos específicos.

C1 - Processo de generalização da representação semanticamente enriquecida Bag of Expressions of Domain (BoED)

Os trabalhos de Marques *et al.* (2015) e Sinoara (2018) apresentam estudos e aplicações que envolvem o conceito relacionado com expressões do domínio, a representação semanticamente enriquecida *Bag of Expressions of Domain (BoED)* e como elas contribuem para melhorar a organização de textos em tarefas de classificação de nível semântico, em domínios específicos. Com o objetivo de melhorar a representação semanticamente enriquecida por expressões do domínio - BoED - e permitir que ela possa ser construída para qualquer domínio do conhecimento e aplicada sobre tarefas de classificação de nível semântico, no Capítulo 3 é desenvolvido um processo de generalização da construção da representação *Bag of Expressions of Domain (BoED)* usando métrica de frequência de expressões. A representação generalizada recebe o nome de *generalized Bag of Expressions of Domain (gBoED)*. Elas construídas pela combinação de termos relacionados ao domíniodo problema e termos relacionados ao tipo de classificação que se deseja realizar. As listas são construídas de forma manual pelos especialistas do domínio, com o objetivo de agregar o conhecimento de domínio em termos e expressões. O conhecimento do domínio é a fonte de enriquecimento de informações que agrega o maior valor ao processo, tornando as representações mais completas e explicáveis.

Experimentos iniciais foram executados com o objetivo de obter uma percepção sobre a viabilidade do uso da representação gBoED em tarefas de classificação de diferentes níveis. Os primeiros resultados mostraram que a combinação da BoW com as expressões do domínio atigem resultados promissores e fornecem evidências de que novas versões da representação semanticamente enriquecida e melhoria no método de classificação poderiam ser realizadas. Com isso, foi proposta uma nova versão da gBoED com métrica relacionada à distância entre termos e um novo método de classificação enriquecida por expressões do domínio foi proposto. Esse resultado está relacionado à questão de pesquisa **Q1** e ao objetivo específico **O1**.

C2 - Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

Com base nos primeiros experimentos executados nos experimentos de viabilidade da representação semanticamente enriquecida por expressões do domínio obtida a partir do processo de generalição do item C1, um novo Método de Classificação Semanticamente Enriquecida por

Expressões do Domínio foi proposto e desenvolvido com o objetivo de melhorar resultados de tarefas de classificação de nível semântico. O método composto por 2 etapas principais realiza o treinamento de modelos por meio da representação tradicional BoW e predição de resultados de novos documentos, seguido pela reclassificação a partir das representações baseadas em expressões do domínio em predições cuja confiança está abaixo de um limiar pré-definido.

Experimentos foram realizados utilizando 10 coleções de documentos distintas em idiomas português e inglês, com variação de domínios, cenários, listas de termos, algoritmos de classificação e parâmetros. De modo geral os resultados foram bastante promissores, com a obtenção de melhorias significativas nos diferentes cenários, assim como a melhoria na explicabilidade dos resultados devido às característica das representações gBoED. A proposta, desenvolvimento, avaliação experimental e análise do impacto do uso do novo método de classificação semanticamente enriquecida, está relacionada à questão de pesquisa **Q1** e ao objetivo específico **O2**.

C3 - Método de extração de termos e representação enriquecida baseados em regras morfossintáticas

A formação das representações semanticamente enriquecidas (gBoEDs) baseiam-se na construção de listas de termos de domínio e identificadores de classe. Na forma mais tradicional de aplicação do método de classificação do item C2, as listas de termos são construídas manualmente pelos especialistas do domínio. Este é um processo que requer um grande esforço por parte dos especialistas.

No Capítulo 4 foi proposto e desenvolvido um método de extração de termos baseado em regras morfossintáticas com o objetivo de tornar o processo de construção das listas mais automatizado e diminuir o esforço para utilização das informações enriquecidas e do método de classificação enriquecida. O método proposto utiliza anotações do tipo *Part-of-Speech (POS)*, do inglês (*POS tags)*, para a construção de regras baseadas nas estruturas morfossintáticas dos documentos nos idiomas português e inglês.

Com base nas mesmas regras morfossintáticas desenvolvidas para o processo de extração de termos, foi proposto um método para construção de uma nova versão da representação gBoED, denominada gBoED_Syntax. Ela utiliza as regras morfossintáticas para combinação dos termos do domínio e identificadores de classe visando um representação mais próxima da estrutura gramatical dos textos dos documentos.

Experimentos foram realizados com o objetivo de avaliar o impacto das novas listas de termos extraídas a partir do método de extração baseado em regras morfossintáticas e, também, avaliar o impacto da nova representação gBoED_Syntax. O experimento foi realizado a partir do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. Como o método de extração baseado em regras morfossintáticas é um processo dependente das

estruturas morfossintáticas de cada domínio, os experimentos foram realizados em 9 coleções de documentos específicas para o domínio de Análise de sentimentos. Neles, foram realizadas variações de idioma (português e inglês), cenários, listas de termos, algoritmos de classificação e parâmetros dos algoritmos. Os resultados mostram apresentam maior nível de explicabilidade da representação gBoED_Syntax em relação às outras versões. Na classificação semanticamente enriquecida os resultados atingem valores intermediários em relação às diferentes versões de representações formadas por listas manuais, algumas vezes até superando os resultados obtidos por listas manuais. Esses resultados são bastante significativos pois mostram que é possível automatizar partes do processo de construação das listas de termos, atingindo resultados satisfatórios em relação a acurácias dos modelos.

A proposta, desenvolvimento e avaliação do método de extração de termos baseado em regras morfossintáticas contempla a questão de pesquisa **Q2** e ao objetivo específico **O3**. A proposta, desenvolvimento e avaliação da representação semanticamente enriquecida com base em regras morfossintáticas contempla a questão de pesquisa **Q1** e ao objetivo específico **O1**.

C4 - Método de extração de termos baseado em modelo de linguagem BERT

De forma semelhante ao item C3, foi realizada a proposta, desenvolvimento e avaliação de um método de extração de termos baseado em modelo de linguagem BERT. Os modelos de linguagem BERT são recursos baseados no conceito de redes neurais, portanto fazem parte das tarefas de aprendizado supervisionado. O método proposto faz parte de uma adaptação do trabalho de Hoang, Bihorac e Rouces (2019) e consiste em 4 etapas principais.

Na etapa 1, os especialistas de domínio preparam listas com termos extraídos manualmente, que servem de treinamento para construção de um modelo de extração baseado em BERT. De modo a tornar o processo mais realista, nos experimentos realizados, as listas iniciais são elaboradas a partir de 10% dos documentos das coleções anotados. Na etapa 2 é realizado o treinamento do modelo. Os modelos são treinados de acordo com o domínio e tipo de lista de termos que se deseja construir. Na etapa 3, o modelo gerado realiza a extração dos termos a partir de novos conjuntos de documentos. Nos experimentos realizados o modelo é aplicado em 100% dos documentos das coleções. A partir dos termos extraídos, na etapa 4, os especialistas do domínio realizam a limpeza e organização das listas.

Experimentos foram realizados com o objetivo de avaliar o impacto das novas listas de termos extraídas a partir do método de extração baseado em modelo de linguagem BERT. O experimento foi realizado a partir do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, comparando-se os resultados dos modelos gerados pelas diferentes representações gBoED e pelas listas manuais, geradas pelo método de análise morfossintáticas e BERT. Foram utilizadas 10 coleções de documentos distintas em idiomas português e inglês, com variação de domínio, cenários, listas de termos, algoritmos de classificação e parâmetros. De modo geral os resultados obtidos usando as listas extraídas pelo método BERT obtiveram

resultados inferiores às listas manuais e extraídas por análise sintática. Porém, ainda apresentam-se bastante promissores, pois a acurácia final dos modelos ainda atingiu resultados intermediários nos diferentes cenários. A proposta, desenvolvimento e avaliação do método de extração de termos baseado em modelo de linguagem BERT contempla a questão de pesquisa **Q2** e ao objetivo específico **O3**.

C5 - Criação de Representações Semanticamente Enriquecidas por Expressões do Domínio

Ao longo do desenvolvimento das soluções geradas nesta tese de doutorado, foram criadas de diferentes abordagens da Representação Semanticamente Enriquecidas por Expressões do Domínio. A primeira delas criada a partir do processo de generalização da representação BoED, utiliza métrica de frequência de expressões do domínio em cada documento, foi denominada gBOED_Freq. A segunda representação baseada-se na métrica que considera o inverso da distância entre um termo do domínio e um identificador de classe em cada sentença de um documento. Esta representação recebe o nome de gBoED_Dist. Por último, a partir do desenvolvimento do método de extração de termos baseado em regras morfossintáticas, foi proposto um método para construção de uma nova versão da representação gBoED, denominada gBoED_Syntax. O desenvolvimento de diferentes versões da representação gBoED contribui o aumento de possibilidades de uso dentro do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio, bem como para evidenciar diferentes características que enriquecem as representações e os modelos. Esse resultado atende ao objetivo específico **O1**, que serve de base para a questão de pesquisa **Q1**.

C6 - Análise de impacto do uso das listas extraídas pelos métodos semiautomáticos no Método de Classificação Semanticamente Enriquecida por Expressões do Domínio

O processo de automatização da extração de termos e construação das listas de termos do domínio e identificadores de classe contribui para uma diminuição do esforço do especialista do domínio e, consequentemente, em diferentes versões das representações semanticamente enriquecidas. Portanto, outra importante contribuição desta tese de doutorado está relacionada à aplicação e análise do impacto da utilização das representações semanticamente enriquecidades por expressões do domínio construídas por listas de termos extraídas pelos métodos semiautomáticos, baseados em regras morfossintáticas e que utilizam modelo de linguagem BERT. Esse resultado atende às questões de pesquisa **Q1** e **Q2**, além de atender ao objetivo específico **O2** e **O3**.

7.2 Publicações

Durante o desenvolvimento deste trabalho, as contribuições obtidas foram divulgadas por meio de publicação de artigos em periódicos e publicação e apresentação de artigos em conferências. Essas publicações são listadas a seguir, apresentando a relação de cada uma com este trabalho e indicando aquelas que estão diretamente relacionadas às questões de pesquisas estabelecidas.

Artigos publicados em periódicos e anais de conferências

SCHEICHER, R.; SINOARA, R. A.; KOGA, N.; REZENDE, S. O. *Uso de expressões do domínio na classificação automática de documentos*. In: ENIAC 2016 - Anais do XIII Encontro Nacional de Inteligência Artificial e Computacional, Recife, Brasil, p. 625-636, 2016.

Nesse artigo foi publicada a proposta da representação *generalized Bag of Expressions of Domain (gBoED)*, desenvolvida neste trabalho. A implementação computacional do método e a execução da avaliação experimental foi realizada em parceria com os demais autores. A autor desta tese foi responsável pela elaboração da proposta e da configuração experimental e participou ativamente da análise dos resultados e escrita do artigo. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1**, objetivo específico **O1**.

SINOARA, R. A.; **SCHEICHER, R. B.**; REZENDE, S. O. *Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity*. In: CIDM'17-Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence, Honolulu, USA, p. 2057–2064, 2017.

Nesse artigo foi publicado o trabalho referente à definição de níveis de complexidade semântica e apresenta uma prova de conceito sobre a aplicação do método LDA em cenários reais de organização de coleções de documentos. A autor desta tese foi responsável pela elaboração e configuração experimental e participou ativamente da análise dos resultados e escrita do artigo.

SCHEICHER, R. B.; SINOARA, R. A.; FELINTO, J. C.; REZENDE, S. O. *Sentiment Classification Improvement Using Semantically Enriched Information*. In: DocEng '19 - Proceedings of the ACM Symposium on Document Engineering 2019, Berlin, Germany, 2019.

Neste artigo foi publicada a proposta da representação *generalized Bag of Expressions of Domain com métrica de Distância entre termos (gBoED_Dist)*. Neste artigo também foi publicada a proposta e avaliação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. O autor desta tese foi responsável pela elaboração da proposta,

implementação e da configuração experimental e participou ativamente da análise dos resultados e escrita do artigo. Os demais autores contribuíram com a implementação e análise dos resultados. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1**, objetivos específicos **O1** e **O2**.

SCHEICHER, R. B.; SINOARA, R. A.; REZENDE, S. O. *Exploração do uso de expressões do domínio na classificação de sentimentos*. In: ERAMIA '20 - Anais da I Escola Regional de Aprendizado de Máquina e Inteligência Artificial de São Paulo, p. 38-42, São Paulo, São Paulo, 2020.

Neste artigo foi publicada uma extensão do artigo anterior referente à representação *generalized Bag of Expressions of Domain com métrica de Distância entre termos (gBoED_Dist)* e ao Método de Classificação Semanticamente Enriquecida por Expressões do Domínio. A extensão foi realizada com a aplição das representações e do método para outras coleções de documentos, ampliando a análise de resultados. O autor desta tese foi responsável pela elaboração da proposta, implementação e da configuração experimental e participou ativamente da análise dos resultados e escrita do artigo. Os demais autores contribuíram com a implementação e análise dos resultados. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1**, objetivos específicos **O1** e **O2**.

XAVIER, F. L. S.; **SCHEICHER, R. B.**; REZENDE, S. O. *Right to information in the era of artificial intelligence in Brazil*. In: IPSA '21 - 26th World Congress of Political Science (New Nationalisms in an Open World), Lisboa, Portugal, 2021.

Neste artigo foi publicado o estudo de caso realizado a partir da exploração de pedidos de acesso à informação oriundos da Controladoria Geral da União (CGU) e da aplicação do Método de Classificação Semanticamente Enriquecida por Expressões do Domínio para um caso real e em língua portuguesa. O autor desta tese foi responsável por parte da elaboração da proposta, implementação e da configuração experimental e participou ativamente da análise dos resultados e escrita do artigo. Os demais autores contribuíram com a implementação e análise dos resultados. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1**, objetivos específicos **O1** e **O2**.

Artigos submetidos

Além dos trabalhos já publicados apresentados, o seguinte trabalho foi submetido, visando a publicação no curto prazo.

XAVIER, F. L. S.; **SCHEICHER, R. B.**; SINOARA, R. A. *Classificação Semântica de Pedidos de Acesso à Informação*. O artigo foi submetido para revisão e publicação no periódico *Revista da CGU*. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1**, objetivos específicos **O1** e **O2**.

SCHEICHER, R. B.; SINOARA, R. A.; REZENDE, S. O. *Rule-based terms extraction and enriched sentiment classification using domain expressions for Portuguese and English*. O artigo foi submetido para revisão e publicação no periódico *Knowledge-based Systems*. O trabalho apresentado nesse artigo está relacionado à questão de pesquisa **Q1** e**Q2**, e objetivos específicos **O1**, **O2** e **O3**.

Artigos em preparação

Todas as contribuições obtidas durante o desenvolvimento deste trabalho serão divulgadas por meio de publicação de artigos em periódicos e em anais de conferências. Os títulos a seguir correspondem aos artigos que se encontram em fase de preparação, visando sua publicação a médio prazo.

Extração de termos usando BERT em tarefas de classificação semântica em diferentes domínios.

O foco deste artigo é a publicação do método de extração de termos baseado em modelo de linguagem BERT, em diferentes domínios, em idioma português e inglês, além dos resultados obtidos. O objetivo é a divulgação das técnicas e do conhecimento sobre expressões do domínio e enriquecimento semântico de representações. A abordagem de extração de termos permite o avança do estado-da-arte neste assunto.

Ferramenta de apoio à extração de termos, construção de representações semanticamente enriquecidas e classificação semântica.

O foco deste artigo é a publicação de uma ferramenta para apoio à extração de termos, construção de representações semanticamente enriquecidas e classificação semântica com base nas técnicas e conceitos que envolvem os experimentos apresentados nesta tese de doutorado.

Ela visa apoiar a construção de listas de termos do domínio e identificadores de classe, em tarefas de classificação semântica e que usando o método de enriquecimento baseado em expressões do domínio. A ferramenta dá suporte à extração de termos, visualização, filtragem e categorização dos termos de acordo com as classes a serem aplicadas na tarefa de classificação.

7.3 Limitações e trabalhos futuros

Nessa seção são discutidas algumas características e limitações das abordagens propostas e das avaliações experimentais realizadas. Também são apresentadas potenciais direções para trabalhos futuros. Esta tese de doutorado possui dois focos principais de atuação voltados para a extração de termos e a construção de representações semanticamente enriquecidas, além da melhoria de resultados de classificação de nível semântico usando informações de domínio.

Com relação à extração de termos e construção de representações semanticamente enriquecidas, dois métodos de extração de termos foram construídos com o objetivo de automatizar o processo de construção das listas de termos. A identificação de possíveis técnicas que reduzam o esforço dos especialistas na construção das listas é de extrema importância para obter representações mais próximas do texto real e, consequentemente, que permitam obter resultados de classificação ainda melhores em cada domínio, além de contribuir com a explicabilidade dos resultados. Em ambos, o especialista do domínio é requisitado para limpeza e organização dos termos extraídos. Essa fase é bastante importante para que as listas possam ser bem consolidadas de acordo com o conhecimento de domínio do especialista. Nesse sentido, em trabalhos futuros poderiam ser realizadas a aplicação de técnicas para ranqueamento e filtragem de termos, por exemplo métricas de *Frequêcia*, *C-Value*, *T-Score*, como em Dai e Song (2019), Stanković *et al.* (2016), Marciniak e Mykowiecka (2014), de modo a facilitar o trabalho dos especialistas. Abordagens híbridas que utilizem *Skip-gram* (SABRA; SABEEH, 2020), por exemplo, combinadas com outras técnicas de extração (CONRADO *et al.*, 2014), seleção de termos, filtragem e ranqueamento são consideradas para diminuir o esforço e a dependência dos especialistas.

Outro recurso importante que visa melhorar o processo de extração de termos, construção de representações e classificação de nível semântico, é a construção de uma ferramenta capaz de encapsular as diversas técnicas abordadas nesta tese de doutorado e permitir que o especialista possa escolher e experimentar diferentes abordagens para o domínio que está trabalhando.

Com relação a identificação de termos, os métodos não são capazes de identificar automaticamente os sinônimos contidos no conjunto de textos. Considerando que a abordagem empregada é livre de domínio, essa tarefa não é trivial. No entanto, o emprego de recursos como bancos de dados lexicais de sinônimos podem auxiliar nessa tarefa. Nesse sentido, esforços futuros serão direcionados para melhorar a identificação automática de sinônimos em língua inglesa e portuguesa, usando recursos como o banco de dados WordNet (LOMBARDI; MARANI, 2015; MILLER, 1998) para o inglês e WordNet.Br (SILVA, 2005) para o português.

Com relação a construção das representações semanticamente enriquecidas, uma limitação e possível trabalho futuro está relacionao à falta de identificação de recursos linguísticos como relacionamentos semânticos do tipo “causalidade” (causa e efeito) e “holonímia e meronímia” (relação parte-todo) (VASQUES; REZENDE; MARTINS, 2018), por exemplo, além de outras estruturas como papéis semânticos (OLIVEIRA, 2020; HARTMANN; DURAN; ALUÍSIO, 2016). Tais recursos linguísticos podem contribuir com o eriquecimento das representações e a melhoria de classificações de nível semântico.

O Método de Classificação Semanticamente Enriquecida por Expressões do Domínio é o outro foco abordado ao longo desta tese. Sua principal limitação está exatamente na dependência de resultados entre as duas etapas. Portanto, trabalhos futuros visam a melhoria com a aplicação de uma abordagem que permita combinar e balancear um sistema tradicional com as informações enriquecidas.

REFERÊNCIAS

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2014. ISBN 1466586745, 9781466586741. Citado na página 51.

AGGARWAL, C. C.; ZHAI, C. (Ed.). **Mining Text Data**. [S.l.]: Springer, 2012. Citado nas páginas 29, 30, 45, 47 e 48.

AGIRRE, E. **Word Sense Disambiguation: Algorithms and applications**. [S.l.]: Springer Netherlands, 2007. v. 33. Citado na página 57.

ALBITAR, S.; ESPINASSE, B.; FOURNIER, S. Semantic enrichments in text supervised classification: application to medical domain. In: **The Twenty-Seventh International Flairs Conference**. [S.l.: s.n.], 2014. Citado na página 80.

ALMEIDA, G. M. B.; L., K. D. S.; MANFRIM, A. M. P.; SOUZA, I. P.; IZUMIDA, F. H.; FELIPPO, A. D.; ZAUBERAS, R. T.; MELCHIADES, F. G.; BOSCHI, A. O. Glossário de revestimento cerâmico. **Cadernos de Terminologia**, n. 04, p. 03–56, 2011. Citado na página 60.

ALMEIDA, G. M. B.; VALE, O. A. Do texto ao termo: interação entre terminologia, morfologia e linguística de corpus na extração semi-automática de termos. **As Ciências do Léxico: Lexicologia, Lexicografia e Terminologia**, Finatto MJB, Isquerdo AN, Isquerdo AN, Finatto MJB. Campo Grande, MS, Brazil . . . , v. 4, p. 483–499, 2008. Citado nas páginas 59, 61 e 64.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: MAMEDE, N. J.; TRANCOSO, I.; BAPTISTA, J.; NUNES, M. das G. V. (Ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 110–117. Citado nas páginas 257 e 258.

AMARAL, D.; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014. Citado na página 57.

BARROS, L. A. **Curso básico de terminologia**. [S.l.]: Edusp, 2004. v. 54. Citado nas páginas 59 e 62.

BATISTA, R. P. Características de terminologia empresarial: um estudo de caso. Universidade do Vale do Rio dos Sinos, 2011. Citado na página 59.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. **Advances in neural information processing systems**, v. 13, 2000. Citado na página 66.

BRASIL. **Lei Geral de Proteçãode Dados (LGPD)**. 2018. Citado na página 31.

____. **Lei de Acesso à Informação (LAI)**. 2020. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei n. 8.112, de 11 de dezembro de 1990; revoga a Lei n. 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. Citado nas páginas 216 e 225.

BREVE, F. A. **Aprendizado de máquina em redes complexas**. Tese (Tese de Doutorado) — Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, 2010. Citado na página 45.

CASTELLVÍ, M. T. C. **Terminology: Theory, methods and applications**. [S.l.]: John Benjamins Publishing, 1999. v. 1. Citado na página 59.

CASTELLVÍ, M. T. C.; BAGOT, R. E.; PALATRESI, J. V. Automatic term detection: A review of current systems. **Recent advances in computational terminology**, Amsterdam/Philadelphia: John Benjamins Publishing Company, v. 2, p. 53–88, 2001. Citado na página 59.

CGU. **Painel Lei de Acesso à Informação**. 2020. Citado nas páginas 216 e 217.

CONRADO, M. da S.; FELIPPO, A. D.; PARDO, T. A. S.; REZENDE, S. O. A survey of automatic term extraction for brazilian portuguese. **Journal of the Brazilian Computer Society**, SpringerOpen, v. 20, n. 1, p. 1–28, 2014. Citado na página 243.

COUNCIL, E. U. P. General data protection regulation. **Official Journal of the European Union**, 2016. Citado na página 31.

DAI, H.; SONG, Y. Neural aspect and opinion term extraction with mined rules as weak supervision. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 5268–5277. Citado nas páginas 134 e 243.

DAUPHIN, Y. N.; FAN, A.; AULI, M.; GRANGIER, D. Language modeling with gated convolutional networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 933–941. Citado na página 66.

DEORAS, A.; KOMBRINK, S. *et al.* Empirical evaluation and combination of advanced language modeling techniques. Citeseer, 2011. Citado na página 66.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. [S.l.]: arXiv, 2018. Citado nas páginas 71, 73, 74, 75, 76 e 177.

DOSHI-VELEZ, F.; KIM, B. **Towards A Rigorous Science of Interpretable Machine Learning**. [S.l.]: arXiv, 2017. Citado na página 48.

DRAGONI, M.; VILLATA, S.; RIZZI, W.; GOVERNATORI, G. Combining Natural Language Processing Approaches for Rule Extraction from Legal Documents. In: PAGALLO, U.; PALMIRANI, M.; CASANOVAS, P.; SARTOR, G.; VILLATA, S. (Ed.). **AI Approaches to the Complexity of Legal Systems**. Cham: Springer International Publishing, 2018. p. 287–300. ISBN 978-3-030-00178-0. Citado na página 133.

DUTKIEWICZ, L. From right to safeguards. the explainability of automated decision-making under the general data protection regulation. In: **Towards a Global Taxonomy of Interpretable AI, Location: online**. [S.l.: s.n.], 2021. Citado na página 31.

EL-DIN, D. M. Enhancement bag-of-words model for solving the challenges of sentiment analysis. **International Journal of Advanced Computer Science and Applications**, v. 7, n. 1, 2016. Citado na página 81.

eSIC. **Sistema Eletrônico do Serviço de Informação ao Cidadão**. 2020. Citado na página 218.

ESTOPÀ, R.; BURGOS, D. *et al.* La identificación de unidades terminológicas en contexto: de la teoría a la práctica. **La identificación de unidades terminológicas en contexto: de la teoría a la práctica**, Documenta Universitaria, p. 1000–1030, 2006. Citado na página 59.

FACELI ANA CAROLINA LORENA, J. G. T. A. d. A. K.; CARVALHO, A. C. P. L. F. de. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2021. Citado na página 51.

FALABR. **Portal Fala Br**. 2020. Disponível em: <https://falabr.cgu.gov.br>. Citado na página 218.

FONSECA, E.; SANTOS, L. Borges dos; CRISCUOLO, M.; ALUíSIO, S. Visão geral da avaliação de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, 2016. Citado na página 57.

FRANTZI, K. T.; ANANIADOU, S.; TSUJII, J. The c-value/nc-value method of automatic recognition for multi-word terms. In: SPRINGER. **International conference on theory and practice of digital libraries**. [S.l.], 1998. p. 585–604. Citado nas páginas 59 e 133.

GAIKWAD, S. V.; CHAUGULE, A.; PATIL, P. Text mining methods and techniques. **International Journal of Computer Applications**, Foundation of Computer Science, v. 85, n. 17, 2014. Citado nas páginas 37, 38, 62 e 63.

GEORGIEVA-TRIFONOVA, T. Text classification based on enriched vector space model. In: **Proceedings of the 18th International Conference on Computer Systems and Technologies**. New York, NY, USA: Association for Computing Machinery, 2017. (CompSysTech'17), p. 103–110. ISBN 9781450352345. Citado na página 79.

GIANOTI, A. C. **Descrição sintático-semântica dos termos do domínio da educação à distância em português do Brasil**. Dissertação (Trabalho de Conclusão de Curso (Graduação em Bacharelado em Liguística)) — Universidade Federal de São Carlos (UFSCar), São Carlos, SP, 2012. Citado na página 61.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **Proceedings of the 16th Conference on Computational Linguistics - Volume 1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996. (COLING '96), p. 466–471. Citado na página 57.

GUNDU, S. R.; ANURADHA, T. Digital data growth and the philosophy of digital universe in view of emerging technologies. **Int. J. Sci. Res. in Computer Science and Engineering Vol**, v. 8, n. 2, 2020. Citado na página 29.

GUNNING, D. Explainable artificial intelligence (xai). **Defense advanced research projects agency (DARPA), nd Web**, v. 2, n. 2, p. 1, 2017. Citado nas páginas 48 e 223.

HAMON, R.; JUNKLEWITZ, H.; SANCHEZ, I.; MALGIERI, G.; HERT, P. D. Bridging the gap between ai and explainability in the gdpr: Towards trustworthiness-by-design in automated decision-making. **IEEE Computational Intelligence Magazine**, v. 17, n. 1, p. 72–85, 2022. Citado na página 31.

HAN, J. Mining heterogeneous information networks: the next frontier. **In KDD'12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 1–3, 2012. Citado na página 46.

HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2016. p. 202–212. Citado na página 243.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEGI, M. R. Text mining in big data analytics. **Big Data and Cognitive Computing**, MDPI, v. 4, n. 1, p. 1, 2020. Citado na página 37.

HICKMAN, L.; THAPA, S.; TAY, L.; CAO, M.; SRINIVASAN, P. Text preprocessing for text mining in organizational research: Review and recommendations. **Organizational Research Methods**, SAGE Publications Sage CA: Los Angeles, CA, v. 25, n. 1, p. 114–146, 2022. Citado na página 37.

HOANG, M.; BIHORAC, O. A.; ROUCES, J. Aspect-based sentiment analysis using BERT. In: **Proceedings of the 22nd Nordic Conference on Computational Linguistics**. Turku, Finland: Linköping University Electronic Press, 2019. p. 187–196. Citado nas páginas 169, 170 e 238.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997. Citado nas páginas 68 e 69.

HOUAISS, A.; VILLAR, M. d. S.; FRANCO, F. M. d. M. Dicionário eletrônico houaiss da língua portuguesa versão 1.0. **Rio de Janeiro: Editora Objetiva**, v. 1, 2001. Citado na página 61.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: . [S.l.: s.n.], 2004. p. 168–177. Citado nas páginas 95, 142, 169 e 174.

ISO1087. **ISO1087: Terminoligie - Vocabularie**. 1990. ISO, Norme Internationale ISO 1087. Geneva, Switzerland. Citado nas páginas 59 e 60.

JANZ, A.; KDZIA, P.; PIASECKI, M. Graph-based complex representation in inter-sentence relation recognition in polish texts. **Cybernetics and Information Technologies**, v. 18, n. 1, p. 152–170, 2018. Cited By 0. Citado na página 44.

JIANHUA, L.; ZHIXIONG, Z.; JIAN, X.; YANDONG, X. Automatic term recognition——an important method for text mining on scientific literature. **Data Analysis and Knowledge Discovery**, v. 24, n. 8, p. 12–17, 2008. Citado na página 133.

JO, T. **Text mining: Concepts, implementation, and big data challenge**. [S.l.]: Springer, 2018. v. 45. Citado nas páginas 37 e 38.

JU, H.; YU, H. Sentiment Classification with Convolutional Neural Network using Multiple Word Representations. In: **12th International Conference on Ubiquitous Information Management and Communication - IMCOM '18**. [S.l.]: ACM Press, 2018. p. 1–7. Citado na página 81.

KAGEURA, K.; UMINO, B. Methods of automatic term recognition: A review. **Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication**, John Benjamins, v. 3, n. 2, p. 259–289, 1996. Citado na página 63.

KUMAR, S.; KAR, A. K.; ILAVARASAN, P. V. Applications of text mining in services management: A systematic literature review. **International Journal of Information Management Data Insights**, v. 1, n. 1, p. 100008, 2021. ISSN 2667-0968. Citado na página 37.

LI, Y.; WEI, B.; LIU, Y.; YAO, L.; CHEN, H.; YU, J.; ZHU, W. Incorporating knowledge into neural network for text representation. **Expert Systems with Applications**, v. 96, p. 103 – 114, 2018. ISSN 0957-4174. Citado na página 30.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. **Entropy**, MDPI, v. 23, n. 1, p. 18, 2020. Citado nas páginas 48 e 223.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan & Claypool Publishers, 2012. ISBN 1608458849, 9781608458844. Citado na página 56.

LOMBARDI, M.; MARANI, A. Synfinder: A system for domain-based detection of synonyms using wordnet and the web of data. In: SIDOROV, G.; GALICIA-HARO, S. N. (Ed.). **Advances in Artificial Intelligence and Soft Computing**. Cham: Springer International Publishing, 2015. p. 15–28. ISBN 978-3-319-27060-9. Citado na página 243.

LU, Q.; GETOOR, L. Link-based classification. In: ACM. **In ICML'2003: Proceedings of the International Conference on Machine Learning**. [S.l.], 2003. p. 496—-503. Citado na página 45.

LUCY, L.; GAUTHIER, J. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. **arXiv preprint arXiv:1705.11168**, 2017. Citado na página 67.

MARCINIAK, M.; MYKOWIECKA, A. Terminology extraction from medical texts in polish. **Journal of biomedical semantics**, v. 5, n. 24, 2014. Citado nas páginas 133 e 243.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. **In Computational Linguistics**, v. 19, n. 2, p. 313–330, 1993. Citado nas páginas 135, 257 e 259.

MARQUES, C. A. N.; MATSUNO, I. P.; SINOARA, R. A.; REZENDE, S. O.; ROZENFELD, H. An exploratory study to evaluate the practical application of pss methods and tools based on text mining. In: **Proceedings of the 20th International Conference on Engineering Design**. [S.l.: s.n.], 2015. p. 7–311–7–320. Citado nas páginas 58, 77, 82, 84 e 236.

MARTIN, J. H. **Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition**. [S.l.]: Pearson/Prentice Hall, 2009. Citado na página 64.

MARTINS, A. C. M.; LOPES, O. A.; CADEMARTORI, S. U. **O STF e a divulgação nominalmente individualizada da remuneração dos servidores púbicos: uma análise do Recurso Extraordinário 652.777-SP**. Tese (Dissertação (Mestrado em Direito)) — Universidade de Brasília, 2017. Citado na página 225.

MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Aspect-based sentiment analysis using semi-supervised learning in bipartite heterogeneous networks. **JIDM**, v. 7, n. 2, p. 141–154, 2016. Citado na página 48.

MCCANN, B.; BRADBURY, J.; XIONG, C.; SOCHER, R. Learned in translation: Contextualized word vectors. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 68.

MENDEL, T. Liberdade de informação: um estudo de direito comparado - 2. ed. **UNESCO**, 2009. Citado na página 215.

MEURISCH, C.; MüHLHäUSER, M. Data protection in ai services: A survey. **ACM Computing Surveys**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 2, mar 2021. ISSN 0360-0300. Citado na página 31.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citado na página 67.

MILLER, G. A. **WordNet: An electronic lexical database**. [S.l.]: MIT press, 1998. Citado na página 243.

MISHRA, B.; KUMAR, V.; PANDA, S.; TIWARI, P. Handbook of research for big data: Concepts and techniques. In: ____. [S.l.: s.n.], 2020. cap. 3, p. 231–244. ISBN 9781771889803. Citado na página 29.

MITCHELL, T. M. **Machine Learning**. New York, NY, USA: McGraw-Hill, 1997. Citado na página 51.

MITKOV, R. **The Oxford handbook of computational linguistics**. [S.l.]: Oxford University Press, 2022. Citado na página 65.

MOONS, E.; TUYTELAARS, T.; MOENS, M.-F. Text-enriched representations for news image classification. In: **Companion Proceedings of the The Web Conference 2018**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 99–100. ISBN 9781450356404. Citado na página 81.

MORO, A.; RAGANATO, A.; NAVIGLI, R. Entity linking meets word sense disambiguation: A unified approach. **Transactions of the Association for Computational Linguistics**, v. 2, p. 231–244, 2014. ISSN 2307-387X. Citado na página 57.

MUJTABA, G.; SHUIB, L.; RAJ, R.; RAJANDRAM, R.; SHAIKH, K.; AL-GARADI, M. Classification of forensic autopsy reports through conceptual graph-based document representation model. **Journal of Biomedical Informatics**, v. 82, p. 88–105, 2018. Cited By 0. Citado na página 44.

MüLLER, A. **Introdução à Linguística II: Princípios de Análise**. [S.l.]: Editora Contexto, 2010. P. 137–159. Citado na página 52.

NAZIR, S.; YOUSAF, M. H.; NEBEL, J.-C.; VELASTIN, S. A. A bag of expression framework for improved human action recognition. **Pattern Recognition Letters**, v. 103, p. 39–45, 2018. ISSN 0167-8655. Citado na página 81.

NEGI, A. A brief survey on text mining, its techniques, and applications. **Int. J. Mob. Comput. Appl**, v. 8, n. 1, p. 1–6, 2021. Citado na página 29.

NIKISHINA, I.; TIKHOMIROV, M.; LOGACHEVA, V.; NAZAROV, Y.; PANCHENKO, A.; LOUKACHEVITCH, N. Taxonomy enrichment with text and graph vector representations. **arXiv preprint arXiv:2201.08598**, 2022. Citado nas páginas 30 e 81.

NÓBREGA, F. A. A.; PARDO, T. A. S. General purpose word sense disambiguation methods for nouns in portuguese. In: BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T. A. S.; NUNES, M. d. G. V. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2014. p. 94–101. ISBN 978-3-319-09761-9. Citado na página 57.

OGADA, K.; MWANGI, W.; CHERUIYOT, W. N-gram based text categorization method for improved data mining. **Journal of Information Engineering and Applications**, Citeseer, v. 5, n. 8, p. 35–43, 2015. Citado na página 81.

OLIVEIRA, A. S. M. Semantic role labeling in portuguese: improving the state of the art with transfer learning and bert-based models. 2020. Citado na página 243.

OLIVEIRA, H. G. A survey on portuguese lexical knowledge bases: Contents, comparison and combination. **Information**, v. 9, n. 2, 2018. Citado nas páginas 51 e 52.

OLIVEIRA, R. Pires de. **Introdução à linguística: domínios e fronteiras**. [S.l.]: Cortez, 2012. Volume 2. P. 23–54. Citado na página 51.

PALATRESI, J. V. Extracción de candidatos a término mediante la combinación de estrategias heterogéneas. **Procesamiento del lenguaje natural, nº 28 (mayo 2002); pp. 111-112**, Sociedad Española para el Procesamiento del Lenguaje Natural, 2002. Citado nas páginas 59 e 64.

PALMER, M.; GILDEA, D.; XUE, N. **Semantic Role Labeling**. [S.l.]: Morgan & Claypool Publishers, 2010. (Synthesis Lectures on Human La). ISBN 9781598298314. Citado na página 57.

PAZIENZA, M. T.; PENNACCHIOTTI, M.; ZANZOTTO, F. M. Terminology extraction: an analysis of linguistic and statistical approaches. In: **Knowledge mining**. [S.l.]: Springer, 2005. p. 255–279. Citado nas páginas 62, 63, 65, 66 e 133.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 67.

PETERS, M. E.; AMMAR, W.; BHAGAVATULA, C.; POWER, R. Semi-supervised sequence tagging with bidirectional language models. **arXiv preprint arXiv:1705.00108**, 2017. Citado na página 68.

PIETROFORTE, A. V. **Introdução à Linguística II: Princípios de Análise**. [S.l.]: Editora Contexto, 2010. P. 111–135. Citado nas páginas 51 e 52.

PONTIKI, M.; GALANIS, D.; ANDROUTSOPOULOS, I.; MANANDHAR, S.; PAPAGEORGIOU, H. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: **8th International Workshop on Semantic Evaluation**. [S.l.: s.n.], 2014. p. 27–35. Citado nas páginas 97, 142 e 175.

PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; MANANDHAR, S.; ANDROUTSOPOULOS, I. Semeval-2015 task 12: Aspect based sentiment analysis. **Proceedings of the 9th International Workshop on Semantic Evaluation**, p. 486–495, 2015. Citado nas páginas 85, 86, 98, 142 e 175.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. *et al.* Improving language understanding by generative pre-training. OpenAI, 2018. Citado na página 71.

RANA, T. A.; CHEAH, Y.-N. A two-fold rule-based model for aspect extraction. **Expert Systems with Applications**, v. 89, p. 273–285, 2017. ISSN 0957-4174. Citado na página 133.

REAL, L.; OSHIRO, M.; MAFRA, A. B2w-reviews01: An open product reviews corpus. In: **Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2019. Citado nas páginas 94, 142 e 174.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. d. Mineração de dados. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicações**. [S.l.]: Editora Manole, 2003. p. 307–335. Citado nas páginas 29, 37 e 38.

RIEMER, N. **Introducing Semantics**. [S.l.]: Cambridge University Press, 2010. (Cambridge Introductions to Language and Linguistics). Citado nas páginas 51 e 54.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional., 2016. Citado nas páginas 47 e 48.

ROSSI, R. G.; LOPES, A. A.; REZENDE, S. O. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. **Information Processing & Management**, v. 52, n. 2, p. 217–257, 2016. Citado nas páginas 44, 81, 84 e 87.

ROSSI, R. G.; LOPES, A. de A.; FALEIROS, T. de P.; REZENDE, S. O. Inductive model generation for text classification using a bipartite heterogeneous network. **Journal of Computer Science and Technology**, v. 29, n. 3, p. 361–375, 2014. ISSN 1860-4749. Citado nas páginas 44, 45, 48, 81, 84 e 87.

ROSSI, R. G.; REZENDE, S. O. Generating features from textual documents through association rules. **ENIA: Encontro Nacional de Inteligência Artificial**, v. 01, p. 311–322, 2011. Citado na página 79.

SABRA, S.; SABEEH, V. A comparative study of n-gram and skip-gram for clinical concepts extraction. In: **2020 International Conference on Computational Science and Computational Intelligence (CSCI)**. [S.l.: s.n.], 2020. p. 807–812. Citado na página 243.

SAGER, J. C. A practical course in terminology processing. **A Practical Course in Terminology Processing**, John Benjamins Publishing Company, p. 1–270, 1990. Citado na página 59.

SANTOS, B. N. D.; MARCACINI, R. M.; REZENDE, S. O. Multi-domain aspect extraction using bidirectional encoder representations from transformers. **IEEE Access**, v. 9, p. 91604–91613, 2021. Citado na página 170.

SCHEICHER, R. B.; SINORARA, R. A.; FELINTO, J. d. C.; REZENDE, S. O. Sentiment classification improvement using semantically enriched information. In: **Proceedings of the ACM Symposium on Document Engineering 2019**. New York, NY, USA: Association for Computing Machinery, 2019. (DocEng '19). ISBN 9781450368872. Citado na página 84.

SCHEICHER, R. B.; SINORARA, R. A.; KOGA, N. J.; REZENDE, S. O. Uso de expressões do domínio na classificação automática de documentos. **XIII Encontro Nacional de Inteligência Artificial e Computacional**, Volume 1, p. 625 – 636, 2016. Citado nas páginas 88 e 89.

SCHEICHER, R. B.; SINORARA, R. A.; REZENDE, S. O. Exploração do uso de expressões do domínio na classificação de sentimentos. In: **Anais da I Escola Regional de Aprendizado de Máquina e Inteligência Artificial de São Paulo**. Porto Alegre, RS, Brasil: SBC, 2020. p. 38–42. ISSN 0000-0000. Citado na página 84.

SCHLUTTER, A.; VOGELSANG, A. Knowledge representation of requirements documents using natural language processing. In: . [S.l.: s.n.], 2018. v. 2075. Cited By 0. Citado na página 44.

SCHWENK, H. Continuous space language models. **Computer Speech & Language**, Elsevier, v. 21, n. 3, p. 492–518, 2007. Citado na página 66.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. ISSN 0360-0300. Citado nas páginas 47, 48, 50, 51 e 56.

SHAH, N.; JESHWANI, S.; BHATT, P. An introduction on interpretable machine learning. In: . [S.l.: s.n.], 2020. Citado nas páginas 48 e 49.

SILVA, B. C. Dias-da. Wordnet. br: An exercise of human language technology research. In: MASARYKOVA UNIV. **Gwc 2006: Third International Wordnet Conference, Proceedings**. [S.l.], 2005. p. 301–303. Citado na página 243.

SINOARA, R. A. **Aspectos semânticos na representação de textos para classificação automática**. Tese (Doutorado) — Universidade de São Paulo, 2018. Citado nas páginas 31, 54, 55, 56, 85, 86, 225 e 236.

SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. **Journal of the Brazilian Computer Society**, v. 23, p. 1–20, 2017. Citado nas páginas 30 e 81.

SINOARA, R. A.; REZENDE, S. O. Best sports: a portuguese collection of documents for semantics-concerned text mining research. **[S.l.]**, 2018. Citado nas páginas 85, 94 e 174.

SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. **Semantic Role-based Representations in Text Classification**. 2016. Citado na página 80.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado na página 177.

SOUZA, J. W. C.; FELIPPO, A. D. **Um exercício em lingüistica de Corpus no âmbito do Projeto TermiNet**. [S.l.], 2010. Citado na página 59.

STANKOVIĆ, R.; KRSTEV, C.; OBRADOVIĆ, I.; LAZIĆ, B.; TRTOVAC, A. Rule-based automatic multi-word term extraction and lemmatization. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 507–514. Citado nas páginas 133 e 243.

SUN, C.; HUANG, L.; QIU, X. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. **arXiv preprint arXiv:1903.09588**, 2019. Citado na página 73.

SUN, Y.; HAN, J. Mining heterogeneous information networks: Principles and methodologies. **Synthesis Lectures on Data Mining and Knowledge Discovery**, 2012. Morgan & Claypool Publishers. Citado na página 46.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2005. 18, 26, 42, 64, 76, 78, 83, 87, e 100 p. Citado na página 43.

TANG, D.; WEI, F.; YANG, N.; ZHOU, M.; LIU, T.; QIN, B. Learning sentiment-specific word embedding for twitter sentiment classification. In: **52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. v. 1, p. 1555–1565. Citado na página 81.

TELINE, M. F.; MANFRIN, A. M. P.; ALUISIO, S. M. *et al.* Extração automática de termos de textos em português: aplicação e avaliação de medidas estatísticas e associação de palavras. São Carlos, SP, Brasil., 2003. Citado na página 66.

TIAN, M.; CUI, R.; HUANG, Z. Automatic extraction method for specific domain terms based on structural features and mutual information. In: **2018 5th International Conference on Information Science and Control Engineering (ICISCE)**. [S.l.: s.n.], 2018. p. 147–150. Citado na página 134.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **CoRR**, abs/1003.1141, 2010. Citado na página 57.

VASQUES, D. G.; REZENDE, S. O.; MARTINS, P. S. A semantic approach to uncovering implicit relationships in textual databases. In: IEEE. **2018 XLIV Latin American Computer Conference (CLEI)**. [S.l.], 2018. p. 490–499. Citado na página 243.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30. Citado nas páginas 70, 71 e 73.

VIEIRA, R.; GONçALVES, P.; SOUZA, J. Processamento computacional de anáfora e correferência. **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 16, n. 1, p. 263–284, 2008. ISSN 2237-2083. Citado na página 57.

VIVALDI, J.; RODRÍGUEZ, H. Evaluation of terms and term extraction systems: A practical approach. **Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication**, John Benjamins, v. 13, n. 2, p. 225–248, 2007. Citado na página 64.

VOUTILAINEN, A. **Part-of-speech tagging**. [S.l.]: The Oxford handbook of computational linguistics, 2003. v. 219. Citado na página 65.

WACHTER, S. Data protection in the age of big data. **Nature Electronics**, Nature Publishing Group, v. 2, n. 1, p. 6–7, 2019. Citado na página 31.

WANG, X.; JIA, Y.; CHEN, R.; FAN, H.; ZHOU, B. Improving text categorization with semantic knowledge in wikipedia. **IEICE TRANSACTIONS on Information and Systems**, The Institute of Electronics, Information and Communication Engineers, v. 96, n. 12, p. 2786–2794, 2013. Citado na página 79.

WIDDOWS, D. **Geometry and meaning**. [S.l.]: Citeseer, 2004. v. 773. Citado na página 225.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K. *et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. **arXiv preprint arXiv:1609.08144**, 2016. Citado na página 73.

XAVIER, F. L. S. **Direito à informação na era da inteligência artificial: classificação dos pedidos de acesso á informação no Brasil**. Dissertação (MBA em Ciência de Dados) — Instituto de Ciências Matemáticas e de Computação, 2021. Citado nas páginas 219, 220, 221, 222, 223, 224, 226, 228 e 231.

XAVIER, F. L. S.; REZENDE, S. O.; SCHEICHER, R. B. Right to information in the era of artificial intelligence in brazil. In: **26th World Congress of Political Science (New Nationalisms in an Open World)**. Lisboa: [s.n.], 2021. Citado na página 231.

XIONG, S. Improving twitter sentiment classification via multi-level sentiment-enriched word embeddings. In: . [S.l.: s.n.], 2016. Citado na página 81.

XU, H.; LIU, B.; SHU, L.; YU, P. S. Bert post-training for review reading comprehension and aspect-based sentiment analysis. **arXiv preprint arXiv:1904.02232**, 2019. Citado nas páginas 73 e 169.

YAN, D.; LI, K.; GU, S.; YANG, L. Network-based bag-of-words model for text classification. **IEEE Access**, IEEE, v. 8, p. 82641–82652, 2020. Citado nas páginas 30 e 81.

YANUAR, M. R.; SHIRAMATSU, S. Aspect extraction for tourist spot review in indonesian language using bert. In: **2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)**. [S.l.: s.n.], 2020. p. 298–302. Citado na página 170.

YU, Y.; SI, X.; HU, C.; ZHANG, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. **Neural Computation**, v. 31, n. 7, p. 1235–1270, 07 2019. ISSN 0899-7667. Citado nas páginas 68, 69 e 70.

ZAVAGLIA, C.; OLIVEIRA, L. H. M. d.; NUNES, M. d. G. V.; ALUÍSIO, S. M. Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. **Anais**, 2007. Citado na página 59.

ANEXO

A

RÓTULOS *PART-OF-SPEECH* (POS) DOS PADRÕES DE ANOTAÇÃO

Nesse anexo são apresentados os rótulos dos padrões de anotação *Part-of-Speech (POS)* utilizadas na extração de termos baseada em padrões sintáticos. Na Tabela 93 são apresentados os rótulos aplicados pelo anotador NLTK treinado a partir do *corpus Mac-Morpho* (ALUÍSIO *et al.*, 2003), para o idioma português. Na Tabela 94 são apresentados os rótulos da ferramenta "*pos_tag*", para o idioma inglês, contida na biblioteca NLTK da linguagem Python 3.7, que segue o padrão *English Penn Treebank tagset* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993). As tabelas possuem três colunas. A primeira corresponde ao rótulo inserido por cada ferramenta, a segunda coluna corresponde à descrição do rótulo e a terceira coluna apresenta um exemplo de palavras ou sentenças que fazem parte de um determinado rótulo.

Tabela 93 – Rótulos *Part-of-Speech (POS)* do padrão Mac-Morpho.

POS Tag	**Descrição**	**Exemplo**
ADJ	Adjetivo	Era bela_ADJ mas vazia_ADJ.
ADV	Advérbio	Ela era bem_ADV vaidosa.
ADV-KS	Advérbio conectivo subordinativo	Sei onde_ADV-KS mora.
ADV-KS-REL	Advérbio relativo subordinativo	Sei o lugar onde_ADV-KS-REL mora.
ART	Artigo (def. ou indef.)	Era um_ART homem de sorte.
KC	Conjunção coordenativa	Comeu e_KC bebeu.
KS	Conjunção subordinativa	Sairei quando_KS puder.
IN	Interjeição	Pare_IN! (Fique parado!)
N	Nome / Substantivo	Plantou uma árvore_N enorme.
NPROP	Nome próprio	Maria_NPROP foi ao mercado.
NUM	Numeral	Haviam duas_NUM irmãs.
PCP	Particípio	Pegou a blusa amassada_PCP e saiu.
PDEN	Palavra denotativa	Somente_PDEN Maria foi à festa.
PREP	Preposição	O livro está sobre_PREP a mesa.
PROADJ	Pronome adjetivo	Cada_PROADJ um com a sua mania.
PRO-KS	Pronome conectivo subordinativo	Diga-me de qual_PRO-KS livro você mais gostou.
PROPESS	Pronome pessoal	Ela_PROPESS tinha três irmãs.
PRO-KS-REL	Pronome relativo conectivo subordinativo	O livro, cujo_PRO-KS-REL autor foi muito festejado na Bienal, não fez sucesso na Europa.
PROSUB	Pronome substantivo	Alguém_PROSUB ligou?
V	Verbo	Eu moro_V em São Carlos.
VAUX	Verbo auxiliar	Sua tia estava_VAUX para chegar_V.
CUR	Símbolo de moeda corrente	Preciso de R$_CUR 10,00_NUM.
|EST |AP |DAD |TEL |DAT |HOR |[|]	Etiquetas complementares (Estrangeirismos; Apostos; Dados; Números de Telefone; Datas; Horas; e Disjunção)	Dizia que estava muito down_ADJ|EST e que não queria sair hoje.
+	Contrações e Ênclises	Desculpe-_V|+ me_PROPESS pelo acontecido.
!	Mesóclises	Daria_V|! lhe_PROPESS o mundo! (dar-lhe-ia)

Fonte: Adaptada de Aluísio *et al.* (2003).

Tabela 94 – Rótulos *Part-of-Speech (POS)* do padrão *English Penn Treebank*.

POS Tag	Descrição	Exemplo
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	les
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PPZ	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
“	Quotation marks	‘ “
“	Opening quotation marks	‘ “
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	– ; : — …

Fonte: Marcus, Santorini e Marcinkiewicz (1993).


ICMC
USP
SÃO CARLOS