UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Exploração de informações contextuais para enriquecimento semântico em representações de textos

João Vítor Antunes Ribeiro

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

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

João Vítor Antunes Ribeiro

Exploração de informações contextuais para enriquecimento semântico em representações de textos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
Dezembro de 2018

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

A636e   Antunes Ribeiro, João Vítor
             Exploração de informações contextuais para
             enriquecimento semântico em representações de textos
             / João Vítor Antunes Ribeiro; orientadora Solange
             Oliveira Rezende. -- São Carlos, 2018.
             131 p.

             Dissertação (Mestrado - Programa de Pós-Graduação
             em Ciências de Computação e Matemática
             Computacional) -- Instituto de Ciências Matemáticas
             e de Computação, Universidade de São Paulo, 2018.

             1. Representação de textos. 2. Semântica. 3.
             Contexto. 4. Classificação. I. Oliveira Rezende,
             Solange, orient. II. Título.

João Vítor Antunes Ribeiro

Exploration of Contextual Information for Semantic Enrichment in Text Representations

Master dissertation submitted to the Institute of
Mathematics and Computer Sciences – ICMC-USP,
in partial fulfillment of the requirements for the
degree of the Master Program in Computer Science
and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
December 2018

*Este trabalho é dedicado aos meus familiares e amigos,
em especial aos meus pais, Yolanda e Marcos, e aos meus avós,
Esther e Waldomiro Antunes, Verônica e José Ribeiro.*

AGRADECIMENTOS

Embora o desenvolvimento desta dissertação compreenda apenas os três últimos anos, sua realização só foi possível devido ao esforço inicial de uma série de pessoas. Por esse motivo, gostaria de direcionar agradecimentos em especial aos meus pais, Marcos e Yolanda, aos meus irmãos, Gabriela e Marcos Junior, à minha parceira, Jéssica, e à toda a minha família e amigos pela lealdade, pela solidariedade e pela integridade. Gostaria de ressaltar a importância das minhas primas Angelita, Clarice e Rosa, e a de toda a sua família para o início dessa jornada acadêmica; agradecimentos pela hospitalidade, pela solidariedade e pela fraternidade.

Em relação ao período da minha graduação, realizada na FCT-UNESP em Presidente Prudente - SP, gostaria de direcionar agradecimentos aos funcionários dessa instituição de ensino, em especial aos meus orientadores, Danilo, Milton e Marco Piteri, pelo aprendizado, pela generosidade e pela confiança. Também gostaria de ressaltar a importância dos demais professores dessa instituição para a minha formação, Almir, Celso, Fabrício, Maurício, Raphael, Rogério e Ronaldo, entre outros; agradecimentos pelo aprendizado, pela competência e pela ponderação. Aos amigos de república, Antonio (“Tonho”), Emerson (“Mersão”), Gabriel (“Gabi-Gol”), Guilherme (“Alemão”), Haroldo, Nicolas (“Linux”), Victor Sato e Vitor (“Mineiro”), aos amigos do CPIDES, Alisson, Ana Maria, Danielle, Livia, Luiza, Marcella, Paula e Stephanie, e aos demais amigos, André Andreotti, Beatriz (“Bia”), Cintia (“JapaGirl”), Douglas Panacho, Fernando Pacanelli, João, Juliana (“Juh”) e Priscila (“Pri”), entre outros, gostaria de direcionar agradecimentos pelo aprendizado, pelo companheirismo e pela solidariedade.

Em relação ao período do meu mestrado, gostaria de direcionar agradecimentos aos funcionários dessa instituição de ensino pela atenção, pela empatia e pela consistência. Em especial à minha orientadora, Solange, gostaria de direcionar agradecimentos pelas oportunidades, pelo aprendizado e pela confiança. Aos amigos em geral, Alan, André, Antonio, Camila, Dario, Diana, Diego, Dildre, Ivone, Jarlinton, Jorge, Rafael (“RG”), Renan, Ricardo Puma, Ricardo Scheicher, Roberta e Vinícius, entre outros, gostaria de direcionar agradecimentos pelo aprendizado, pelo companheirismo e pela solidariedade.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Código de Financiamento #132666/2016-2. Agradecimentos especiais aos desenvolvedores dos recursos em geral utilizados nesta dissertação por disponibilizarem gratuitamente tal conteúdo. Agradecimentos também aos integrantes da minha banca, Alípio, Edson e Thiago, pelas correções, pelos debates e pelas contribuições.

RESUMO

ANTUNES, J. **Exploração de informações contextuais para enriquecimento semântico em representações de textos**. 2018. 131 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Em decorrência da crescente quantidade de documentos disponíveis em formato digital, a importância da análise computacional de grandes volumes de dados torna-se ainda mais evidente na atualidade. Embora grande parte desses documentos esteja disponível em formato de língua natural, a análise por meio de processos como a Mineração de Textos ainda é um desafio a ser superado. Normalmente, abordagens tradicionais de representação de textos como a *Bag of Words* desconsideram aspectos semânticos e contextuais das coleções de textos analisadas, ignorando informações que podem potencializar o desempenho das tarefas realizadas. Os principais problemas associados a essas abordagens são a alta esparsidade e dimensionalidade que prejudicam consideravelmente o desempenho das tarefas realizadas. Como o enriquecimento de representações de textos é uma das possibilidades efetivas para atenuar esses tipos de problemas, nesta dissertação foi investigada a aplicação conjunta de enriquecimentos semânticos e contextuais. Para isso foi proposta uma nova técnica de representação de textos, cuja principal novidade é a abordagem utilizada para calcular a frequência dos atributos (contextos) baseando-se em suas similaridades. Os atributos extraídos por meio dessa técnica proposta são considerados dependentes já que são formados por conjuntos de termos correlacionados que podem compartilhar informações semelhantes. A efetividade da técnica foi avaliada na tarefa de classificação automática de textos, na qual foram explorados diferentes procedimentos de enriquecimento textual e versões de modelos de linguagem baseados em *word embeddings*. De acordo com os resultados obtidos, há evidências favoráveis a respeito da efetividade e da aplicabilidade da técnica de representação de textos proposta. Segundo os testes de significância estatística realizados, a aplicação de enriquecimentos textuais baseados em Reconhecimento de Entidades Nomeadas e em Desambiguação Lexical de Sentido pode contribuir efetivamente para o aumento do desempenho da tarefa de classificação automática de textos, principalmente nas abordagens em que também são considerados textos de fontes externas de conhecimento como a Wikipédia. Constatou-se empiricamente que a efetividade dessa técnica proposta pode ser superior às abordagens tradicionais em cenários de aplicação baseados em informações semânticas das coleções de textos, caracterizando-a como uma alternativa promissora para a geração de representações de textos com alta densidade de informações semânticas e contextuais que se destacam pela interpretabilidade.

Palavras-chave: Representação de textos, Semântica, Contexto, Classificação.

ABSTRACT

ANTUNES, J. **Exploration of Contextual Information for Semantic Enrichment in Text Representations**. 2018. 131 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Due to the increasing number of available documents in digital format, the importance of computational analysis of large volumes of data becomes even more evident recently. Although most of these documents are available in natural language format, analysis through processes such as text mining is still a challenge to be overcome. Normally, traditional text representation approaches such as the bag of words disregard semantic and contextual aspects of the analyzed text collections, ignoring information that can enhance the performance of the tasks performed. The main problems associated with these approaches are the high sparsity and dimensionality that considerably impair the performance of the tasks performed. As the text representations enrichment is one of the effective possibilities to attenuate these types of problems, in this dissertation the joint application of semantic and contextual enrichment was investigated. For that a new text representation technique was proposed, whose main novelty is the approach used to calculate the frequency of attributes (contexts) based on their similarities. The attributes extracted by this proposed technique are considered dependent because they are formed by sets of correlated terms that can share similar information. The effectiveness of the technique was evaluated in the automatic text classification task, in which different procedures of textual enrichment and versions of language models based on word embeddings were explored. According to the results, there is favorable evidence regarding the effectiveness and applicability of the proposed text representation technique. According to the statistical significance tests, the application of textual enrichment based on named entity recognition and word sense disambiguation can effectively contribute to the increase of the performance of the automatic text classification task, especially in the approaches that are also considered texts from external knowledge sources such as Wikipedia. It has been empirically verified that the effectiveness of this proposed technique can be superior to the traditional approaches in application scenarios based on semantic information of the text collections, characterizing it as a promising alternative for the generation of text representations with high density of semantic and contextual information that stand out for their interpretability.

Keywords: Text Representation, Semantics, Context, Classification.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo do processo tradicional da etapa de Pré-processamento.	31
Figura 2 – Métricas de avaliação de desempenho da tarefa de classificação.	37
Figura 3 – Categorias de modelos composicionais distribuídos.	40
Figura 4 – Evolução do interesse científico por aspectos textuais semânticos na MT. . .	44
Figura 5 – Modelos do Word2Vec e exemplo de associações entre vetores densos. . . .	54
Figura 6 – Arquitetura da técnica de representação de textos CIRT.5.	56
Figura 7 – Exemplo de enriquecimento textual baseado em REN.	57
Figura 8 – Exemplo de delimitação de escopos baseados em sentenças.	58
Figura 9 – Exemplo de construção de um modelo baseado em <i>word embeddings</i>	59
Figura 10 – Exemplo de extração de contextos textuais baseados em sentenças.	60
Figura 11 – Exemplo de correlações entre contextos textuais.	60
Figura 12 – Performance das abordagens de representação de textos: 1º nível.	75
Figura 13 – Performance das abordagens de representação de textos: 2º nível.	76
Figura 14 – Performance das abordagens de representação de textos: 1º e 2º níveis. . . .	77
Figura 15 – Ranque das abordagens de representação de textos: 1º nível.	79
Figura 16 – Ranque das abordagens de representação de textos: 2º nível.	79
Figura 17 – Ranque das abordagens de representação de textos: 1º e 2º níveis.	80
Figura 18 – Correlações dos parâmetros da técnica CIRT.5: contextos e limiares.	83
Figura 19 – Performance das abordagens CIRT.5 em razão dos limiares de similaridade. .	85
Figura 20 – Distribuição de classes da base de dados BS-4.	108
Figura 21 – Distribuição de classes da base de dados SE.	109
Figura 22 – Distribuição de classes da base de dados BBC.	110

LISTA DE TABELAS

Tabela 1 – Coleções de textos preparadas para a tarefa de classificação.	68
Tabela 2 – Configuração paramétrica das representações de textos avaliadas.	71
Tabela 3 – Acurácia máxima na versão “raw” das coleções: 1º nível.	114
Tabela 4 – Acurácia máxima na versão “raw” das coleções: 1º nível (cont.).	115
Tabela 5 – Acurácia máxima na versão “word” das coleções: 1º nível.	116
Tabela 6 – Acurácia máxima na versão “word” das coleções: 1º nível (cont.).	117
Tabela 7 – Acurácia máxima na versão “id” das coleções: 1º nível.	118
Tabela 8 – Acurácia máxima na versão “id” das coleções: 1º nível (cont.).	119
Tabela 9 – Acurácia máxima na versão “raw” das coleções: 2º nível.	120
Tabela 10 – Acurácia máxima na versão “raw” das coleções: 2º nível (cont.).	121
Tabela 11 – Acurácia máxima na versão “word” das coleções: 2º nível.	122
Tabela 12 – Acurácia máxima na versão “word” das coleções: 2º nível (cont.).	123
Tabela 13 – Acurácia máxima na versão “id” das coleções: 2º nível.	124
Tabela 14 – Acurácia máxima na versão “id” das coleções: 2º nível (cont.).	125
Tabela 15 – Acurácia máxima na versão “raw” das coleções: 1º e 2º níveis.	126
Tabela 16 – Acurácia máxima na versão “raw” das coleções: 1º e 2º níveis (cont.).	127
Tabela 17 – Acurácia máxima na versão “word” das coleções: 1º e 2º níveis.	128
Tabela 18 – Acurácia máxima na versão “word” das coleções: 1º e 2º níveis (cont.).	129
Tabela 19 – Acurácia máxima na versão “id” das coleções: 1º e 2º níveis.	130
Tabela 20 – Acurácia máxima na versão “id” das coleções: 1º e 2º níveis (cont.).	131

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
BoV	<i>Bag of Vectors</i>
BoW	<i>Bag of Words</i>
CBoW	<i>Continuous Bag of Words</i>
CF-IDF	<i>Context Frequency - Inverse Document Frequency</i>
CIET.5	<i>Contextual Information Extraction Technique based on 5 steps</i>
CIRT.5	<i>Contextual Information Representation Technique based on 5 steps</i>
DLS	Desambiguação Lexical de Sentido
IA	Inteligência Artificial
ICMC	Instituto de Ciências Matemáticas e de Computação
IDF	<i>Inverse Document Frequency</i>
IMBHN	<i>Inductive Model based on Bipartite Heterogeneous Networks</i>
KNN	<i>K - Nearest Neighbors</i>
LABIC	Laboratório de Inteligência Computacional
LDA	<i>Latent Dirichlet Allocation</i>
LSA	<i>Latent Semantic Analysis</i>
LSI	<i>Latent Semantic Indexing</i>
MS	Mapeamento Sistemático
MT	Mineração de Textos
NASARI	<i>Novel Approach to a Semantically-Aware Representation of Items</i>
NB	<i>Naïve Bayes</i>
PLN	Processamento de Língua Natural
REN	Reconhecimento de Entidades Nomeadas
RNA	Rede Neural Artificial
SVM	<i>Support Vector Machines</i>
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency - Inverse Document Frequency</i>

SUMÁRIO

1	INTRODUÇÃO	21
1.1	Contextualização e definição do problema	21
1.2	Motivações e lacunas identificadas	23
1.3	Hipóteses definidas	26
1.4	Principais objetivos e novidades da pesquisa	26
1.5	Organização desta dissertação	27
2	MINERAÇÃO E ANÁLISE DE TEXTOS	29
2.1	Considerações iniciais	29
2.2	Mineração de Textos	30
2.2.1	<i>Etapas do processo tradicional de Mineração de Textos . . . . .</i>	<i>30</i>
2.2.2	<i>Tarefa de classificação automática de textos . . . . .</i>	<i>33</i>
2.2.2.1	<i>Paradigmas de Aprendizado de Máquina e algoritmos de classificação . . . . .</i>	<i>34</i>
2.2.2.2	<i>Avaliação de desempenho da tarefa de classificação . . . . .</i>	<i>36</i>
2.3	Análise de aspectos textuais semânticos	37
2.3.1	<i>Níveis de complexidade semântica . . . . .</i>	<i>40</i>
2.3.2	<i>Mapeamento da semântica na área de Mineração de Textos . . . . .</i>	<i>42</i>
2.3.2.1	<i>Roteiro de pesquisa do processo de mapeamento realizado . . . . .</i>	<i>42</i>
2.3.2.2	<i>Meta-análise da literatura . . . . .</i>	<i>43</i>
2.4	Representação de textos	46
2.4.1	<i>Tipos de informações e de relacionamentos textuais . . . . .</i>	<i>46</i>
2.4.2	<i>Abordagens e técnicas de representação de textos . . . . .</i>	<i>47</i>
2.5	Considerações finais	49
3	ENRIQUECIMENTO DE REPRESENTAÇÕES DE TEXTOS	51
3.1	Considerações iniciais	51
3.2	Abordagens e técnicas de enriquecimento textual	52
3.3	CIRT.5, uma proposta de representação de textos	56
3.3.1	<i>Etapas de Preparação: enriquecimento dos textos . . . . .</i>	<i>57</i>
3.3.2	<i>Etapas de Delimitação: delimitação dos escopos de interesse . . . . .</i>	<i>58</i>
3.3.3	<i>Etapas de Modelagem: construção do modelo de linguagem . . . . .</i>	<i>58</i>
3.3.4	<i>Etapas de Contextualização: extração dos contextos . . . . .</i>	<i>59</i>
3.3.5	<i>Etapas de Representação: correlação entre os contextos . . . . .</i>	<i>60</i>

3.4	Aplicações baseadas na técnica CIRT.5	61
3.4.1	<i>Representação de textos</i>	62
3.4.2	<i>Extração de contextos textuais</i>	64
3.5	Considerações finais	66
4	AVALIAÇÃO EMPÍRICA, RESULTADOS E DISCUSSÃO	67
4.1	Considerações iniciais	67
4.2	Organização e configuração experimental	68
4.2.1	<i>Bases de dados e recursos textuais utilizados</i>	68
4.2.2	<i>Abordagens de representação de textos avaliadas</i>	70
4.2.3	<i>Algoritmos de classificação aplicados</i>	72
4.3	Avaliação das abordagens de representação de textos	73
4.4	Propriedades dos parâmetros da técnica CIRT.5	82
4.5	Considerações finais	86
5	CONCLUSÕES E PERSPECTIVAS FUTURAS	87
5.1	Considerações iniciais	87
5.2	Principais contribuições desta dissertação	88
5.3	Aplicações e limitações das técnicas propostas	91
5.4	Trabalhos e pesquisas pretendidas	92
5.5	Considerações finais	92
	REFERÊNCIAS	93
	GLOSSÁRIO	105
	APÊNDICE A FONTES DE TEXTOS UTILIZADAS	107
A.1	Coleções de textos	108
A.2	Fontes externas de conhecimento	110
	APÊNDICE B MELHORES RESULTADOS OBTIDOS	113

INTRODUÇÃO

Neste capítulo são introduzidos conceitos relacionados ao problema de enriquecimento de representações de textos, que é o foco desta dissertação. Após a contextualização e definição do problema abordado, são apresentadas as motivações e lacunas identificadas no tema, seguidas das definições das hipóteses que fundamentaram o desenvolvimento desta dissertação. Por último, são descritos os principais objetivos de pesquisa definidos e as novidades da pesquisa, além da descrição da organização desta dissertação.

1.1 Contextualização e definição do problema

Para o processo de descoberta de conhecimento em grandes volumes de dados é importante utilizar representações que preservem os elementos originais da coleção de documentos analisada, já que a perda das informações originais pode prejudicar a aquisição de conhecimento. Na descoberta de conhecimento de dados textuais, por exemplo, a identificação de padrões pode ser afetada tanto pela ausência de informações semânticas quanto pela ausência de informações contextuais dos textos (HARRIS, 1954; AGGARWAL; ZHAI, 2012b).

A mineração e análise computacional de textos pode ser realizada por meio de técnicas da Mineração de Textos (MT), área de Inteligência Artificial (IA) destinada à extração de padrões de dados escritos em língua natural, como notícias, *e-mails* e *reviews* de produtos ou de serviços. Conforme abordado em Srivastava e Sahami (2009, p. xxi), o processo de mineração de grandes volumes de dados textuais pode explorar tanto aspectos estatísticos (p. ex., frequência dos termos) quanto aspectos linguísticos (p. ex., significado dos termos) dos textos.

A área de MT pode ser caracterizada como um processo para a identificação de padrões potencialmente úteis e compreensíveis embutidos em dados textuais (EBECKEN; LOPES; COSTA, 2003, p. 338), como tendências e pontos atípicos (AGGARWAL; ZHAI, 2012b, p. 2). Os tipos de padrões extraídos por meio desse processo dependem exclusivamente das tarefas a

serem realizadas, que embora compartilhem o mesmo fluxo de execução diferem sobretudo em seu objetivo, seja para classificar, para agrupar, para sumarizar ou para recuperar informações (SRIVASTAVA; SAHAMI, 2009; WEISS; INDURKHYA; ZHANG; DAMERAU, 2010).

As etapas do processo de MT são caracterizadas respectivamente pela Identificação do Problema, pelo Pré-processamento, pela Extração de Padrões, pelo Pós-processamento e pela Utilização do Conhecimento (EBECKEN; LOPES; COSTA, 2003). Embora todas essas etapas sejam importantes, o Pré-processamento é fundamental para representar a maior quantidade de informações e de correlações (relações de correspondência) entre os atributos dos textos (vide a Subseção 2.2.1). Além de tratar eventuais inconsistências linguísticas (p. ex., erros ortográficos e abreviações), a etapa de Pré-processamento também é reponsável por promover as representações de textos¹ utilizadas na Extração de Padrões.

As técnicas tradicionais de representação de textos normalmente consideram apenas informações léxicas e/ou sintáticas dos textos, a exemplo da *Bag of Words* (BoW), principal representante do Modelo de Espaço Vetorial (SEBASTIANI, 2002). Originalmente essa técnica considera somente informações léxicas para correlacionar os textos pela frequência de correlação de seus termos², considerados independentes (SALTON; WONG; YANG, 1975). Apesar de ignorar informações semânticas e contextuais, abordagens de representação de textos inspiradas na técnica BoW são efetivas em muitas aplicações baseadas em informações de tópicos ou de categorias (AGGARWAL; ZHAI, 2012b, p. 3).

Entre os problemas causados por técnicas baseadas no Modelo de Espaço Vetorial pode-se destacar a alta esparsidade (proporção de zeros) e dimensionalidade (número de atributos) das representações de textos, que prejudicam o desempenho de algoritmos de Aprendizado de Máquina (AM) devido a fatores como sobreajuste (*overfitting*) (AGGARWAL; ZHAI, 2012b, p. 3). Alternativamente, abordagens baseadas em redes permitem incorporar diferentes tipos de correlações entre os atributos dos textos, e em alguns casos chegam a superar a performance de técnicas tradicionais (JI *et al.*, 2010; ROSSI; LOPES; FALEIROS; REZENDE, 2014). Entretanto, esse tipo de abordagem geralmente é mais efetivo em domínios específicos, além de possuir um maior custo computacional em relação às técnicas tradicionais (NEWMAN, 2010).

Uma possibilidade de atenuar os problemas de alta esparsidade e dimensionalidade é o enriquecimento das representações de textos (vide a Seção 3.2). Os procedimentos de enriquecimento textual permitem explorar diferentes tipos de informações textuais, léxicas, sintáticas e/ou semânticas (vide a Subseção 2.4.1), com o objetivo de auxiliar na melhora da performance de algoritmos de AM por meio da intensificação do volume e da qualidade das informações consideradas nas representações de textos (SHEHATA, 2009; SCHULZ; RISTOSKI; PAULHEIM, 2013; NOURASHRAFEDDIN; MILIOS; ARNOLD, 2014).

¹ A representação também pode ser considerada uma etapa independente do processo de descoberta de conhecimento denominada Transformação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

² Representação denominada “matriz documento por termo” (TAN; STEINBACH; KUMAR, 2005).

1.2 Motivações e lacunas identificadas

Normalmente técnicas tradicionais de representação de textos como a BoW consideram apenas informações léxicas e/ou sintáticas dos textos, ignorando informações semânticas e contextuais que podem ser importantes para expressar correlações entre seus termos. Como as possibilidades de aplicações da MT crescem com a expansão e a variedade de documentos disponíveis em formato digital, a incorporação de informações semânticas e contextuais em representações de textos torna-se um desafio a ser superado (AGGARWAL; ZHAI, 2012b).

Conforme abordado em Ebecken, Lopes e Costa (2003, p. 340), uma das principais motivações para o enriquecimento semântico de representações de textos é a possibilidade do aumento da efetividade de algoritmos de AM em tarefas da MT baseadas nos significados dos textos. Segundo os autores, a intensificação de informações úteis associadas aos textos tende a melhorar a caracterização/representação de seu conhecimento, ressaltando a importância em se considerar aspectos semânticos (significado das palavras), pragmáticos (impacto do significado das palavras de acordo com o contexto) e do conhecimento de mundo (conhecimento geral do domínio) além dos aspectos morfológicos (estrutura da língua), sintáticos (estrutura das sentenças) e do discurso (interpretação da língua).

Em relação ao conhecimento pragmático presente nos textos, conforme destacado em Harris (1954, p. 39) os significados das palavras estão diretamente relacionados aos contextos em que elas ocorrem. Considerando contexto como “alguma informação que pode ser utilizada para caracterizar a situação de uma entidade” (DEY, 2001), contextos textuais podem ser descritos por conjuntos de palavras correlacionadas (informações contextuais) que caracterizam o conteúdo dos escopos de interesse (entidades) dos textos.

Assim como apresentado na Seção 3.3, as abordagens de representação de textos avaliadas empiricamente por meio desta dissertação foram baseadas na delimitação das sentenças dos textos analisados, consideradas como os escopos de interesse das coleções de textos analisadas. As sentenças podem ser definidas sintaticamente pela “presença de um verbo principal conjugado”, ou semanticamente pela “expressão de um pensamento completo” (OLIVEIRA, 2001, p. 99). Na prática, entretanto, essas diferenças não influenciam no entendimento deste trabalho.

Conforme apresentado na Subseção 2.3.2, a exploração de informações semânticas e contextuais tem sido amplamente abordada em trabalhos da literatura de MT, principalmente com o objetivo de enriquecer representações de textos. Entretanto, durante a revisão bibliográfica realizada foram identificados poucos trabalhos (apenas 9,2% do total) que abordam conjuntamente aspectos semânticos e contextuais para a geração de representações de textos enriquecidas, o que se caracteriza como uma das principais lacunas identificadas na literatura. Também foram identificados poucos trabalhos (apenas 4% do total) que abordam aspectos semânticos e/ou contextuais de coleções de textos do idioma português, representando mais uma das lacunas identificadas e exploradas nesta dissertação (SINOARA; ANTUNES; REZENDE, 2017).

Em Sinoara, Antunes e Rezende (2017) a utilização de fontes externas de conhecimento (p. ex., Wikipédia³ e WordNet⁴) é retratada como uma das abordagens mais utilizadas para o enriquecimento de representações de textos. Segundo os autores, entre as principais aplicações de fontes externas de conhecimento destacam-se as tarefas de Reconhecimento de Entidades Nomeadas (REN) e de Desambiguação Lexical de Sentido (DLS) da área de Processamento de Língua Natural (PLN).

Enquanto a tarefa de REN visa identificar entidades nomeadas previamente conhecidas (p. ex., pessoas e localizações) associadas aos termos dos textos, a tarefa de DLS (do inglês *Word Sense Disambiguation*) visa determinar o sentido mais adequado de uma palavra em relação ao contexto em que ela ocorre (JURAFSKY; MARTIN, 2009). No exemplo a seguir, baseado em um texto formado por um único parágrafo com duas sentenças, são destacadas as principais diferenças entre textos sem enriquecimento textual e textos semanticamente enriquecidos por meio das tarefas de REN e de DLS.

1. **Texto original sem enriquecimento** – Extração de termos simples:

- *O Laboratório de Inteligência Computacional foi fundado em 1995.*
- *É parte do Departamento de Ciência da Computação do Instituto de Ciências Matemáticas e de Computação da **Universidade de São Paulo**, campus de São Carlos.*

2. **Texto semanticamente enriquecido (REN)** – Extração de termos simples e compostos:

- *Laboratório Inteligência_Computacional foi fundado*
- *parte Departamento_de_Ciência_da_Computação Instituto Ciências_Matemáticas Computação **Universidade_de_São_Paulo** campus São_Carlos*

3. **Texto semanticamente enriquecido (DLS)** – Extração de termos desambiguados:

- *laboratório Inteligência_computacional ser estabelecer*
- *componente departamento_de_ciência_da_computação instituto ciência_matemática ciência_da_computação **Universidade_de_São_Paulo** campus São_Carlos*

No primeiro caso, os termos são extraídos do texto considerando apenas sua separação por espaços em branco ou quebras de linha, o que ignora a ocorrência de possíveis termos compostos como “Universidade de São Paulo”. No segundo caso, os termos são extraídos do texto considerando a ocorrência de conceitos ou de entidades nomeadas previamente conhecidas, o que permite a identificação de termos compostos como “Universidade_de_São_Paulo”. Já no último caso, os termos são extraídos do texto considerando o contexto em que ocorrem, o que também permite a identificação de termos compostos como no caso anterior.

³ <<http://wikipedia.org>>

⁴ <<https://wordnet.princeton.edu>>

Embora no exemplo anterior as tarefas de REN e de DLS sejam retratadas como uma solução efetiva para o enriquecimento textual, existem outras estratégias computacionalmente menos custosas que também poder ser utilizadas para essa finalidade. Os procedimentos de enriquecimento textual são abordados mais detalhadamente na Seção 3.2.

As abordagens de enriquecimento textual também podem ser baseadas em modelos de linguagem, utilizados para explorar correlações estatísticas entre os termos dos textos já que “a noção de um modelo de linguagem é inerentemente probabilística” (MANNING; RAGHAVAN; SCHÜTZE, 2008, p. 238). Portanto, “um modelo de linguagem é uma função que coloca uma medida de probabilidade sobre *strings* [termos] extraídas [extraídos] de algum vocabulário”. Os modelos de linguagem auxiliam na predição probabilística dos termos dos textos, favorecendo a identificação de termos correlacionados e consequentemente na redução da esparsidade e da dimensionalidade das representações de textos (JURAFSKY; MARTIN, 2009).

Como tradicionalmente os atributos das representações de textos são considerados independentes, até pelo fato de normalmente serem caracterizados por um único termo, informações compartilhadas entre conjuntos de termos relacionados acabam sendo ignoradas. Tendo em vista de que a utilização de modelos de linguagem favorece a descoberta de conjuntos de termos correlacionados, aparentemente seria possível gerar representações de textos cujos atributos são caracterizados por conjuntos de termos. Essa alternativa permitiria levar em consideração tanto a frequência literal dos atributos nos textos, quanto a frequência relativa entre atributos similares.

Além de modelos de linguagem tradicionais como os *N-grams*, que são baseados na ordem direta dos termos nos textos, na área de MT também têm sido frequentemente aplicados modelos baseados em Rede Neural Artificial (RNA) para a representação distribuída de termos (p. ex., *word embeddings*) (SINOARA; ANTUNES; REZENDE, 2017). Considerada uma das motivações desta dissertação, modelos baseados em *word embeddings* permitem explorar correlações não lineares entre os textos por meio de seus termos, considerando ainda a influência de textos de fontes externas de conhecimento como a Wikipédia (MIKOLOV; CHEN; CORRADO; DEAN, 2013; MIKOLOV; LE; SUTSKEVER, 2013; MIKOLOV *et al.*, 2013).

Conforme abordado em Mesnil, He, Deng e Bengio (2013), as abordagens de representação de textos baseadas em *word embeddings* também podem ser exploradas por meio de técnicas de *deep learning*, caracterizadas pela alta densidade no treinamento dos modelos tanto pelo volume de informações processadas quanto pela complexidade de configuração dos parâmetros utilizados. Essa possibilidade de refinamento também é considerada uma das motivações em se utilizar modelos baseados em *word embeddings* nesta dissertação, já que possibilita uma exploração mais detalhada das coleções de textos em domínios específicos de aplicação.

Na seção a seguir são apresentadas as hipóteses que fundamentaram o desenvolvimento desta dissertação. Essas hipóteses assumem o enriquecimento semântico e contextual das representações de textos como uma alternativa efetiva que contribui para uma melhora na performance de algoritmos de AM na tarefa de classificação automática (não manual) de textos.

1.3 Hipóteses definidas

Considerando o problema de enriquecimento de representações de textos, foram definidas duas hipóteses de pesquisa, apresentadas a seguir, baseadas nas motivações e lacunas identificadas na literatura. O foco desta dissertação é o estudo e a avaliação de informações semânticas e contextuais, exploradas por meio das tarefas de REN e de DLS, para o enriquecimento de representações de textos utilizando modelos de linguagem baseados em *word embeddings*.

- **Hipótese 1** – As representações de textos baseadas em atributos caracterizados por contextos (conjuntos de termos correlacionados) contribuem para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos;
- **Hipótese 2** – O enriquecimento semântico dos atributos das representações de textos contribui para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos.

Baseando-se nessas duas hipóteses definidas, o objetivo geral desta dissertação é contribuir para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos. Na seção a seguir são apresentados os objetivos específicos relacionados à teoria, à prática e às aplicações de abordagens de enriquecimento de representações de textos, definidos com o intuito de verificar as hipóteses de pesquisa apresentadas.

1.4 Principais objetivos e novidades da pesquisa

Em consequência das hipóteses definidas, nesta dissertação é apresentada uma nova técnica de representação de textos. Além da abordagem de enriquecimento semântico e contextual de representações de textos, a principal novidade dessa técnica proposta é a abordagem inédita utilizada para calcular a frequência dos atributos (contextos), considerados dependentes, que se baseia na similaridade entre os contextos por meio da intersecção entre seus termos.

Para verificar a **Hipótese 1**, a técnica de representação de textos proposta foi comparada com soluções existentes na literatura por meio da tarefa de classificação automática de textos. Para verificar a **Hipótese 2**, as performances das abordagens de representação de textos avaliadas foram comparadas considerando diferentes versões das bases de dados utilizadas: textos com termos simples, textos com termos compostos e textos com termos desambiguados. Como consequência dessas verificações foram definidos os seguintes **objetivos primários**, que enfatizam os diferentes tipos de contribuições pretendidas nesta dissertação:

- **Objetivo 1 (teórico)** – Contribuir para a fundamentação teórica do uso de fontes externas de conhecimento para o enriquecimento semântico e contextual de representações de textos, analisando-o por meio da tarefa de classificação automática de textos;

- **Objetivo 2 (prático)** – Propor, desenvolver e avaliar técnicas não supervisionadas de representação de textos que exploram informações semânticas e contextuais, analisando-as por meio da tarefa de classificação automática de textos;
- **Objetivo 3 (aplicacional)** – Aplicar as técnicas desenvolvidas em diferentes cenários de representação de textos, explorando possíveis variações de performance entre os idiomas português e inglês em diferentes bases de dados e comparando os resultados obtidos com soluções existentes na literatura.

Baseado nos objetivos primários definidos, consequentemente também foram definidos **objetivos secundários** para o desenvolvimento desta dissertação. Esses objetivos estão relacionados à implementação e disponibilização de ferramentas ou *scripts* para auxiliar na aplicação de procedimentos de enriquecimento textual e na geração de representações de textos, principalmente em relação às soluções propostas. Nesse sentido, recomenda-se a utilização de repositórios de controle de versão de arquivos como o GitHub⁵, ou até mesmo por meio do *site* do Laboratório de Inteligência Computacional (LABIC)⁶. O laboratório LABIC, cujo autor desta dissertação atualmente integra, faz parte do Departamento de Ciência da Computação do Instituto de Ciências Matemáticas e de Computação (ICMC).

Na seção a seguir é descrita a organização dos assuntos discutidos nesta dissertação, onde são apresentados desde a fundamentação teórica da área de MT, enfatizando a tarefa de classificação automática de textos, até a análise de aspectos textuais semânticos relacionados ao enriquecimento de representações de textos.

1.5 Organização desta dissertação

Neste primeiro capítulo são apresentadas a contextualização e a definição do problema de enriquecimento de representações de textos, seguidas das motivações e lacunas identificadas na literatura, das hipóteses e dos principais objetivos definidos nesta dissertação. No Capítulo 2 são abordados os conceitos relacionados à mineração e análise de textos, enfatizando os trabalhos relacionados ao estudo da semântica na MT e à tarefa de classificação automática de textos. No Capítulo 3 são apresentadas as diferentes abordagens de enriquecimento textual existentes na literatura e apresentada a nova técnica de representação de textos proposta. No Capítulo 4 são descritos os experimentos realizados seguidos da análise dos resultados obtidos. Por último, no Capítulo 5 são destacadas as principais conclusões obtidas por meio desta dissertação, incluindo as implicações dos resultados obtidos para a área de MT, as projeções e as pesquisas futuras. No Apêndice A são descritas as fontes de dados utilizadas nos experimentos, enquanto no Apêndice B são apresentados os melhores resultados obtidos.

---

⁵ <<https://github.com>>

⁶ <<http://labic.icmc.usp.br>>

MINERAÇÃO E ANÁLISE DE TEXTOS

Neste capítulo são apresentados conceitos importantes associados ao estudo da semântica para a mineração e análise de documentos de texto, abordando os fundamentos da área de MT com foco em representações de textos. Também são descritos os diferentes níveis de complexidade semântica e o impacto de cada nível no desempenho da tarefa de classificação automática de textos, além de serem apresentados os principais trabalhos relacionados ao tema. Por último são analisados os aspectos linguísticos que envolvem as técnicas de representação de textos e suas diferentes abordagens de correlação entre os atributos, como os tipos de informações textuais e seus relacionamentos.

2.1 Considerações iniciais

A comunicação é um processo fundamental para a humanidade e depende de uma série de condições para ser bem sucedida, como a definição de uma língua comum formada por um alfabeto e um conjunto de regras para a escrita, a leitura e a fala; tais regras definem entre outros fatores os relacionamentos entre as palavras (SAUSSURE, 1966, p. 7–17). Por meio de correlações semânticas entre as palavras torna-se possível identificar sutilezas na comunicação que facilitam seu entendimento, como a aparição de um conjunto de palavras já conhecidas e seu provável significado em contextos similares (BAKER, 2001, p. 351).

Análogo ao processo natural de interpretação de textos, o processo de MT também se baseia na descoberta de conhecimento a partir de atalhos de comunicação formados por palavras-chave, assim como tratado nas seções a seguir. Em razão do escopo desta dissertação ser delimitado pelo enriquecimento de representações de textos, são enfatizadas algumas das principais técnicas de representação de textos avaliadas por meio da tarefa de classificação automática de textos. Também são apresentados e discutidos os diferentes níveis de complexidade semântica e alguns dos principais trabalhos relacionados ao tema da semântica na MT.

2.2 Mineração de Textos

Devido à popularização da *Internet* e ao desenvolvimento da computação desde a década de 1990, neste século houve um aumento substancial no volume de dados digitais armazenados e disponíveis em todo o mundo (FUCHS; BOERSMA; ALBRECHTSLUND; SANDOVAL, 2013; Cisco Systems, 2017). Conforme abordado em Fuchs, Boersma, Albrechtslund e Sandoval (2013), embora essa evolução possa ser interpretada como avanço tecnológico, a análise manual dos dados torna-se inviável devido a uma série de fatores como a diversidade, a volatilidade e o gigantesco volume dos dados, em sua maioria não estruturados. Nesse sentido a MT caracteriza-se como uma importante área de estudo na análise e organização automática de textos (EBECKEN; LOPES; COSTA, 2003, p. 337–338).

A área de MT pode ser definida como um processo específico da área de Mineração de Dados, destinado à identificação de padrões potencialmente úteis e compreensíveis embutidos em dados textuais (HEARST, 1999; EBECKEN; LOPES; COSTA, 2003; KAO; POTEET, 2007). Por meio do processo de MT torna-se possível descobrir e analisar padrões interessantes em dados não estruturados do tipo texto, incluindo tendências e pontos atípicos (AGGARWAL; ZHAI, 2012b, p. 2). Como a descoberta de padrões interessantes depende principalmente do objetivo da análise a ser realizada, é essencial a escolha de algoritmos de AM que sejam condizentes com a tarefa a ser executada, como o agrupamento, a classificação e a sumarização (EBECKEN; LOPES; COSTA, 2003, p. 351). Na subseção a seguir são descritas detalhadamente as etapas do processo tradicional de MT, que vão desde a identificação do problema a ser resolvido até a utilização do conhecimento adquirido para solucioná-lo.

2.2.1 Etapas do processo tradicional de Mineração de Textos

Assim como o processo da área de Mineração de Dados, o processo tradicional de MT pode ser dividido em cinco etapas complementares (EBECKEN; LOPES; COSTA, 2003). Conforme ilustrado no topo da Figura 1, na etapa inicial de **Identificação do Problema** são definidos os objetivos que irão guiar todo o processo de mineração em razão da meta do problema traçado. No **Pós-processamento** os padrões extraídos na etapa de **Extração de Padrões** são analisados para apoiar a tomada de decisões na etapa de **Utilização do Conhecimento**.

O foco desta dissertação reside na etapa de **Pré-processamento**, que é fundamental para todo o processo de MT já que é responsável por representar computacionalmente os textos analisados e suas características. Por meio das representações de textos busca-se considerar o maior número de correlações entre os termos dos textos para viabilizar a identificação de padrões úteis, embora as técnicas tradicionais de representação de textos normalmente abordem apenas informações léxicas e/ou sintáticas dos textos, implicando na perda de informações semânticas e contextuais que possivelmente afetam a qualidade dos padrões extraídos (SRIVASTAVA; SAHAMI, 2009, p. 11–13).

Figura 1 – Exemplo do processo tradicional da etapa de Pré-processamento.

Fonte: Elaborada pelo autor.

A etapa de Pré-processamento do processo de MT tem o objetivo principal de permitir o tratamento dos textos da coleção a ser analisada, permeando desde a correção de inconsistências linguísticas até a formatação dos textos por meio de estruturas de dados eficientes que permitam extrair padrões interessantes. Tradicionalmente o Pré-processamento pode ser dividido nas seguintes fases complementares: Preparação de Textos, Extração de Termos, Atribuição de Pesos, Seleção de Termos e Representação de Textos (EBECKEN; LOPES; COSTA, 2003).

Na fase de **Preparação de Textos** são tratadas eventuais inconsistências linguísticas que possam dificultar a identificação de correlações recorrentes entre os termos durante a etapa de Extração de Padrões (p. ex., erros de digitação, abreviações, acentuações, uso de gírias e de caracteres especiais). Durante essa fase normalmente são removidos termos considerados irrelevantes, como conectivos e preposições, por meio de listas de *stopwords* denominadas *stoplists*. Também é frequente a realização de padronizações da caixa dos termos, geralmente convertidos ou para minúsculas ou para maiúsculas para facilitar a identificação de padrões nas coleções de textos.

Na fase de **Extração de Termos** o objetivo é identificar os atributos que melhor descrevem uma coleção de textos. Para alcançar esse objetivo podem ser aplicadas técnicas de enriquecimento textual para diminuir a diversidade do vocabulário e possivelmente intensificar correlações entre os termos dos textos (vide a Seção 3.2). Embora essas abordagens contribuam para diminuir a esparsidade e a dimensionalidade das abordagens de representação de textos, enriquecimentos como simplificações podem distorcer o conhecimento presente nos textos e consequentemente prejudicar a qualidade dos padrões extraídos (WEISS; INDURKHYA; ZHANG; DAMERAU, 2010, p. 15–46).

A fase de **Atribuição de Pesos** torna possível estimar a expressividade relativa dos termos considerando tanto sua representatividade na coleção analisada quanto os tamanhos dos textos. A expressividade dos termos em cada texto pode ser transmitida por meio de métricas como a *Term Frequency* (TF) (utilizada para calcular a frequência relativa dos termos nos textos), a *Inverse Document Frequency* (IDF) (utilizada para estimar a relevância dos termos como o inverso de sua ocorrência nos textos) ou a *Term Frequency - Inverse Document Frequency* (TF-IDF) (utilizada para estimar a importância relativa de um termo por meio da combinação de sua frequência relativa e relevância nos textos) (WEISS; INDURKHYA; ZHANG; DAMERAU, 2010, p. 25–32).

Supondo que diferentes textos possuem tamanhos diferentes, é importante normalizar o cálculo da métrica TF com base nos tamanhos dos textos em que um termo ocorre, já que a frequência literal dos termos pode enviesar a representatividade das informações dos textos (WEISS; INDURKHYA; ZHANG; DAMERAU, 2010, p. 30). Assim, dado D como a quantidade total de documentos da coleção analisada, e $frequencia(t, d)$, $termos(d)$ e $ocorrencia(t)$ como funções que calculam respectivamente a frequência literal de um termo t em um documento d , a quantidade de termos do documento d e a quantidade de documentos em que ocorre o termo t , as métricas de atribuição de pesos podem ser expressas pelas seguintes equações:

$$TF(t, d) = \frac{frequencia(t, d)}{termos(d)} \quad IDF(t, d) = \ln \left(\frac{D}{ocorrencia(t)} \right) \quad (2.1)$$

$$TF - IDF(t, d) = \frac{frequencia(t, d)}{termos(d)} * \ln \left(\frac{D}{ocorrencia(t)} \right) = TF(t, d) * IDF(t, d) \quad (2.2)$$

Embora ambas as métricas TF, IDF e TF-IDF permitam exprimir a representatividade dos termos dos textos em diferentes cenários, a métrica TF-IDF é mais frequentemente utilizada por transmitir a relevância dos termos tanto nos textos em que ocorrem quanto em toda a coleção analisada (JONES, 1972; JONES, 1973; BEEL; GIPP; LANGER; BREITINGER, 2016). Na métrica TF-IDF os termos que ocorrem mais vezes em poucos textos são consideradas mais significativos do que os termos que ocorrem mais vezes em muitos textos da coleção analisada (LESKOVEC; RAJARAMAN; ULLMAN, 2014, p. 8–9).

A fase de **Seleção de Termos** caracteriza-se pela restrição de termos considerados pouco relevantes para descrever a coleção analisada. Baseado na Lei de Zipf, que afirma que uma pequena porcentagem de termos (menos de 20% do total) é suficiente para expressar o conhecimento contido em textos relativamente grandes (ZIPF, 1949), o conjunto de termos extraídos normalmente é ranqueado decrescentemente de acordo com os seus pesos atribuídos. Esse ranqueamento torna possível a aplicação dos Cortes de Luhn, permitindo a remoção de termos considerados muito comuns ou que aparecem raramente nos textos (LUHN, 1958).

Conforme abordado em Crain, Zhou, Yang e Zha (2012), a aplicação de técnicas de redução de dimensionalidade também é importante na fase de Seleção de Termos, pois auxiliam efetivamente no aumento da performance dos algoritmos de AM. Técnicas como a *Latent Semantic Analysis* (LSA) (também conhecida como *Latent Semantic Indexing* (LSI)) (LANDAUER; FOLTZ; LAHAM, 1998) e a *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003), por exemplo, auxiliam na redução de dimensionalidade exprimindo conjuntos de termos semanticamente correlacionados em conceitos específicos representados por meio de tópicos.

A fase de **Representação de Textos** é realizada após todos os tratamentos textuais realizados na coleção analisada. Como o tratamento de textos é importante para atenuar possíveis inconsistências linguísticas e intensificar as correlações entre os termos, o objetivo das técnicas de representação de textos pode ser caracterizado como a representação das correlações mais relevantes entre os textos por meio de seus termos. Na Seção 2.4 são apresentados mais detalhadamente os procedimentos de algumas das principais técnicas de representação de textos presentes na literatura e utilizadas nos experimentos apresentados nesta dissertação.

Assim como ilustrado no topo da Figura 1, após a etapa de Pré-processamento é realizada a etapa de Extração de Padrões, que visa permitir a aplicação de algoritmos de AM para a descoberta de padrões potencialmente úteis contidos nas representações de textos geradas. Nesta dissertação foi escolhida a tarefa de classificação automática de textos, uma das aplicações mais frequentemente consideradas para a avaliação de estudos envolvendo aspectos textuais semânticos na MT (vide a Subsubseção 2.3.2.2). Na subseção a seguir são tratados mais detalhadamente conceitos importantes associados à tarefa de classificação automática de textos, incluindo a descrição de alguns dos principais algoritmos de classificação utilizados na literatura.

2.2.2 Tarefa de classificação automática de textos

A classificação é uma das tarefas de predição mais amplamente realizadas no AM, e tem como objetivo geral descobrir as classes, dentre um conjunto pré-definido de rótulos discretos, de exemplos não rotulados. A tarefa de classificação é realizada por meio do aprendizado indutivo, em que são criados modelos de conhecimento com base na generalização de conceitos específicos dos exemplos. Portanto, trata-se de um aprendizado supervisionado que busca reproduzir o conhecimento contido em um conjunto de exemplos para aplicá-lo na organização ou predição de novos exemplos (MONARD; BARANAUSKAS, 2003, p. 91).

Embora a tarefa de classificação possa ser aplicada a quaisquer tipos de dados, na área de MT os exemplos analisados são os textos pertencentes à uma coleção específica que se pretende minerar (p. ex., notícias, *e-mails* ou *reviews*). As correlações entre os termos das coleções de textos podem ser representadas por meio de dados binários ou numéricos, e os textos podem ser atribuídos à uma única classe (conhecida como classificação simples-rótulo ou versão *hard*) ou à múltiplas classes (conhecida como classificação multirrótulo ou versão *soft*) (FACELI; LORENA; GAMA; CARVALHO, 2011; AGGARWAL; ZHAI, 2012a).

Na subseção a seguir são apresentados alguns dos principais algoritmos utilizados na literatura da área de MT para a tarefa de classificação automática de textos, discutidos em detrimento dos paradigmas de AM empregados por cada algoritmo. Em seguida são apresentadas as métricas de avaliação de desempenho mais frequentemente empregadas na tarefa de classificação, discutidas em detrimento dos aspectos de performance analisados.

2.2.2.1 Paradigmas de Aprendizado de Máquina e algoritmos de classificação

A tarefa de classificação automática de textos pode adotar diferentes estratégias de resolução do problema. Essas abordagens de classificação descrevem um conjunto de paradigmas de AM (baseado em exemplos, conexionista, simbólico e evolutivo) que permitem associar características específicas de textos com uma estratégia de resolução dados os tipos de resultados buscados e a compreensibilidade das soluções obtidas, entre outros fatores (MONARD; BARANAUSKAS, 2003, p. 92–94). Levando em consideração os diferentes paradigmas de AM, os algoritmos de classificação podem ser organizados nas seguintes classes de métodos: métodos baseados em distância, métodos baseados em otimização, métodos probabilísticos e métodos baseados em procura (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 58–136).

O *K - Nearest Neighbors* (KNN) (ALTMAN, 1992) pode ser ressaltado como um dos algoritmos mais representativos dos **métodos baseados em distância**. O algoritmo KNN depende do número *K* de vizinhos mais próximos e de uma medida de similaridade para estimar a classe de um novo exemplo (p. ex., distância euclidiana e cosseno), definida pelo voto majoritário dado pelos exemplos já conhecidos e previamente rotulados (AGGARWAL; ZHAI, 2012a, p. 200–203). Esse classificador destaca-se pela sua simplicidade e flexibilidade, embora suas principais desvantagens sejam a sensibilidade à redundância de atributos ou à presença de atributos irrelevantes, além do alto custo computacional em representações de alta dimensionalidade (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 63–64).

Entre as abordagens de classificação mais representativas dos **métodos baseados em otimização** destacam-se as técnicas baseadas em RNAs¹ (p. ex., *Multi-layer Perceptron* (AGGARWAL; ZHAI, 2012a)) e o classificador *Support Vector Machines* (SVM) (CORTES; VAPNIK, 1995). Embora o uso de RNAs seja efetivo em tarefas de classificação automática de textos, normalmente os ganhos de performance são insuficientes para justificar as maiores complexidades computacionais associadas à essas abordagens (AGGARWAL; ZHAI, 2012a, p. 197–199), juntamente com a incompreensibilidade dos modelos gerados e de suas soluções apresentadas (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 121). Entretanto, o classificador SVM pode ser designado como uma opção computacionalmente menos custosa, mas que também apresenta uma boa capacidade de generalização baseada no conhecimento aprendido (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 135).

¹ Alguns autores consideram as técnicas de classificação baseadas em RNAs como métodos estatísticos paramétricos (MONARD; BARANAUSKAS, 2003, p. 93).

O objetivo do classificador SVM é encontrar o hiperplano de separação ideal que maximiza a margem da base de treinamento dos exemplos analisados, de modo a separar os exemplos em duas classes distintas (VAPNIK; CHERVONENKIS, 1971). Embora a performance do SVM seja efetiva em grandes conjuntos de exemplos, permitindo uma classificação relativamente mais rápida do que os modelos conexionistas baseados em RNAs, entre as suas principais desvantagens destacam-se a sensibilidade a ruídos e o elevado tempo dispendido no treinamento de modelos com alta dimensionalidade, bem como a dificuldade de interpretação das soluções geradas (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 135).

Ainda como exemplo dos métodos baseados em otimização, é possível citar o *Inductive Model based on Bipartite Heterogeneous Networks* (IMBHN), um algoritmo baseado em redes bipartidas que permite correlacionar termos e textos (ROSSI; LOPES; FALEIROS; REZENDE, 2014). Conforme abordado em Rossi (2015), o modelo de classificação do algoritmo IMBHN é treinado a partir da indução dos pesos dos termos para as classes dos textos, e pode ser aplicado por meio das seguintes versões de implementação: IMBHN^C, que “somente altera as informações de classe dos termos se as informações de classe correntes produzirem um erro de classificação” e IMBHN^R, que “realiza uma regressão para induzir as informações de classe dos termos” (ROSSI, 2015, p. 94). O autor ainda ressalta que ambas as versões do algoritmo IMBHN são capazes de produzir resultados significativamente superiores aos obtidos por algoritmos tradicionais como o SVM e o KNN, além de apresentar soluções mais compreensíveis (ROSSI, 2015, p. 67–114).

Entre os algoritmos mais representativos dos **métodos probabilísticos** destaca-se o *Naïve Bayes* (NB), que utiliza probabilidades condicionais para classificar novos exemplos baseando-se em probabilidades de ocorrência conhecidas de determinados atributos. Extensamente abordado em trabalhos das áreas de Matemática, de Estatística e de Computação, o Teorema de Bayes permite descrever a probabilidade de acontecimento de um novo evento dado a probabilidade de acontecimento de eventos já ocorridos (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 70–82). Embora o classificador NB considere os atributos como sendo condicionalmente independentes, ele não é sensível a ruídos e características irrelevantes, além de permitir treinamentos e classificações rápidas e apresentar soluções compreensíveis (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 76).

Nos **métodos baseados em procura** a classificação de novos exemplos é efetuada com base em uma organização de exemplos já conhecidos por meio de estruturas simbólicas (p. ex., árvores de decisão), utilizando critérios de particionamento baseados em entropia ou índice de Gini, por exemplo (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 83–106). A predição das classes é realizada por meio de buscas e de comparações com os atributos da estrutura simbólica, tendo o C4.5² como um dos algoritmos mais difundidos na literatura. Entre as vantagens do C4.5 destacam-se a compreensibilidade das soluções obtidas, a flexibilidade e robustez no tratamento

² O C4.5 é uma derivação do algoritmo *Iterative Dichotomiser 3*, um exemplo de classificador tradicional baseado em árvores de decisão (QUINLAN, 1993).

dos dados e a eficiência no tempo de execução. Suas principais desvantagens são a ineficiência em tratar valores ausentes e atributos contínuos, e a instabilidade causada por pequenas variações no conjunto de treinamento (FACELI; LORENA; GAMA; CARVALHO, 2011, p. 95–97).

2.2.2.2 Avaliação de desempenho da tarefa de classificação

Considerando que a generalização de conceitos textuais normalmente está sujeita à ausência de informações linguísticas nas representações de textos, por exemplo, os algoritmos de AM empregados nas tarefas de MT consequentemente estão sujeitos à eventuais erros ou inconsistências relacionadas à seus desempenhos. Na tarefa de classificação a avaliação do desempenho dos algoritmos de AM normalmente é realizada sob diferentes aspectos de performance como a acurácia, a precisão e a revocação dos resultados dos classificadores (EBECKEN; LOPES; COSTA, 2003; MANNING; RAGHAVAN; SCHÜTZE, 2008).

Para exemplificar o fluxo de execução da tarefa de classificação, considera-se uma coleção de exemplos a serem atribuídos à uma classe arbitrária ϕ . Diante dessa tarefa denominam-se **exemplos recuperados** aqueles atribuídos como sendo da classe ϕ e **exemplos relevantes** aqueles cuja classe é sabidamente ϕ . Assim como apresentado na lista a seguir, os exemplos da coleção analisada podem ser mais precisamente distinguidos em função da assertividade de sua classe atribuída, como por meio dos conjuntos de exemplos que sabidamente são da classe ϕ ou que não pertencem à ela.

- **Verdadeiro Positivo (VP)** – Exemplo corretamente atribuído ao rótulo ϕ ;
- **Falso Negativo (FN)** – Exemplo erroneamente não atribuído ao rótulo ϕ ;
- **Falso Positivo (FP)** – Exemplo erroneamente atribuído ao rótulo ϕ ;
- **Verdadeiro Negativo (VN)** – Exemplo corretamente não atribuído ao rótulo ϕ .

Conforme ilustrado na Figura 2, a métrica **Acurácia** permite avaliar a relação dos exemplos classificados corretamente por um algoritmo, medindo a efetividade dos classificadores. Já as métricas **Precisão** e **Revocação** estão relacionadas à análise de relevância das classificações, medindo respectivamente a utilidade (probabilidade de que exemplos relevantes sejam recuperados) e a completude ou sensibilidade (probabilidade de que exemplos recuperados sejam relevantes) dos classificadores (MANNING; RAGHAVAN; SCHÜTZE, 2008, p. 362).

Para correlacionar as métricas Acurácia e Precisão de forma equilibrada é utilizada a métrica **F1**, também conhecida como *F-measure* ou *F-score*. Assim como ressaltado em Sokolova e Lapalme (2009), há duas maneiras complementares de calcular a razão F1: média micro e média macro. Enquanto na **média micro** é calculada a soma dos termos (VP, FP e FN) individuais de cada uma das métricas Precisão e Revocação, na **média macro** é calculada a média dos valores de Precisão e de Revocação considerando todas as classes da coleção analisada.

Segundo os autores, enquanto a estratégia média macro trata de forma igualitária todas as classes a estratégia média micro tende a favorecer a classe majoritária da coleção.

Figura 2 – Métricas de avaliação de desempenho da tarefa de classificação.

Fonte: Elaborada pelo autor.

Na seção a seguir são discutidos aspectos linguísticos importantes, com ênfase na semântica, para auxiliar na potencialização da performance dos algoritmos de AM em tarefas da área de MT. Inicialmente são apresentados os diferentes tipos de complexidade semântica na tarefa de classificação automática de textos e retratados os diferentes modos de se incluir esse tipo de informação nas representações de textos. Também são apresentados alguns dos principais trabalhos relacionados ao tema “estudo da semântica na área de MT”, correlacionados por meio de uma meta-análise da literatura.

2.3 Análise de aspectos textuais semânticos

Considerando as diferenças de cada idioma, toda língua é regida por um conjunto de normas gramaticais que definem como deve ser realizada a comunicação entre dois ou mais indivíduos de maneira padronizada (CHOMSKY, 1959; SAUSSURE, 1966; ECKERSLEY; MACAULAY, 1987). Conceitualmente essas regras permitem definir três diferentes tipos de gramática: as **gramáticas normativas**, formadas por regras que devem ser seguidas; as **gramáticas descritivas**, formadas por regras que são seguidas; e as **gramáticas internalizadas**, formadas por regras que os falantes da língua dominam. A aplicação conjunta dessas definições gramaticais possibilita analisar os diferentes aspectos das línguas tanto pela norma culta quanto de suas variações informais (POSSENTI, 1996, p. 63–73).

Tradicionalmente o estudo da gramática é dividido em três partes básicas – a fonética (estudo dos sons da fala), a morfologia (estudo da estrutura das palavras) e a sintaxe (estudo da estrutura das frases) – e duas complementares: a semântica (estudo do significado das palavras) e a estilística (estudo dos estilos de escrita) (BECHARA, 2012). Como o foco desta dissertação compreende a exploração de informações semânticas e contextuais para o enriquecimento de

representações de textos, é enfatizado o estudo do significado das palavras, ou mais precisamente o estudo de aspectos textuais semânticos das coleções de textos analisadas.

Conforme ressaltado em Cançado (2008, p. 14–16), assim como as demais partes da gramática o estudo da semântica está relacionado às habilidades linguísticas dos indivíduos baseadas em conhecimentos específicos sobre a língua enquanto ferramenta, e sobre a linguagem enquanto uso da língua. Sob o ponto de vista teórico, os tipos de abordagens utilizados na investigação do significado estudado pela semântica podem ser organizados nas seguintes vertentes: abordagem referencial, abordagem mentalista e abordagem pragmática.

Na **abordagem referencial** são tratados fenômenos que lidam com a referência do mundo e as palavras, como as implicações lógicas, as relações entre sentenças, as ambiguidades, as noções de sentido e as referências. Já na **abordagem mentalista** são tratados fenômenos que propõem que o sentido também acontece em um nível intermediário entre o mundo e as palavras, como os protótipos, as metáforas e os papéis temáticos, também denominados papéis semânticos. Por fim, na **abordagem pragmática** são tratados fenômenos que estudam os usos situados da língua, como as implicaturas conversacionais e os atos de fala (CANÇADO, 2008).

Em razão da ausência consensual sobre a definição de significado, a semântica tem sido estudada em diversas perspectivas por diferentes teorias linguísticas. Em relação à abordagem referencial a principal corrente teórica é a **semântica formal**, focada em aspectos linguísticos associados à veracidade do significado. Como representante da abordagem mentalista pode-se citar a **semântica lexical**, focada em aspectos linguísticos associados à ligação entre estruturas sintáticas e semânticas. Já na abordagem pragmática pode-se citar a **semântica argumentativa**, focada em aspectos linguísticos associados à argumentação independente da veracidade do significado (CANÇADO, 2008, p. 139–145).

Embora o estudo de aspectos textuais semânticos por meio de abordagens computacionais normalmente esteja associado à corrente teórica da semântica lexical, principalmente em relação aos trabalhos relacionados à MT, por motivo de convenção a terminologia “semântica” e suas variações são utilizadas sem distinção nesta dissertação. Os tipos de informações e de relacionamentos textuais, incluindo as relações semânticas entre palavras, são abordados detalhadamente na Subseção 2.4.1.

Sob a perspectiva distribucional, as palavras e suas classes funcionais podem ser distribuídas como **associações sintagmáticas** (eixo de relações de coocorrência entre as palavras) ou como **paralelos paradigmáticos** (eixo de relações que não coocorrem entre as palavras) (SCHÜTZ; PEDERSEN, 1993). Na sentença “*O cachorro está latindo*”, por exemplo, “*latindo*” estabelece uma relação sonora explícita com o objeto “*cachorro*” (associação sintagmática) pois o verbo “*latir*” está contido no texto. Entretanto, o objeto direto da sentença também poderia estar “*correndo*” ou “*dormindo*” (paralelos paradigmáticos), pois essas ações correspondem às possíveis atividades de “*cachorro*”.

Para a descoberta de padrões potencialmente úteis em coleções de textos é importante a escolha de técnicas de representação de textos que permitam considerar associações sintagmáticas e/ou paralelos paradigmáticos, já que normalmente as técnicas de representação de textos são embasadas em hipóteses que afirmam que padrões estatísticos podem ser mapeados para expressar os significados de textos. Conforme listado a seguir, essa afirmação combina diferentes hipóteses de estruturação linguística que permitem interpretar uma coleção de textos como uma distribuição de características específicas baseadas em modelos composicionais distribuídos (TURNERY; PANTEL, 2010, p. 146).

- **Hipótese distribucional** – Os termos que coocorrem em contextos similares tendem a possuir significados similares (HARRIS, 1954);
- **Hipótese de *bag of words*** – As frequências dos termos dos textos tendem a indicar algum aspecto de seu significado (SALTON; WONG; YANG, 1975);
- **Hipótese distribucional estendida** – Os padrões que coocorrem com pares de termos similares tendem a possuir significados similares (LIN; PANTEL, 2001);
- **Hipótese de relações latentes** – Os pares de termos que coocorrem em padrões similares tendem a possuir relações semânticas (TURNERY, 2008).

Assim como abordado em Kartsaklis (2014, p. 9–26), os modelos composicionais distribuídos podem ser categorizados a partir dos tipos de informações e de relacionamentos considerados nos modelos de linguagem, que podem ser lineares, multilineares ou não lineares. Conforme ilustrado na Figura 3, técnicas tradicionais de representação de textos como a BoW geralmente são baseadas em modelos de linguagem lineares que consideram contribuições igualitárias dos termos, ignorando informações semânticas importantes dos textos.

As abordagens de representação de textos baseadas em **modelos lineares** (p. ex., BoW) geralmente fundamentam-se na realização de operações vetoriais básicas como a adição e a multiplicação. Normalmente essas abordagens compõem os textos como combinações de vetores de mesma magnitude, o que apesar de implicar em perda de informações sintáticas e semânticas, proporciona um desempenho efetivo em tarefas baseadas em termos específicos como os tópicos e as categorias (MITCHELL; LAPATA, 2010).

As abordagens de representação de textos baseadas em **modelos multilineares** (p. ex., LSA/LSI e LDA) permitem considerar correlações multilineares entre os termos dos textos, que podem ser mapeados por meio de operações vetoriais como produto interno e contração de vetores. Essas abordagens costumam ser mais flexíveis do que as abordagens baseadas em modelos lineares, apesar de geralmente apresentarem resultados abstratos e menos interpretáveis em razão da multilinearidade das soluções encontradas (COECKE; SADRZADEH; CLARK, 2010; BARONI; ZAMPARELLI, 2010).

Figura 3 – Categorias de modelos composicionais distribuídos.

Fonte: Adaptada de Kartsaklis (2014, p. 70).

As abordagens de representação de textos baseadas em **modelos não lineares** (p. ex., *word embeddings* e *deep learning*) geralmente estão relacionadas ao paradigma conexionista, que tende a realizar otimizações dependentes de um domínio específico. Embora essas abordagens sejam computacionalmente mais custosas e apresentem soluções menos compreensíveis do que as abordagens baseadas nas demais categorias de modelos, elas permitem agregar um grande volume de informações semânticas e contextuais que normalmente são perdidas em técnicas tradicionais de representação de textos (COLLOBERT; WESTON, 2008; SOCHER; HUVAL; MANNING; NG, 2012; KALCHBRENNER; GREFFENSTETTE; BLUNSOM, 2014).

Na subseção a seguir são apresentados os diferentes níveis de complexidade semântica que podem ser explorados em tarefas da MT como a classificação automática de textos. A definição desses níveis de complexidade semântica é fundamental para o entendimento das motivações desta dissertação, e principalmente para a compreensão da aplicabilidade da técnica de representação de textos proposta, apresentada na Seção 3.3.

2.3.1 Níveis de complexidade semântica

Conforme abordado em Sinoara, Antunes e Rezende (2017, p. 1–3), a efetividade dos algoritmos de AM nas tarefas de MT depende principalmente do tipo de conhecimento a ser extraído, que pode estar relacionado à associações sintagmáticas e/ou à paralelos paradigmáticos. Por exemplo, em coleções de textos que apresentam paráfrases (textos equivalentes em sentido, mas formadas por vocabulários diferentes), ambiguidades (textos indefinidos ou com mais de um sentido) ou antagonismos (textos com oposição de ideias), considerar apenas o vocabulário dos textos pode ser insuficiente para classificá-los corretamente. Esse tipo de problema torna-se evidente quando a tarefa de classificação envolve alguma informação implícita nos textos, como explicado nos seguintes documentos:

- D_1 = “a companhia α adquiriu a companhia β ”;
- D_2 = “a companhia β adquiriu a companhia α ”.

Embora os documentos D_1 e D_2 sejam formados pelo mesmo vocabulário (conjunto de termos $\{a; adquiriu; companhia; \alpha; \beta\}$), eles apresentam sentidos diferentes de relação antagônica. Esse caso evidencia o fato de que representações de textos baseadas apenas no vocabulário original dos textos (conhecimento léxico) podem ser insuficientes para diferenciar seu conteúdo, já que desconsideram informações potencialmente úteis baseadas em conhecimentos sintáticos e/ou semânticos (SINOARA; ANTUNES; REZENDE, 2017, p. 2).

Considerando a análise de aspectos textuais semânticos, as tarefas de MT podem ser organizadas em dois níveis de complexidade complementares, listados a seguir. As **tarefas baseadas em tópicos (1º nível)** podem ser realizadas de maneira eficaz considerando apenas o vocabulário original dos textos, o que normalmente impossibilita a descoberta de informações implícitas como no exemplo anterior. Já as **tarefas baseadas em semântica (2º nível)** dependem essencialmente da consideração do significado dos termos, normalmente atrelado às relações que não coocorrem linearmente entre os mesmos (SINOARA; SCHEICHER; REZENDE, 2017).

- **1º nível** – Tarefas para resolução de problemas baseados em informações de tópicos;
- **2º nível** – Tarefas para resolução de problemas baseados em informações semânticas.

Conforme abordado em Sinoara, Scheicher e Rezende (2017), os dois níveis de complexidade semântica podem ser explorados isoladamente ou em conjunto. Por exemplo, considera-se uma coleção de textos de notícias esportivas relacionadas à uma competição internacional específica de futebol. Considera-se ainda que cada texto dessa coleção corresponde ao resumo de cada partida disputada, permitindo identificar a fase da competição e o país da seleção vencedora. Embora as fases da competição possam ser diretamente identificadas por meio de tópicos específicos contidos nos textos (p. ex., fase de grupos e fase eliminatória), a identificação da seleção vencedora depende da análise de aspectos textuais semânticos para determinar o vencedor caso não tenha havido um empate.

Para avaliar o desempenho das abordagens de representação de textos e determinar a aplicabilidade de cada técnica, na avaliação empírica apresentada no Capítulo 4 a tarefa de classificação automática de textos foi considerada em ambos os níveis de complexidade semântica, explorados isoladamente e em conjunto.

Na subseção a seguir são apresentados alguns dos principais trabalhos relacionados ao “estudo da semântica na área de MT”. Conforme constatado em Sinoara, Antunes e Rezende (2017), esse tema tem sido frequentemente associado ao enriquecimento de representações de textos com o objetivo de intensificar o volume de informações consideradas no processo de MT e consequentemente contribuir para uma melhora na performance dos algoritmos de AM.

2.3.2 Mapeamento da semântica na área de Mineração de Textos

Buscando verificar o atual interesse científico pelo estudo de aspectos textuais semânticos na área de MT, durante o desenvolvimento desta dissertação foi realizado um levantamento bibliográfico em forma de Mapeamento Sistemático (MS), em que foram identificadas as principais técnicas de representação de textos, as métricas de avaliação, as bases de dados utilizadas como *benchmark* e os resultados obtidos com o estudo da semântica (SINOARA; ANTUNES; REZENDE, 2017). Diferente de uma Revisão Sistemática, focada na análise detalhada de um volume reduzido de publicações, o processo de MS aborda uma análise exploratória de grandes volumes de publicações (KITCHENHAM; CHARTERS, 2007; PETERSEN; FELDT; MUJTABA; MATTSSON, 2008), nesse caso baseada nos resumos e títulos dos trabalhos.

Em razão dos extensos resultados obtidos por meio do processo de MS realizado, publicados detalhadamente em Sinoara, Antunes e Rezende (2017), as subseções a seguir descrevem exclusivamente o roteiro de pesquisa seguido e a meta-análise efetuada da literatura. Para a realização dessa meta-análise da literatura também foram considerados trabalhos identificados além do processo sistemático realizado com o objetivo de complementar a apresentação das informações obtidas; esses trabalhos não foram contabilizados nas estatísticas quantitativas apresentadas.

2.3.2.1 Roteiro de pesquisa do processo de mapeamento realizado

A realização do MS foi baseada no procedimento descrito em Kitchenham e Charters (2007), diferindo apenas na quantidade de indivíduos que conduziram as fases de seleção e de extração³, e na consideração de estudos secundários (*surveys*). O protocolo definido para guiar essa pesquisa levou em consideração a principal questão a ser respondida: “como a semântica é considerada nos trabalhos de MT?”. Conforme apresentado no Quadro 1, foram definidos tópicos distintos para melhor detalhamento dos aspectos relacionados à essa questão de pesquisa (SINOARA; ANTUNES; REZENDE, 2017, p. 4).

Quadro 1 – Questões de pesquisa que fundamentaram o MS realizado.

ID	Questões de pesquisa
1	Quais são os principais domínios de aplicação?
2	Quais idiomas são mais frequentemente abordados?
3	Quais fontes externas de conhecimento têm sido frequentemente utilizadas?
4	Quais tarefas de MT são mais frequentemente realizadas?
5	Quais métodos e algoritmos são comumente utilizados?
6	Quais representações de textos têm sido mais frequentemente empregadas?
7	Qual é a frequência de participação de usuários em processos de MT?

³ Ao invés da análise por um indivíduo e a supervisão por outro, os autores sugerem que a análise da literatura seja realizada por pelo menos duas pessoas.

Para recuperar sistematicamente os trabalhos relacionados ao tema foi utilizada a expressão de busca “*semantic* AND text* AND (mining OR representation OR clustering OR classification OR association rules)*”, aplicada aos títulos e palavras-chave dos trabalhos sempre que possível. O levantamento desses trabalhos foi efetuado por meio das seguintes bases digitais amplamente difundidas na área da Computação: *ACM Digital Library*⁴, *IEEE Xplore*⁵, *Science Direct*⁶, *Scopus*⁷ e *Web of Science*⁸ (SINOARA; ANTUNES; REZENDE, 2017, p. 4).

O processo de execução do MS realizado foi dividido em três fases complementares: busca, seleção e extração de informações. Os trabalhos obtidos na fase inicial de **busca** (3.984 resultados relativos aos *trabalhos publicados até fevereiro de 2016*) passaram por um processo manual de triagem para identificar as publicações consideradas irrelevantes ao tema. Durante a fase de **seleção** foram removidas publicações duplicadas (725 resultados) ou consideradas inadequadas (1.566 resultados) segundo um conjunto específico de critérios de rejeição formado por (SINOARA; ANTUNES; REZENDE, 2017, p. 4–5):

- Publicações com uma única página, pôsteres, apresentações, resumos e editoriais;
- Publicações escritas em idiomas diferentes do português e do inglês;
- Publicações não relacionadas à MT e/ou à semântica textual;
- Publicações com acesso restrito ou inacessíveis.

Na última fase do processo de MS realizado, a **extração de informações**, foram analisados os 1.693 trabalhos resultantes da fase de seleção buscando responder às questões de pesquisa apresentadas no Quadro 1. Alternativamente, os trabalhos selecionados nessa última fase também encontram-se disponíveis para visualização e exploração no *site* do LABIC⁹, agrupados de acordo com o assunto ou com as questões de pesquisa realizadas (SINOARA; ANTUNES; REZENDE, 2017, p. 5–16).

2.3.2.2 Meta-análise da literatura

Conforme ilustrado na Figura 4, considerando os trabalhos resultantes (aceitos) da fase de seleção do processo de MS realizado, o estudo da semântica na MT tem crescido gradativamente desde o início deste século, alcançando um pico no ano de 2011 com 223 publicações. Entre os termos mais frequentemente citados nas publicações destacam-se respectivamente: LSA/LSI, SVM, PLN, RNA, NB, LDA e KNN (SINOARA; ANTUNES; REZENDE, 2017, p. 6–12).

⁴ <<https://dl.acm.org>>

⁵ <<https://ieeexplore.ieee.org>>

⁶ <<http://www.sciencedirect.com>>

⁷ <<https://www.scopus.com>>

⁸ <<https://webofknowledge.com>>

⁹ <http://sites.labic.icmc.usp.br/pinda_sm>

Figura 4 – Evolução do interesse científico por aspectos textuais semânticos na MT.

Fonte: Adaptada de Sinoara, Antunes e Rezende (2017, p. 6).

A síntese sobre como a semântica é considerada nos trabalhos de MT está descrita a seguir. Essa meta-análise da literatura foi baseada nas respostas dos tópicos anteriormente definidos apresentados no Quadro 1 (SINOARA; ANTUNES; REZENDE, 2017).

1. **Domínios de aplicação** – “Saúde e Ciências da Vida” (33,9%) e “Web” (22,7%);
2. **Idiomas** – Inglês (45%) e chinês (26,4%); português (4%);
3. **Fontes externas de conhecimento** – WordNet (29,9%) e Wikipédia (16,6%);
4. **Tarefas de MT** – Classificação (27,4%) e agrupamento (17%);
5. **Métodos e algoritmos** – SVM, PLN (REN e DLS), RNA, NB e KNN;
6. **Representações de textos** – BoW, LSI/LSA, LDA, *word embeddings* e *deep learning*;
7. **Participação de usuários** – Pouco frequente, de maneira geral (5%).

Em relação aos domínios de aplicação, as áreas com maior concentração de estudos referem-se aos temas “Saúde e Ciências da Vida” (33,9%), “Web” (22,7%) e “Notícias” (9,3%). Em contraste com o idioma português, abordado em apenas 4% das publicações analisadas, os idiomas inglês (45%) e chinês (26,4%) são os idiomas mais frequentemente abordados nos trabalhos de MT. Já as tarefas mais comumente realizadas na MT são a classificação (27,4%) e o agrupamento (17%). A participação de usuários no processo de MT é pouco mencionada nos trabalhos analisados, possivelmente devido à dificuldade de avaliação relacionada a subjetividades. (SINOARA; ANTUNES; REZENDE, 2017, p. 9–12).

Assim como ressaltado em Sinoara, Antunes e Rezende (2017, p. 10–11), as fontes externas de conhecimento têm sido comumente utilizadas em tarefas de PLN como o REN e a DLS. As *wordnets*¹⁰ e as enciclopédias livres (p. ex., Wikipédia) destacam-se como os tipos de fontes externas mais abordados nos trabalhos, mencionadas respectivamente em 29,9% e 16,6% das publicações. De acordo com os autores, uma das principais diferenças entre esses dois tipos de fontes externas de conhecimento é o processo pelo qual elas são construídas, já que os relacionamentos de *wordnets* são gerados por especialistas de diversos domínios e, a exemplo da Wikipédia, as enciclopédias livres são criadas de forma colaborativa, o que contribui para a ocorrência de relacionamentos dúbios nos textos.

As *wordnets* são redes de relacionamentos textuais baseados em *synsets* (conjuntos de sinônimos) que representam de forma agrupada sentidos que podem ser expressos por nomes, verbos, adjetivos e advérbios (MILLER *et al.*, 1990; FELLBAUM, 1998). Conforme abordado em Camacho-Collados, Pilehvar e Navigli (2016), as *wordnets* e as enciclopédias podem ser utilizadas de forma complementar assim como ocorre na BabelNet¹¹, uma rede semântica de informações ontológicas representadas por *synsets*, disponível em diversos idiomas e composta pela combinação de fontes como a WordNet e a Wikipédia (NAVIGLI; PONZETTO, 2012).

Em relação às técnicas de representação de textos, normalmente têm-se abordado o enriquecimento dos termos como uma tentativa de minimizar a perda de informações semânticas associadas aos textos, seja pelo tratamento automático de informações contextuais (FREITAS *et al.*, 2013; MIZZARO; PAVAN; SCAGNETTO; VALENTI, 2014) ou de abordagens relacionadas à semântica latente dos textos (BLEI; NG; JORDAN, 2003; KUHN; DUCASSE; GÍRBA, 2007). A exploração de abordagens baseadas em *word embeddings* (SOHRAB; MIWA; SASAKI, 2015; WANG *et al.*, 2016) ou em *deep learning* (SCHMIDHUBER, 2015; LEBRET; COLLOBERT, 2015) também se destaca como uma das tendências para a obtenção de representações mais ricas de coleções de textos (SINOARA; ANTUNES; REZENDE, 2017, p. 13–14).

Assim como relatado na Seção 1.2, foram identificados poucos trabalhos (apenas 9,2% do total) que abordam conjuntamente aspectos semânticos e contextuais para a geração de representações de textos enriquecidas. Pelo fato de informações contextuais serem caracterizadas como uma importante fonte de conhecimento dos textos, esse volume reduzido de publicações se caracteriza como uma das principais lacunas identificadas na literatura.

Na seção a seguir são discutidos os diferentes aspectos linguísticos associados à fase de Representação de Textos do Pré-processamento. Primeiramente são apresentados e discutidos os tipos de informações e de relacionamentos textuais normalmente tratados por meio do processo de MT. Em seguida são apresentadas algumas das principais técnicas de representação de textos identificadas durante o processo de MS realizado, consideradas como *baseline* na avaliação experimental da técnica de representação de textos proposta.

¹⁰ Redes baseadas na WordNet, originalmente desenvolvida para o idioma inglês (FELLBAUM, 1998).

¹¹ <<http://babelnet.org>>

2.4 Representação de textos

Em razão de dados textuais serem naturalmente não estruturados, torna-se necessária a utilização de técnicas para representar coleções de textos por meio de estruturas de dados eficientes, como por meio de estruturas matriciais que são formatos normalmente aceitos por algoritmos de AM. Normalmente as abordagens de representação de textos são baseadas em técnicas que correlacionam os textos e seus atributos por meio de dados numéricos, que podem expressar desde a frequência de um termo em um texto (p. ex., BoW) até um ponto de uma dimensão espacial (p. ex., *word embeddings*).

O processo tradicional de representação de textos implica na divisão dos textos da coleção em unidades menores de informação, considerando-os como uma distribuição de termos (*tokens*) separados por espaços em branco (p. ex., caracteres de espaço ou quebras de linha). Conforme tratado na Subseção 2.2.1, esse processo de tokenização consequentemente implica na utilização de técnicas de extração ou de seleção de termos que visam intensificar as correlações entre os termos dos textos considerando diferentes tipos de informações textuais, que são discutidas na subseção a seguir.

2.4.1 Tipos de informações e de relacionamentos textuais

Em um paralelo com as línguas naturais, as linguagens de programação também são formadas por um alfabeto e um conjunto de regras englobadas por uma gramática, que nesse caso tem o objetivo de intermediar a comunicação entre pessoas e máquinas. Como essa intermediação deve ser exata para que os programas funcionem sem erros, os códigos escritos em linguagem de máquina passam por um processo metódico de validação denominado compilação, composto respectivamente pelos seguintes níveis de análises: a análise léxica, responsável pela verificação dos símbolos utilizados; a análise sintática, responsável pela verificação da estrutura do programa; e a análise semântica, responsável pela verificação da lógica do programa (AHO; LAM; SETHI; ULLMAN, 2006).

Análogo ao processo de compilação das linguagens de programação, as abordagens de representação de textos intermediam o conhecimento presente em coleções de textos por meio da análise e compreensão dos tipos de relacionamentos textuais. Neste caso, entretanto, o objetivo da análise computacional das coleções consiste em obter conhecimento linguístico sobre os textos processados (CHOMSKY, 1956; JURAFSKY; MARTIN, 2009).

No processo de MT o objetivo da análise computacional das coleções de textos resume-se à descoberta de conhecimento útil acerca dos documentos processados. Para a extração de padrões de dados textuais não-estruturados podem ser considerados diferentes tipos de informações associados aos possíveis relacionamentos entre os termos dos textos. De maneira geral, esses relacionamentos podem estar relacionados a informações de natureza léxica, sintática e/ou semântica (KAO; POTEET, 2007; AGGARWAL; ZHAI, 2012b).

As abordagens de representação de textos baseadas em **informações léxicas** permitem explorar relações associadas à estrutura dos termos, como na radicalização e na lematização (vide a Seção 3.2). Já as abordagens de representação de textos baseadas em **informações sintáticas** permitem explorar relações associadas às estruturas dos textos, como a associação de verbos e de substantivos por meio de categorias de palavras com propriedades gramaticais similares (*part-of-speech*) (BICK, 2000). Por fim, as abordagens de representação de textos baseadas em **informações semânticas** permitem explorar relações associadas aos significados dos termos nos textos, como os agentes (papel temático de quem realiza uma ação) ou as pessoas (classe ontológica de todas as pessoas) identificados nos textos (HENNIG; UMBRATH; WETZKER, 2008; FILHO; PARDO, 2014).

Embora cada tipo de informação textual permita explorar um aspecto específico de relacionamento entre os termos dos textos, também é possível considerá-las de maneira complementar (KAMBHATLA, 2004). Como o significado das palavras pode variar de acordo com seu contexto, é importante considerar os diferentes aspectos textuais semânticos associados ao sentidos dos termos nos textos, como as conotações (sentido literal) e a denotações (sentido figurado). Assim como retratado em Riemer (2010), esses aspectos semânticos podem ser analisados por um conjunto específico de relações entre palavras como a polissemia (palavras com múltiplos significados), a antonímia (oposição entre palavras) e a sinonímia (palavras com sentidos equivalentes).

Conforme ressaltado em Jurafsky e Martin (2009, p. 616–621), é importante que as técnicas de representação de textos sejam efetivas o bastante para capturar o significado dos textos, constituindo-se das características de veracidade (o conteúdo das representações deve ser coerente com os textos), de clareza (não ambiguidade das correlações representadas), de forma canônica (entradas com o mesmo significado devem ter a mesma representação) e de expressividade (as representações devem permitir a extração de padrões potencialmente úteis).

2.4.2 Abordagens e técnicas de representação de textos

Conforme abordado na Seção 2.3, as abordagens de representação de textos podem ser baseadas em modelos de linguagem lineares, multilineares ou não lineares, e cada técnica pode abordar aspectos diferentes dos textos, permitindo a exploração de informações léxicas, sintáticas e/ou semânticas. Como retratado a seguir, algumas dessas técnicas são voltadas para a representação matricial de textos enquanto outras são voltadas para a representação vetorial de termos, fato que não impede a representação de coleções de textos (vide a Subseção 4.2.2).

A BoW é uma das técnicas tradicionais de representação de textos mais amplamente utilizadas e difundidas na literatura, e representa uma coleção de textos por meio de atributos dimensionais únicos associados à cada termo do seu vocabulário. Normalmente a representação matricial gerada pela BoW é do tipo documento-termo¹², em que cada linha corresponde a um

¹² Também existem os tipos termo-contexto e par-padrão (TURNERY; PANTEL, 2010, p. 142).

documento de texto e cada coluna corresponde a um atributo, considerando apenas os termos simples. A correlação entre os termos dá-se pela frequência de cada atributo em cada documento, e como os atributos são considerados independentes essas representações tornam-se suscetíveis a perda de informações semânticas e a problemas de alta esparsidade e dimensionalidade (AGGARWAL; ZHAI, 2012b).

Para atenuar os problemas associados à técnicas como a BoW, as representações de textos geradas por meio da técnica LDA representam os documentos como misturas de tópicos que geram termos com certas probabilidades, destacando-se pela sua praticidade em lidar com vários tipos e volumes de dados sem um aumento computacional significativo (CRAIN; ZHOU; YANG; ZHA, 2012, p. 142–156). A técnica LDA permite explorar a semântica latente dos textos por meio dos K principais tópicos (conjuntos de termos) que os descrevem considerando as estatísticas de coocorrência de cada termo em cada texto. As representações de textos baseadas nessa técnica são geradas pelas proporções (entre 0 e 1) desses K tópicos em cada documento, já que os tópicos extraídos são considerados os atributos da representação matricial (BLEI; NG; JORDAN, 2003).

Em contraste com as técnicas de representação de textos BoW e LDA, respectivamente baseadas em modelos lineares e multilineares, as representações baseadas em *word embeddings* permitem prever termos envoltos em um mesmo contexto sem calcular suas coocorrências (BARONI; DINU; KRUSZEWSKI, 2014). Baseadas em modelos não lineares inspirados em RNAs, esse tipo de representação permite gerar vetores únicos para cada termo levando em consideração o contexto em que eles ocorrem nos textos, aprendido durante a fase de treinamento dos modelos. Além de contribuir para uma diminuição da esparsidade e da dimensionalidade das representações, já que o número de dimensões do modelo é fixado no início do processo, as representações de textos baseadas em *word embeddings* agregam informações semânticas e contextuais que geralmente são perdidas em outras representações (MIKOLOV; CHEN; CORRADO; DEAN, 2013).

Modelos de linguagem baseados em *word embeddings* têm sido aplicados em diversos domínios de aplicações e em diversos idiomas. Um exemplo dessas abordagens é a *Novel Approach to a Semantically-Aware Representation of Items* (NASARI), uma abordagem de representação de conceitos e de entidades nomeadas que combina informações linguísticas de redes semânticas e informações estatísticas da coleção analisada. A abordagem NASARI permite representar correlações entre termos de múltiplos idiomas sob duas perspectivas distintas: representações explícitas (técnicas $NASARI_{lexical}$ e $NASARI_{unified}$) e representações latentes (técnica $NASARI_{embed}$) (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016).

A técnica de representação de termos $NASARI_{lexical}$ gera vetores léxicos baseados em lematizações, ponderadas de acordo com uma medida de especificidade léxica análoga à métrica TF-IDF (vide a Equação 2.2). Na técnica $NASARI_{unified}$ os vetores de representação dos termos são gerados por meio de agrupamentos de termos correlacionados que compartilham o mesmo

hiperônimo, isto é, que estabelecem relação de generalização (p. ex., “meio de transporte” é hiperônimo de “bicicleta”), o que na prática permite associar termos de línguas diferentes em um mesmo espaço semântico. Na técnica $NASARI_{embed}$ a representação dos termos é gerada por meio de modelos baseados em *word embeddings* treinados a partir de grandes volumes de textos, o que implica na obtenção de vetores densos que agregam informações semânticas e contextuais (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016).

2.5 Considerações finais

Embora técnicas tradicionais de representação de textos como a BoW sejam efetivas em muitas aplicações da MT, normalmente elas desconsideram informações semânticas e contextuais, o que pode vir a prejudicar a performance dos algoritmos de AM utilizados. Assim, os procedimentos de enriquecimento textual destacam-se como uma alternativa efetiva para intensificar o volume e a qualidade de informações consideradas nas representações de textos. No capítulo seguinte é apresentada uma nova técnica de representação de textos baseada em contextos, proposta para permitir aliar o uso de fontes externas de conhecimento tanto para o enriquecimento semântico das coleções de textos quanto para a criação de modelos de linguagem.

ENRIQUECIMENTO DE REPRESENTAÇÕES DE TEXTOS

Neste capítulo são apresentados fatores importantes associados ao processo de enriquecimento de representações de textos, ressaltando algumas das abordagens de enriquecimento textual mais utilizadas na literatura. Primeiramente são apresentadas as técnicas de enriquecimento textual mais frequentemente empregadas em tarefas da MT, ressaltando suas principais características. Em seguida é apresentada a nova técnica de representação de textos proposta, baseada em informações contextuais e no enriquecimento semântico de textos. Por fim, também são discutidas possíveis aplicações baseadas nessa nova técnica proposta.

3.1 Considerações iniciais

Conforme abordado em Aggarwal e Zhai (2012b), as técnicas de representação de textos que desconsideram informações semânticas geralmente estão suscetíveis a problemas de sobreajuste (*overfitting*) causados pela alta esparsidade e dimensionalidade. Entretanto, técnicas tradicionais como a BoW têm sido amplamente utilizadas e difundidas devido à uma série de vantagens como a facilidade e a escalabilidade de uso. Por esse motivo, o enriquecimento textual surge como uma importante alternativa para intensificar a agregação de informações relevantes às representações de textos.

Na seção a seguir são descritas as abordagens e técnicas de enriquecimento textual usualmente realizadas na área de MT, ressaltando suas principais características e aplicabilidades na representação de textos. Conforme apresentado na Subseção 2.2.1, conceitualmente essas técnicas de enriquecimento textual estão associadas à fase de Extração de Termos, contida na etapa de Pré-processamento do processo de MT, incluindo também técnicas associadas à correlação entre termos por meio de modelos de linguagem.

3.2 Abordagens e técnicas de enriquecimento textual

O enriquecimento textual pode ser caracterizado como um processo computacional que busca intensificar o volume e a qualidade de informações consideradas em representações de textos, sejam elas relacionadas à forma dos termos (enriquecimento léxico), à constituição das sentenças (enriquecimento sintático) ou aos significados dos termos (enriquecimento semântico) (EBECKEN; LOPES; COSTA, 2003; JURAFSKY; MARTIN, 2009; AGGARWAL; ZHAI, 2012b). Assim como ressaltado em Sinoara, Antunes e Rezende (2017), por meio de técnicas de enriquecimento textual pode-se aprimorar a qualidade e a eficiência das representações de textos em razão da descoberta de atributos textuais mais relevantes para cada domínio específico.

O objetivo do enriquecimento textual é auxiliar na identificação dos principais atributos que descrevem um conjunto de textos, buscando reduzir possíveis perdas de informações léxicas, sintáticas e/ou semânticas. Como as técnicas de representação de textos normalmente consideram cada termo como um atributo independente, é importante a aplicação de procedimentos que auxiliam na identificação de termos compostos ou de conceitos e de entidades equivalentes que podem ser representados como um único atributo (AGGARWAL; ZHAI, 2012b). A seguir são apresentadas as abordagens de enriquecimento textual mais comumente aplicadas na MT, incluindo técnicas baseadas apenas em informações do vocabulário original dos textos, explicitamente e/ou implicitamente, e técnicas baseadas em fontes externas de conhecimento, que auxiliam na identificação de informações externas sobre os textos.

Entre as abordagens de enriquecimento textual baseadas apenas em informações do vocabulário original da coleção analisada, a **combinação de termos** pode ser considerada a forma mais imediata de extração de termos relacionados ou compostos (também denominados *collocations* (JURAFSKY; MARTIN, 2009)), sendo por vezes efetiva na realização de tarefas baseadas em vocabulários reduzidos ou específicos (CAVNAR; TRENKLE, 1994). Embora a identificação de termos compostos possa ser realizada por meio de combinações contínuas (p. ex., *N-grams* (BROWN *et al.*, 1992)) ou intermitentes (p. ex., *Skip-grams* (HUANG *et al.*, 1993)), as abordagens de representação de textos que consideram essas combinações tendem a crescer consideravelmente conforme a expansão do vocabulário, aumentando sua esparsidade e dimensionalidade (FÜRNKRANZ, 1998; GUTHRIE *et al.*, 2006).

Conforme abordado em Ebecken, Lopes e Costa (2003), abordagens de enriquecimento textual baseadas em **simplificação de termos**, como a radicalização, a lematização e a substantivação, buscam unificar termos similares e consequentemente intensificar as correlações entre os termos. O processo de radicalização (LOVINS, 1968) tem como objetivo reduzir os termos para seu radical ou forma inflexionável (p. ex., “conhecimento” para “conhec”). Já o processo de lematização (SCHMID, 1994) busca converter as palavras para seu lema ou forma canônica (p. ex., “conhecido” para “conhecer”), enquanto a substantivação ou nominalização (GONZALEZ; LIMA; LIMA, 2006) busca transformar as palavras em substantivos (p. ex., “conhecer” para “conhecimento”) por meio de artigos, verbos e adjetivos por exemplo.

No caso das abordagens de enriquecimento textual baseadas em **tópicos**, os textos podem ser representados em razão dos termos mais frequentes ou por meio de categorias específicas previamente conhecidas. No primeiro caso, os tópicos (atributos) geralmente são identificados por meio de técnicas baseadas em modelos multilineares (p. ex., LSA/LSI e LDA) (WANG; SUN; CHEN; ZHAI, 2006). No segundo caso, geralmente são utilizadas fontes externas de conhecimento que possibilitam a agregação de informações semânticas nas representações de textos por meio atributos complementares, como páginas da Wikipédia comuns aos temas dos textos (LI *et al.*, 2016).

Em coleções de textos formadas por vocabulários muito diversificados, as abordagens de enriquecimento textual baseadas em fontes externas de conhecimento podem auxiliar na identificação de correlações entre termos equivalentes, uma vez que o vocabulário original pode ser insuficiente para considerar correlações semânticas entre os termos dos textos (vide a Subseção 2.3.1). Entre as principais abordagens de enriquecimento textual baseadas em fontes externas de conhecimento podem ser destacadas as técnicas baseadas em: expressões de domínio, dicionários, tesouros, ontologias, contextos, classes gramaticais, papéis temáticos, REN e DLS.

As abordagens de enriquecimento textual baseadas em **expressões de domínio** buscam intensificar a relevância de atributos das representações de textos geralmente por meio da adição de termos específicos do domínio analisado. Embora a identificação de expressões relacionados ao domínio possa necessitar do auxílio manual de especialistas, técnicas baseadas em regras de associação, por exemplo, podem ser efetivas em tarefas como hierarquia de tópicos (ROSSI; REZENDE, 2011) e agrupamento (MARQUES *et al.*, 2015a; MARQUES *et al.*, 2015b).

Nas abordagens de enriquecimento textual baseadas em **dicionários**, normalmente os termos dos textos são correlacionados de acordo com uma definição formal ou padronizada de seus significados (CHANG; YANG; YANG, 2016; GARTEN *et al.*, 2018), como é o caso da normalização de textos (HAN; COOK; BALDWIN, 2013; ALMEIDA; SILVA; SANTOS; HIDALGO, 2016). Analogamente, nas abordagens baseadas em **tesouros** essas correlações são normalmente realizadas por meio de listas de termos com significados relacionados ou por meio de dicionários de termos de um domínio específico e/ou restrito (YAMAKAWA; JING; FELDMAN, 2010). De acordo com alguns autores, os tesouros normalmente não apresentam definições detalhadas dos vocabulários dos textos, salientando apenas diferenças mínimas entre os termos (TSUJII; ANANIADOU, 2005).

As **ontologias** apresentam descrições detalhadas de conceitos (p. ex., classes e coisas) em determinados domínios, incluindo hierarquias, axiomas, restrições e relações (HARB; HAJLAOUI, 2010; BERTIN; ATANASSOVA, 2012). Na tarefa de extração de informações, por exemplo, algumas abordagens baseadas em ontologias consideram os contextos como importantes fontes de informações para determinar a especificidade dos termos em cada domínio de aplicação (WIMALASURIYA; DOU, 2010). Conforme abordado em Fellbaum (1998), a rede WordNet de relacionamentos textuais baseados em *synsets* pode ser caracterizada como uma

combinação de dicionário e tesouro, embora em alguns casos fontes externas de conhecimento como a WordNet e a Wikipédia também possam estar relacionadas à construção de modelos ontológicos (SUCHANEK; KASNECI; WEIKUM, 2008; NAVIGLI; PONZETTO, 2012).

As abordagens de enriquecimento textual baseadas em conhecimento pragmático geralmente exploram os **contextos** em domínios específicos de aplicação dos termos, permitindo potencializar abordagens baseadas em outros tipos de enriquecimento textual (TAMILIN *et al.*, 2010; MIZZARO; PAVAN; SCAGNETTO; VALENTI, 2014). Em Dang, Luk e Allan (2014) e em Campos *et al.* (2018), por exemplo, informações contextuais são utilizadas de maneira efetiva para potencializar a identificação de termos relevantes nos textos. Já em Aljaber, Martinez, Stokes e Bailey (2011) as informações contextuais são extraídas das referências de trabalhos citados em textos científicos, cujos contextos são caracterizados pelos termos que ocorrem no entorno das citações. Segundo os autores, a expansão dos atributos da representação de textos por meio dos termos presentes nos contextos de citação melhora os resultados da tarefa de classificação automática de textos, principalmente se combinados com termos extraídos de tesouros, já que apresentam maior volume de informações semânticas.

Informações contextuais também têm sido investigadas na composição distribuída de representações de termos (LI; YANG, 2018) e de documentos (LE; MIKOLOV, 2014), permitindo explorar similaridades entre atributos textuais outrora considerados independentes (BORDES; GLOROT; WESTON; BENGIO, 2012; KIM; KIM; CHO, 2017). Utilizando modelos baseados em *word embeddings* (vide a Figura 5), por exemplo, os termos dos textos são representados por meio de vetores únicos formados por um número pré-definido de valores reais conhecidos como dimensões do modelo de linguagem. Nesse caso, para a composição da representação matricial de uma coleção de textos podem ser utilizadas diferentes estratégias de combinação vetorial, como a média aritmética das representações individuais de cada termo em cada um dos textos, denominada *centroide* (SOHRAB; MIWA; SASAKI, 2015).

Figura 5 – Modelos do Word2Vec e exemplo de associações entre vetores densos.

Fonte: Adaptada de Mikolov, Chen, Corrado e Dean (2013), Mikolov *et al.* (2013).

Como na fase de treinamento de modelos de linguagem baseados em *word embeddings* pode-se considerar tanto os textos da coleção analisada quanto textos de fontes externas de conhecimento, os vetores gerados acabam agregando um grande volume de informações semânticas e contextuais (MIKOLOV; CHEN; CORRADO; DEAN, 2013). Conforme ilustrado na Figura 5b, a projeção ortogonal desses vetores permite correlacionar cada termo em razão do contexto em que ocorrem, possibilitando associações diretas como por exemplo os países e suas capitais (MIKOLOV *et al.*, 2013).

Entre os principais modelos baseados em *word embeddings* destacam-se os do Word2Vec e do GloVe, que diferem sobretudo na maneira como geram os vetores densos dos termos (BARONI; DINU; KRUSZEWSKI, 2014). Enquanto o GloVe baseia-se em contagens de coocorrências dos termos (PENNINGTON; SOCHER; MANNING, 2014), os modelos do Word2Vec (vide a Figura 5a) são baseados em RNAs que permitem prever termos envoltos em um mesmo contexto a partir de janelas deslizantes (MIKOLOV; CHEN; CORRADO; DEAN, 2013).

O Word2Vec é composto pelos modelos *Continuous Bag of Words* (CBoW) e *Skip-gram*, respectivamente destinados à especificação e à generalização dos contextos dos termos (MIKOLOV; CHEN; CORRADO; DEAN, 2013). Entretanto, o CBoW é mais recomendável para a aplicação em grandes coleções de textos por ser mais eficiente do que o outro modelo nesse cenário (MIKOLOV; LE; SUTSKEVER, 2013), sendo frequentemente utilizado em trabalhos da literatura (IACOBACCI; PILEHVAR; NAVIGLI, 2015; IACOBACCI; PILEHVAR; NAVIGLI, 2016; MANCINI; CAMACHO-COLLADOS; IACOBACCI; NAVIGLI, 2017).

Quanto à identificação de informações linguísticas específicas nos textos, podem ser exploradas desde a anotação de **classes gramaticais** como substantivos e adjuntos (BICK, 2000; MARNEFFE; MACCARTNEY; MANNING, 2006) e de **papéis temáticos** como agentes e pacientes (VASQUES; ZAMBON; BAIOCO; MARTINS, 2016; SINOARA; ROSSI; REZENDE, 2016; SCHEICHER; SINOARA; KOGA; REZENDE, 2016) até o **REN** (LESER; HAKENBERG, 2005; SINOARA *et al.*, 2014) e a **DLS** (TAMILIN *et al.*, 2010; BIEMANN; FARALLI; PANCHENKO; PONZETTO, 2018), que permitem reconhecer sistematicamente a classe apropriada de cada termo em um contexto específico. Para determinar o sentido mais provável de cada palavra alvo, por exemplo, os procedimentos de DLS utilizam repositórios de sentidos (p. ex., WordNet) e normalmente se baseiam apenas no contexto local das palavras alvo (HEARST, 1991; AGIRRE; STEVENSON, 2007), embora também possam considerar informações externas de outros documentos (AGIRRE; SOROA, 2009; NÓBREGA; PARDO, 2013).

Na próxima seção é apresentada a nova técnica de representação de textos proposta nesta dissertação, que permite considerar informações semânticas e contextuais independente de aspectos específicos como o idioma, o domínio e o tamanho dos textos. Por meio dessa nova técnica é possível explorar combinações de diferentes modelos de linguagem e a aplicação de diferentes abordagens de enriquecimento textual, implicando em representações de textos mais interpretáveis e com maior densidade de informações semânticas.

3.3 CIRT.5, uma proposta de representação de textos

A nova técnica de representação de textos, denominada *Contextual Information Representation Technique based on 5 steps* (CIRT.5), tem como objetivo possibilitar a geração de representações de textos semanticamente enriquecidas baseadas em contextos, ao invés de termos. Diferente de abordagens baseadas em BoW, em relação à independência dos atributos das representações de textos, e de modo similar ao que ocorre na técnica LDA, em relação à exploração de relacionamentos latentes entre os termos dos textos, os atributos gerados por meio da técnica CIRT.5 são caracterizados por conjuntos de termos correlacionados (contextos), que possibilitam considerar a dependência dos atributos no cálculo de suas frequências.

Conforme ilustrado na Figura 6, a técnica CIRT.5 é composta por cinco etapas complementares que caracterizam a etapa de Pré-processamento do processo de MT, permitindo considerar informações semânticas e contextuais por meio de abordagens de enriquecimento textual como REN, DLS e modelos baseados em *word embeddings*. Nesse processo de representação também podem ser utilizados textos de fontes externas de conhecimento no treinamento dos modelos de linguagem, permitindo a agregação de termos ausentes do vocabulário da coleção analisada e, conseqüentemente, de correlações latentes entre os termos dos textos.

Figura 6 – Arquitetura da técnica de representação de textos CIRT.5.

Fonte: Elaborada pelo autor.

Nas subseções a seguir as etapas da técnica CIRT.5 são descritas detalhadamente por meio de um exemplo prático, ilustrando passo a passo a arquitetura ilustrada na Figura 6, simbolicamente referenciada no canto superior esquerdo de cada figura. Por motivo de conveniência, a coleção de textos analisada foi extraída da página inicial do LABIC e os textos de fontes externas de conhecimento foram extraídos da página da Wikipédia sobre o ICMC. No caso dos textos da Wikipédia, as ilustrações a seguir evidenciam parcialmente o conteúdo considerado no exemplo de funcionamento da técnica CIRT.5 (vide a Subseção 3.3.3).

3.3.1 Etapa de Preparação: enriquecimento dos textos

A primeira etapa da técnica CIRT.5 (**Preparação**) diz respeito à preparação dos textos utilizados na representação de textos, com o objetivo de potencializar o processo de extração de termos por meio de abordagens de enriquecimento textual. Embora essa seja uma etapa opcional, a aplicação de procedimentos de enriquecimento semântico baseados em REN ou DLS permite reduzir problemas de interpretação, causados pela ocorrência de termos ambíguos, de acordo com o contexto em que os termos ocorrem (JURAFSKY; MARTIN, 2009).

Conforme ilustrado na Figura 7, na técnica CIRT.5 recomenda-se o enriquecimento textual dos textos da coleção analisada e das fontes externas de conhecimento utilizadas para o pré-treinamento do modelo de linguagem, ignorando ainda termos considerados irrelevantes (p. ex., pontuações e conectivos). Esse pré-treinamento diz respeito ao processo inicial de treinamento do modelo a partir de um conjunto de documentos suficientemente representativo do domínio analisado. O pré-treinamento de modelos de linguagem normalmente pode ser realizado de maneira efetiva a partir de grandes volumes de dados textuais de domínio geral como páginas da Wikipédia (IACOBACCI; PILEHVAR; NAVIGLI, 2015; MANCINI; CAMACHO-COLLADOS; IACOBACCI; NAVIGLI, 2017).

Figura 7 – Exemplo de enriquecimento textual baseado em REN.

Fonte: Elaborada pelo autor.

No caso da aplicação de abordagens de REN ou de DLS, que implicam na identificação de informações relevantes como classes de palavras, conceitos e entidade nomeadas, devem ser consideradas duas variáveis principais: a identificação do segmento que melhor representa uma entidade ou conceito e o limiar de similaridade de desambiguação, já que podem ser identificados vários conceitos ou entidades para um mesmo segmento de texto (MORO; RAGANATO; NAVIGLI, 2014). Por exemplo, na sentença “O Laboratório de Inteligência Computacional foi fundado em 1995” o conceito “Inteligência Computacional” intuitivamente é mais representativo do que os termos isolados “Inteligência” e “Computacional”, estando associado ao maior segmento contínuo de termos que representa o sentido de interesse dessa sentença.

3.3.2 Etapa de Delimitação: delimitação dos escopos de interesse

A segunda etapa da técnica CIRT.5 (**Delimitação**) diz respeito à delimitação de escopos textuais, com o objetivo de considerar informações contextuais associadas localmente à cada região de interesse dos textos da coleção analisada. Normalmente as técnicas de representação de textos ignoram possíveis relacionamentos locais entre os termos. Relacionamentos associados à regiões como sentenças e parágrafos, por exemplo, acabam sendo transformadas em um único fragmento de texto. Conforme ilustrado na Figura 8, recomenda-se a delimitação de escopos dos textos da coleção analisada e das fontes externas de conhecimento.

Figura 8 – Exemplo de delimitação de escopos baseados em sentenças.

Fonte: Elaborada pelo autor.

A delimitação de escopos textuais define o âmbito e o número de contextos identificados para cada texto, já que cada contexto é originado a partir de um escopo. Ao invés dos textos serem tratados como conjuntos de atributos independentes, eles passam a ser caracterizados por atributos dependentes representados por contextos (conjuntos de termos correlacionados). A utilização de atributos compostos permite explorar de forma explícita as correlações entre os textos por meio de informações implícitas ou latentes.

Como os contextos são gerados por meio de inferências aos modelos de linguagem, a delimitação de escopos impacta diretamente na generalização dos termos mais correlacionados. Embora normalmente modelos de linguagem baseados em *word embeddings* sejam treinados a partir das sentenças dos textos, uma alternativa imediata seria a utilização dos parágrafos dos textos. Pelo fato de parágrafos serem facilmente extraídos a partir de quebras de linha, essa acaba sendo uma opção computacionalmente menos custosa do que a utilização de sentenças.

3.3.3 Etapa de Modelagem: construção do modelo de linguagem

A terceira etapa da técnica CIRT.5 (**Modelagem**) diz respeito à construção de um modelo de linguagem, com o objetivo de permitir a identificação dos termos mais correlacionados em cada escopo da coleção analisada. Conforme ilustrado na Figura 9, recomenda-se que o modelo

de linguagem utilizado seja previamente treinado a partir de textos de fontes externas de conhecimento e posteriormente treinado a partir dos textos da coleção analisada. Essa abordagem torna possível considerar correlações entre termos ausentes do vocabulário original, possivelmente contribuindo para o enriquecimento semântico das representações de textos.

Figura 9 – Exemplo de construção de um modelo baseado em *word embeddings*.

Fonte: Elaborada pelo autor.

Conforme ressaltado na Subseção 3.3.1, a etapa de construção do modelo de linguagem deve levar em consideração versões enriquecidas dos textos para potencializar a descoberta de correlações mais densas e menos espúrias entre os contextos. A arquitetura do modelo de linguagem deve permitir a descoberta dos termos mais correlacionados considerando todos os textos da coleção analisada, ou seja, dos termos mais próximos no espaço semântico gerado (GABRILOVICH; MARKOVITCH, 2007; LEVY; GOLDBERG, 2014). A utilização de modelos baseados em *word embeddings*, por exemplo, torna possível correlacionar os termos pela proximidade espacial de suas representações vetoriais, medida pelo cosseno do ângulo formado entre os vetores densos.

3.3.4 Etapa de Contextualização: extração dos contextos

A quarta etapa da técnica CIRT.5 (**Contextualização**) diz respeito à extração de contextos textuais, com o objetivo de identificar os termos mais correlacionados à cada região ou escopo dos textos da coleção analisada. Conforme ilustrado na Figura 10, os contextos extraídos dos textos podem incluir termos originalmente não pertencentes ao vocabulário da coleção analisada, considerados a partir dos textos de fontes externas de conhecimento utilizados no pré-treinamento do modelo de linguagem. Nesse caso ambas as sentenças foram associadas ao termo “LABIC”, originalmente ausente da coleção de textos analisada.

Como os contextos são baseados nos escopos previamente delimitados, é provável que os textos tenham números de contextos diferentes, exceto se os escopos forem delimitados em razão do conteúdo total dos textos. O processo de extração de contextos pode ser caracterizado pela

seguinte interrogação: “quais são os T termos mais correlacionados ao conjunto de termos X ?”. Nessa questão, T é o número de termos retornados pelo modelo de linguagem e X é conjunto original de termos provenientes de cada escopo previamente delimitado.

Figura 10 – Exemplo de extração de contextos textuais baseados em sentenças.

Fonte: Elaborada pelo autor.

3.3.5 Etapa de Representação: correlação entre os contextos

A quinta etapa da técnica CIRT.5 (**Representação**) diz respeito à correlação entre os contextos textuais extraídos, com o objetivo de representar a coleção analisada por meio das similaridades identificadas entre os contextos. Conforme ilustrado na Figura 11, os atributos (contextos) da representação de textos são considerados dependentes uns dos outros pelo fato de serem formados por termos comuns a outros atributos. Assim, o incremento das frequências literais (quantidade total de repetições) dos atributos pode ser qualquer valor entre 0 e 1, dependendo da similaridade calculada entre cada par de contextos.

Figura 11 – Exemplo de correlações entre contextos textuais.

Fonte: Elaborada pelo autor.

A técnica CIRT.5 também possibilita a moderação no cálculo dos valores dos atributos, permitindo desconsiderar similaridades inferiores a um limiar previamente estabelecido. Essa moderação possibilita a investigação do impacto causado por informações semânticas latentes em abordagens de representação de textos. Por exemplo, considerando um documento de texto D_1 formado pelo conjunto de contextos $\{C_1 = \{t_1; t_2; t_3\}, C_2 = \{t_2; t_4\}\}$, quando a similaridade $S_{C_1, C_2} = \frac{\text{tamanho}(C_1 \cap C_2)}{\text{tamanho}(C_2)}$ (tal que $S \in \mathbb{R}, 0 \leq S \leq 1$) entre os contextos C_1 e C_2 é maior ou igual a um limiar L (tal que $L \in \mathbb{R}, 0 \leq L \leq 1$), então o valor do atributo C_2 de D_1 é acrescido de S_{C_1, C_2} . Analogamente, quando a similaridade S_{C_2, C_1} entre os contextos C_2 e C_1 é maior ou igual ao limiar L , então o valor do atributo C_1 é acrescido de S_{C_2, C_1} .

No exemplo apresentado na Figura 11, o limiar de similaridade considerado é igual a 0,25, o que significa que qualquer similaridade maior ou igual a 25% entre dois contextos arbitrários será considerada como incremento na frequência de ambos os atributos. Nesse caso o tamanho de todos os contextos foi definido previamente como sendo 5, sendo que os dois contextos em destaque possuem similaridade igual a 0,4, já que compartilham os mesmos termos “Laboratório” e “LABIC” (40% de similaridade, já que são 2 termos de 5). Normalmente só seriam contabilizadas as frequências literais dos atributos nos documentos, o que implicaria que a frequência de ambos os contextos em destaque seria igual a 1 no documento D_1 . Porém, nessa nova técnica a similaridade entre os atributos também é considerada como parte de sua frequência, o que nesse caso implica em um acréscimo de 0,4 para ambos os contextos.

Na próxima seção são apresentadas duas novas aplicações baseadas na técnica CIRT.5, ambas explorando os modelos de linguagem baseados em *word embeddings*. Na primeira aplicação são descritos os procedimentos relacionados ao uso da técnica CIRT.5 para a representação de textos, avaliada empiricamente no Capítulo 4. Na segunda aplicação são descritos os procedimentos relacionados ao uso da técnica CIRT.5 para a extração de contextos textuais, avaliada empiricamente e publicada em Sundermann, Antunes, Domingues e Rezende (2018).

3.4 Aplicações baseadas na técnica CIRT.5

Conforme abordado na Seção 3.2, as abordagens de enriquecimento textual inspiradas em modelos baseados em *word embeddings* destacam-se de abordagens tradicionais devido a sua eficiência e escalabilidade em lidar com grandes volumes de textos. Embora essas abordagens normalmente necessitem de maior tempo e custo computacional na fase de treinamento dos modelos de linguagem, o grande volume de textos considerados permite gerar representações com alta densidade de informações semânticas e contextuais (MIKOLOV *et al.*, 2013).

As aplicações apresentadas nas subseções a seguir são baseadas na técnica CIRT.5 e foram implementadas utilizando o modelo de linguagem CBoW do Word2Vec. Para a implementação dos *scripts* criados foi utilizado o *toolkit* Gensim¹, formado por um conjunto de bibliotecas

¹ <<https://radimrehurek.com/gensim/>>

de código aberto e acesso livre disponível na linguagem de programação Python (ŘEHŮŘEK; SOJKA, 2010). Também foi utilizada a API Babelfy², um *software* de acesso livre disponível na linguagem de programação Java que permite a aplicação dos processos de REN e de DLS por meio da rede semântica BabelNet (MORO; CECCONI; NAVIGLI, 2014).

3.4.1 Representação de textos

Dependendo da aplicação a técnica CIRT.5 pode ser utilizada para a representação de textos, conforme retratado no Algoritmo 1, ou para a extração de contextos textuais, conforme retratado no Algoritmo 2. Em ambos os casos pode ser utilizado o *script* S-Enrich.py³, que além de permitir a aplicação dos processos de REN ou de DLS por meio da API Babelfy também possibilita delimitar os escopos textuais de interesse baseando-se nas sentenças dos textos, uma tarefa de PLN conhecida como *Sentence Boundary Detection* (CONDORI; PARDO, 2015). Para a segmentação dos textos em sentenças o *script* S-Enrich.py utiliza a API NLTK⁴, que possibilita a identificação de sentenças em diversos idiomas, incluindo português e inglês.

Algoritmo 1 – Técnica de representação de textos CIRT.5.

Requer: documentos D da coleção a ser analisada, documentos E das fontes externas de conhecimento, tamanho T dos contextos e limiar L de similaridade.

Assegure: matriz R de representação de textos.

```

1: procedimento CIRT.5( $D, E$ )
2:    $D, E \leftarrow \text{PREPARAR}(D, E)$ 
3:    $D, E \leftarrow \text{DELIMITAR}(D, E)$ 
4:    $M \leftarrow \text{MODELAR}(D, E)$                                 ▷ Cria um novo modelo  $M$ 
5:    $D, A \leftarrow \text{CONTEXTUALIZAR}(M, D, T)$                     ▷ Contextos formados por  $T$  termos
6:    $m \leftarrow \text{tamanho}(D)$                                     ▷ Tamanho do conjunto  $D$  de documentos
7:    $n \leftarrow \text{tamanho}(A)$                                     ▷ Tamanho do conjunto  $A$  de atributos
8:    $R \leftarrow \text{INSTANCIARMATRIZ}(m, n)$                         ▷ Nova matriz com  $m$  linhas e  $n$  colunas
9:   para todo  $d \in D$  faça
10:    para todo  $a \in A$  faça                                    ▷ Percorre o conjunto  $A$  de todos os atributos
11:      para todo  $c \in d$  faça                                  ▷ Percorre os contextos  $c$  de cada documento  $d$ 
12:         $s \leftarrow \text{SIMILARIDADE}(a, c)$ 
13:        se  $s \geq L$  então                                       ▷ Moderação por meio do limiar  $L$ 
14:           $R(d, a) \leftarrow R(d, a) + s$                        ▷ Incrementa a frequência do atributo  $R(d, a)$ 
15:        fim se
16:      fim para
17:    fim para
18:  fim para
19:  retorna  $R$                                                     ▷ Representação de textos do conjunto de documentos  $D$ 
20: fim procedimento

```

---

² <<http://babelfy.org>>

³ <<https://github.com/joao8tunes/S-Enrich>>

⁴ <<https://www.nltk.org/api/nltk.tokenize>>

Assim como retratado no Algoritmo 1 (linhas 2 e 3), o *script* `S-Enrich.py` está associado às duas primeiras etapas da técnica CIRT.5: **Preparação** e **Delimitação**. Os procedimentos associados às demais etapas (a partir da linha 4, inclusive) podem ser executados por meio do *script* `CIRT.5_embed.py`⁵, implementado por meio do módulo `Word2Vec` do *toolkit* `Gensim`⁶.

Após a execução das etapas de Preparação e de Delimitação por meio do *script* `S-Enrich.py`, os textos originais tanto da coleção analisada quanto das fontes externas de conhecimento utilizadas tornam-se semanticamente enriquecidos e contextualmente delimitados por suas sentenças. Termos considerados irrelevantes são removidos dos textos na etapa de Preparação (p. ex., pontuações e conectivos). Na etapa de **Modelagem** da técnica CIRT.5⁷ (linha 4) esses textos são empregados no treinamento do modelo de linguagem utilizado, que pode considerar textos de fontes externas de conhecimento (*E*). Nesse caso podem ser utilizados os *scripts* `Wiki2Model.py` e `Wiki2Model_S-Enrich.py`⁸, que permitem transformar os textos da Wikipédia em modelos baseados no `Word2Vec`, considerando os termos originais ou os termos enriquecidos por meio de procedimentos baseados em REN e DLS, respectivamente.

Embora o pré-treinamento do modelo de linguagem a partir de textos de fontes externas de conhecimento como a Wikipédia seja uma recomendação da técnica CIRT.5, os modelos também podem ser treinados apenas a partir dos textos da coleção de documentos analisada (*D*). Em ambos os casos pode ser utilizado o *script* `CIRT.5_embed.py`, que permite tanto o treinamento de novos modelos a partir da coleção analisada quanto o retreinamento de modelos existentes baseado nesses mesmos textos. Nesse último caso, por exemplo, modelos de linguagem pré-treinados a partir de páginas da Wikipédia podem ser gerados com ou sem enriquecimento semântico baseado em procedimentos de REN e DLS por meio dos *scripts* `Wiki2Model_S-Enrich.py` (gera 2 modelos semanticamente enriquecidos por meio do CBoW do `Word2Vec`) e `Wiki2Model.py` (gera 1 modelo por meio do CBoW do `Word2Vec`), respectivamente.

Na etapa de **Contextualização** da técnica CIRT.5 (linha 5) o modelo de linguagem treinado (*M*) é utilizado para extrair os contextos de cada escopo delimitado. No *script* `CIRT.5_embed.py` essa operação é realizada por meio da função `most_similar`, que calcula os *T* termos mais próximos aos termos originais de cada escopo considerando como medida de similaridade o cosseno dos ângulos formados entre os vetores densos. Dessa maneira, a similaridade entre dois vetores é máxima quando os mesmos apontam na mesma direção (ângulo de 0°) e mínima quando os mesmos são perpendiculares (ângulo de 90°) (MIKOLOV; CHEN; CORRADO; DEAN, 2013; MIKOLOV *et al.*, 2013).

Na etapa de **Representação** da técnica CIRT.5 (a partir da linha 6, inclusive) são considerados todos os contextos extraídos da etapa anterior, sem repetições. Esse conjunto de contextos caracteriza-se como os atributos (*A*) da representação de textos (*R*), e seus valores dependem do

⁵ <https://github.com/joao8tunes/CIRT.5_embed>

⁶ <<https://radimrehurek.com/gensim/models/word2vec>>

⁷ Também denominada *CIRT.5_{embed}* no domínio de modelos baseados em *word embeddings*.

⁸ <<https://github.com/joao8tunes/Wiki2Model>>

grau de similaridade em cada documento de texto (d) da coleção analisada. Após a instanciação da matriz de representação de textos (linha 8) são percorridos todos os contextos (c) de cada texto para calcular sua similaridade (s) com cada atributo (a) identificado.

Como há uma moderação no cálculo do valor literal dos atributos (vide a Subseção 3.3.5), similaridades menores do que o limiar (L) previamente definido são desconsideradas na técnica CIRT.5. A tendência é que a definição de limiares menores do que 1 (100%) auxilie na redução da esparsidade das representações de textos, já que permite considerar incrementos não unitários dos atributos, considerados dependentes. Em relação à dimensionalidade das representações geradas, também podem ser consideradas normalizações como a métrica TF-IDF (vide a Equação 2.2), que permite transmitir a relevância do contexto tanto nos textos em que ocorrem quanto em toda a coleção analisada. Análogo à variação dessa métrica apresentada em Goossen *et al.* (2011), nesta dissertação a variação da métrica TF-IDF para contextos é denominada *Context Frequency - Inverse Document Frequency* (CF-IDF) por motivo de adequação terminológica.

3.4.2 Extração de contextos textuais

A técnica *Contextual Information Extraction Technique based on 5 steps* (CIET.5) é uma proposta de extração de contextos textuais, e tem como objetivo transformar uma coleção de textos em uma coleção de conjuntos de contextos formados por termos correlacionados. Conforme retratado no Algoritmo 2, assim como na técnica CIRT.5, as duas primeiras etapas da técnica CIET.5 (linhas 2 e 3) são a **Preparação** e a **Delimitação**, que podem ser executadas por meio do mesmo *script* `S-Enrich.py`. Na técnica CIET.5 também recomenda-se o enriquecimento textual tanto dos textos da coleção analisada quanto dos textos de fontes externas de conhecimento utilizados no pré-treinamento do modelo de linguagem, implicando na remoção de termos considerados irrelevantes (p. ex., pontuações e conectivos).

Após a realização das etapas de Preparação e Delimitação da técnica CIET.5⁹ (a partir da linha 4, inclusive) pode ser utilizado o *script* `CIET.5_embed.py`¹⁰, que assim como o *script* `CIRT.5_embed.py` foi implementado considerando os modelos do Word2Vec. Durante a etapa de **Modelagem** da técnica CIET.5 (linha 4) também podem ser considerados textos de fontes externas de conhecimento para o pré-treinamento do modelo de linguagem, utilizando por exemplo textos da Wikipédia por meio do *script* `Wiki2Model.py` ou `Wiki2Model_S-Enrich.py`.

A etapa de **Contextualização** (linha 5) é idêntica em ambas as técnicas propostas, embora os contextos extraídos sejam aplicados com propósitos diferentes. Na etapa de **Extração** da técnica CIET.5 (a partir da linha 6, inclusive) são realizados os procedimentos de correlação entre os contextos por meio do limiar de similaridade L . Diferente da técnica CIRT.5, a saída da técnica CIET.5 é o conjunto (R) dos grupos de contextos (C) extraídos de cada documento de texto (d), e não uma representação de textos utilizando esses contextos como atributos.

⁹ Também denominada CIET.5_{embed} no domínio de modelos baseados em *word embeddings*.

¹⁰ <https://github.com/joao8tunes/CIET.5_embed>

Algoritmo 2 – Técnica de extração de contextos CIET.5.

Requer: documentos D da coleção a ser analisada, documentos E das fontes externas de conhecimento, tamanho T dos contextos e limiar L de similaridade.

Assegure: conjunto R de grupos de contextos textuais.

```

1: procedimento CIET.5( $D, E$ )
2:    $D, E \leftarrow \text{PREPARAR}(D, E)$ 
3:    $D, E \leftarrow \text{DELIMITAR}(D, E)$ 
4:    $M \leftarrow \text{MODELAR}(D, E)$                                 ▷ Cria um novo modelo  $M$ 
5:    $D, A \leftarrow \text{CONTEXTUALIZAR}(M, D, T)$                   ▷ Contextos formados por  $T$  termos
6:    $R \leftarrow \{\}$                                            ▷ Novo conjunto de grupos de contextos
7:   para todo  $d \in D$  faça
8:      $C \leftarrow \{\}$ 
9:     para todo  $a \in A$  faça                                ▷ Percorre o conjunto  $A$  de todos os atributos
10:      para todo  $c \in d$  faça                                ▷ Percorre os contextos  $c$  de cada documento  $d$ 
11:         $s \leftarrow \text{SIMILARIDADE}(a, c)$ 
12:        se  $s \geq L$  então                                    ▷ Moderação por meio do limiar  $L$ 
13:           $C \leftarrow C + a$ 
14:        fim se
15:      fim para
16:    fim para
17:     $R \leftarrow R + C$ 
18:  fim para
19:  retorna  $R$                                                 ▷ Contextos do conjunto de documentos  $D$ 
20: fim procedimento

```

Como uma abordagem de enriquecimento textual, a técnica CIET.5 foi inicialmente explorada de maneira empírica na área de Sistemas de Recomendação (SUNDERMANN; ANTUNES; DOMINGUES; REZENDE, 2018). Segundo os autores, a aplicação da técnica CIET.5 provê uma melhora significativa na performance de sistemas de recomendação sensíveis ao contexto em relação aos que não utilizam contexto. Nesse caso, a performance da técnica CIET.5 foi avaliada em uma aplicação específica baseada nas sentenças de *reviews*, considerando textos da Wikipédia para o pré-treinamento dos modelos de linguagem utilizados.

Devido às múltiplas combinações paramétricas possíveis da técnica CIET.5, e ao fato do foco desta dissertação estar associado à análise de abordagens de representações de textos, essa nova técnica de extração de contextos textuais não é avaliada nesta dissertação. Para a realização de uma análise empírica da técnica CIET.5 como um procedimento de enriquecimento semântico e contextual de coleções de textos, teriam que ser avaliados os impactos causados por diferentes tamanhos de contextos e valores de limiares na performance de abordagens de representação de textos. Na tarefa de classificação automática de textos, por exemplo, a avaliação da técnica CIET.5 pode ser realizada de maneira análoga aos procedimentos descritos no próximo capítulo, em que são apresentados indícios promissores a respeito da efetividade da técnica CIRT.5 na tarefa de classificação automática de textos.

3.5 Considerações finais

Conforme abordado neste capítulo, as abordagens de enriquecimento textual são alternativas efetivas para intensificar o volume e a qualidade de informações consideradas em representações de textos. Abordagens de enriquecimento textual baseadas em *word embeddings* e/ou em procedimentos de REN e DLS, por exemplo, podem ser exploradas em aplicações baseadas em domínios específicos como a utilização da técnica CIET.5 em Sistemas de Recomendação, ou em domínios gerais como a utilização da técnica CIRT.5 em tarefas do processo de MT. No capítulo seguinte são descritas as avaliações experimentais realizadas para verificar a efetividade da técnica de representação de textos CIRT.5 na tarefa de classificação automática de textos. A performance dessa técnica é avaliada em diferentes idiomas considerando os diferentes níveis de complexidade semântica. Também é investigado o impacto causado pelo uso de fontes externas de conhecimento, tanto para o enriquecimento textual das coleções quanto para a melhoria da performance de algoritmos supervisionados de AM.

AVALIAÇÃO EMPÍRICA, RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados e discutidos os experimentos realizados para verificar o desempenho e a aplicabilidade da técnica de representação de textos proposta na tarefa de classificação automática de textos. Além da organização e da configuração dos experimentos, que descrevem as bases de dados utilizadas, as abordagens de representação de textos avaliadas e os algoritmos de classificação aplicados, também são apresentados e discutidos os impactos causados por fatores como o uso de fontes externas de conhecimento e o tratamento de diferentes idiomas por meio de enriquecimentos semânticos e contextuais. Ao final são apresentadas prováveis correlações dos parâmetros da técnica CIRT.5, permitindo uma análise exploratória mais refinada dos cenários onde há ganhos ou perdas de performance.

4.1 Considerações iniciais

A avaliação de abordagens de enriquecimento textual consiste basicamente em avaliar a performance de representações de textos geradas a partir de coleções de textos enriquecidos, considerando diferentes tarefas ou cenários de aplicação. Considerando a tarefa de classificação automática de textos, a avaliação dessas abordagens pode ser realizada em razão da performance das representações de textos, analisando possíveis tipos de ganho como reduções de esparsidade e dimensionalidade, e o desempenho da tarefa de classificação em si.

Geralmente o impacto de diferentes abordagens de representação de textos é avaliado por medidas que calculam a qualidade da classificação realizada, como a Acurácia, a Precisão e a Revocação. Entretanto, também podem ser avaliados o custo/benefício dos tempos de geração dos classificadores e da tarefa de classificação em si, assim como a esparsidade, a dimensionalidade e compreensibilidade das representações. Na seção a seguir são abordados alguns dos principais indicadores de desempenho identificados para medir a performance da técnica CIRT.5.

4.2 Organização e configuração experimental

Baseado nas etapas do processo tradicional de MT apresentado na Subseção 2.2.1, as subseções a seguir relatam os procedimentos respectivamente realizados nas etapas de Identificação do Problema, de Pré-processamento e de Extração de Padrões. Inicialmente são apresentadas as bases de dados utilizadas e os recursos textuais aplicados durante os experimentos realizados. Em seguida são apresentadas as abordagens de representação de textos avaliadas, baseando-se no desempenho de alguns dos algoritmos de classificação mais utilizados na literatura.

Na Seção 4.3 e na Seção 4.4 são apresentadas e discutidas respectivamente as performances das abordagens de representação de textos avaliadas e as propriedades paramétricas da técnica de representação de textos CIRT.5. Em ambas as avaliações são analisados os impactos causados por diferentes fatores no desempenho de algoritmos supervisionados de AM na tarefa de classificação automática de textos, como o enriquecimento textual das coleções de textos e o uso de fontes externas de conhecimento.

4.2.1 Bases de dados e recursos textuais utilizados

Para permitir a análise empírica da técnica CIRT.5 em diferentes cenários de aplicação, foram pré-processadas 3 bases de dados textuais relativas a 2 idiomas. Para analisar as abordagens de representação de textos no idioma português, foi selecionada a base de dados *BEST sports* v1.1¹ (SINOARA; REZENDE, 2018). Para o idioma inglês foram selecionadas as bases de dados *SE-ABSA15*² (PONTIKI *et al.*, 2015) e *BBC*³ (GREENE; CUNNINGHAM, 2006). Por motivo de comodidade, os nomes das duas primeiras bases de dados são abreviados respectivamente por *BS-4* (versão da *BEST sports* v1.1 com as 4 classes majoritárias, menos desbalanceada) e *SE* (versão original da *SE-ABSA15*); análogo à base de dados *SE*, a base de dados *BBC* também corresponde à sua versão original contendo todos os documentos (vide a Tabela 1).

Tabela 1 – Coleções de textos preparadas para a tarefa de classificação.

Coleção	Base de dados	Tema	Tipo de classe	# Documentos / Classes
BS-4	<i>BEST sports</i> v1.1	Notícias	Tópico e/ou Semântica	283 (881) / 4 (66)
SE	<i>SE-ABSA15</i>	<i>Reviews</i>	Categoria e/ou Polaridade	815 / 9
BBC	BBC	Notícias	Tópico	2.225 / 5

Fonte: Dados da pesquisa.

Nota – Preparação baseada em enriquecimento textual e divisão de sentenças.

Observação – Os valores entre parênteses se referem aos atributos originais das bases de dados.

¹ <<http://sites.labic.icmc.usp.br/rsinoara/bestsports>>

² <<http://alt.qcri.org/semeval2015/task12>>

³ <<http://mlg.ucd.ie/datasets/bbc>>

Conforme apresentado na Tabela 1, os textos das bases de dados selecionadas podem ser classificados por Tópico ou Categoria (p. ex., nome de um esporte ou categoria de um produto), por Semântica ou Polaridade (p. ex., o resultado de um país em uma competição esportiva ou uma opinião sobre um produto) ou por Tópico e Semântica ou Categoria e Polaridade, tarefas de permeiam respectivamente o 1º nível, o 2º nível e os 1º e 2º níveis de complexidade semântica (vide a Subseção 2.3.1).

Recapitulando as duas hipóteses de pesquisa definidas nesta dissertação e apresentadas na Seção 1.3, tem-se que a **Hipótese 1**, verificada por meio da avaliação empírica da técnica CIRT.5, diz que “as representações de textos baseadas em atributos caracterizados por contextos (conjuntos de termos correlacionados) contribuem para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos”. Já a **Hipótese 2** diz que “o enriquecimento semântico dos atributos das representações de textos contribui para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos”. Nesse caso, para verificar a Hipótese 2 as bases de dados selecionadas foram analisadas nos seguintes cenários de aplicação considerando a delimitação das sentenças como o escopo de interesse:

- **Versão “raw” das coleções** – Textos formados por termos simples ou compostos delimitados por espaços em branco ou quebras de linha (p. ex., “Inteligência” e “Computacional”);
- **Versão “word” das coleções** – Textos formados por termos simples ou compostos que representam conceitos ou entidades nomeadas (p. ex., “Inteligência Computacional”);
- **Versão “id” das coleções** – Textos formados por *synset ids* que representam conceitos ou entidades nomeadas desambiguadas (p. ex., “bn:00787812n”).

A análise de representações de textos geradas a partir de diferentes versões de cada coleção selecionada foi motivada para verificar o impacto causado por diferentes procedimentos de enriquecimento textual na performance de algoritmos supervisionados de AM. A versão “raw” das coleções foi considerada com o objetivo de avaliar a efetividade das abordagens de representação de textos em cenários sem enriquecimento textual, enquanto as versões “word” e “id” foram consideradas com o objetivo de avaliar a efetividade de representações de textos geradas a partir de textos enriquecidos. Embora as versões “word” e “id” das coleções possam apresentar termos com correspondência direta, na versão “id” termos como “Inteligência Artificial” e “IA” podem ser associados por um mesmo *synset id*, ou seja, pelo identificador único do sentido associado a um conjunto de termos sinônimos denominados *synsets* (vide a Subsubseção 2.3.2.2).

Em ambas as versões “word” e “id” foi utilizado o *script S-Enrich.py* apresentado na Subseção 3.4.1, executado com os parâmetros padrões da API Babelfy tanto do tipo de correspondência (exata) quanto do limiar de similaridade de desambiguação (0,7), considerando para o reconhecimento de um termo o maior segmento contínuo identificado (vide a Subseção 3.3.1).

Portanto, considerando os níveis de complexidade semântica e os tipos de preparação de textos explorados, foram geradas 21 coleções de *benchmark* a partir das bases de dados selecionadas (vide a Seção A.1).

Seguindo as recomendações da técnica CIRT.5 para intensificar a identificação de correlações latentes entre os textos (vide a Subseção 3.3.3), foram utilizados *backups* textuais da Wikipédia⁴ como textos de fontes externas de conhecimento utilizados para o pré-treinamento dos modelos de linguagem baseados em *word embeddings*. Esses *backups* foram individualmente extraídos e processados⁵ para ambos os idiomas tratados (vide a Seção A.1).

4.2.2 Abordagens de representação de textos avaliadas

Para comparar a técnica CIRT.5 com algumas das principais técnicas de representação de textos utilizadas na literatura (vide a Subseção 2.4.2), foram avaliadas ao todo 6 abordagens de representação de textos. Além das tradicionais técnicas BoW e LDA, que normalmente permitem explorar apenas informações locais (L) associadas ao vocabulário original das coleções, também foi avaliada a técnica NASARI_{embed} por meio do modelo *Embed* (Wiki)⁶, e a técnica CIRT.5_{embed} explorada também por meio de informações externas (E-L) das coleções.

As representações de textos baseadas em *word embeddings* e apenas em informações externas (E) das coleções foram geradas por meio da abordagem *Bag of Vectors* (BoV), que correlaciona os textos por meio do centroide dos vetores de seus termos presentes nos modelos; essa abordagem de representação de textos é baseada em Sohrab, Miwa e Sasaki (2015). Conforme apresentado na Tabela 2, no caso da técnica NASARI_{embed} essas representações de textos foram denominadas BoV_NASARI. Analogamente, as representações de textos geradas por meio dos modelos pré-treinados da Wikipédia foram denominadas BoV_Wiki (vide a Seção A.2).

As representações de textos do tipo BoW foram geradas por meio do *script* BoW.py⁷, que permite utilizar métricas de atribuição de pesos como a TF-IDF, adotada como padrão. Conforme apresentado na Tabela 2, em todas as representações de textos originadas das versões “raw” e “word” foram utilizadas *stoplists*⁸. Exceto para a execução da abordagem LDA, na versão “raw” das coleções foram aplicados os processos de padronização de caixas (*case-insensitive*), de radicalização⁹ e de validação de termos (filtragem), considerando apenas os termos compostos pelos caracteres do conjunto $\{a - z; A - Z; 0 - 9; -, _; ' \}$. Para a execução das abordagens BoW, CIRT.5 L e CIRT.5 E-L também foram aplicados cortes de frequência mínima baseados no número de textos em que cada termo ocorre, definido como 2.

⁴ Obtidos em 26/09/2017 por meio da seguinte URL: <<https://dumps.wikimedia.org>>.

⁵ <<https://github.com/attardi/wikiextractor>>

⁶ Obtido em 26/05/2018 por meio da seguinte URL: <<http://lcl.uniroma1.it/nasari>>.

⁷ <<https://github.com/joao8tunes/BoW>>

⁸ <<https://www.ranks.nl/stopwords>>

⁹ <<http://www.nltk.org/api/nltk.stem>>

Tabela 2 – Configuração paramétrica das representações de textos avaliadas.

Representação	Versão	Corte	Stoplist	Validação	Ponderação	Informação
BoW	“raw”	≥ 2	Sim	Sim	TF-IDF	L
	“word”	≥ 2	Sim	Não	TF-IDF	L
	“id”	≥ 2	Não	Não	TF-IDF	L
LDA	“raw”		Sim	Sim	Proporção	L
	“word”		Sim	Não	Proporção	L
	“id”		Não	Não	Proporção	L
BoV_NASARI	“raw”		Sim	Sim	Centroide	E
	“word”		Sim	Não	Centroide	E
BoV_Wiki	“raw”		Sim	Sim	Centroide	E
	“word”		Sim	Não	Centroide	E
	“id”		Não	Não	Centroide	E
CIRT.5	“raw”	≥ 2	Sim	Sim	CF-IDF	L
	“raw”	≥ 2	Sim	Sim	CF-IDF	E-L
	“word”	≥ 2	Sim	Não	CF-IDF	L
	“word”	≥ 2	Sim	Não	CF-IDF	E-L
	“id”	≥ 2	Não	Não	CF-IDF	L
	“id”	≥ 2	Não	Não	CF-IDF	E-L

Fonte: Dados da pesquisa.

Para a execução da abordagem LDA foi utilizado o *toolkit* MALLET¹⁰ (versão 2.0.8), formado por um conjunto de bibliotecas de código aberto e acesso livre disponível na linguagem de programação Java. Seguindo a recomendação apresentada em McCallum (2002), para as representações de textos do tipo LDA foram considerados os 300 tópicos mais representativos de cada versão das coleções, baseado na implementação da técnica LDA por meio da amostragem de Gibbs (GEMAN; GEMAN, 1984).

Por motivo de compatibilidade paramétrica do *toolkit* MALLET, exclusivamente na execução da abordagem LDA foram aplicados processos de validação de termos diferentes para a versão “raw” das coleções. Para o idioma inglês foram considerados apenas os termos compostos por caracteres do conjunto $\{a - z; A - Z; -, ' \}$. Para o idioma português, entretanto, além de termos compostos por caracteres do mesmo conjunto considerado para o idioma inglês, também foram considerados caracteres acentuados.

Para a análise exploratória da técnica CIRT.5 foram avaliadas duas abordagens de representação de textos: a abordagem CIRT.5 L, utilizando modelos treinados apenas com o conteúdo local das coleções; e a abordagem CIRT.5 E-L, utilizando modelos pré-treinados por meio de fontes externas de conhecimento. Essas duas abordagens foram executadas por meio do *script* CIRT.5_embed.py apresentado na Subseção 3.4.1. Para a abordagem CIRT.5 E-L os modelos foram gerados a partir de textos da Wikipédia, individualmente treinados e aplicados

¹⁰ <<http://mallet.cs.umass.edu>>

para cada idioma tratado e para cada versão avaliada, totalizando 6 modelos pré-treinados¹¹ (vide o Apêndice A). Baseado em experimentos preliminares, para o pré-treinamento desses modelos foram aplicados cortes de frequência mínima de 10 ocorrências por termo para atenuar possíveis inconsistências linguísticas.

Em ambas as abordagens de representação de textos da técnica CIRT.5 foi utilizado o modelo de linguagem CBoW com 300 dimensões, considerando ainda os valores padrões do *toolkit* Gensim para os tamanhos das janelas deslizantes (escopos em que são aprendidos os contextos) e para as épocas de treinamento dos modelos, ambos definidos como 5. Também foram aplicados cortes de frequência mínima de ocorrência dos termos em pelo menos 2 documentos, e uma variação para contextos da métrica de atribuição de pesos TF-IDF (CF-IDF). Em relação aos principais parâmetros da técnica CIRT.5, foram considerados diferentes tamanhos dos contextos extraídos (conjunto de valores {1;5;10;15;20;25}) e diferentes limiares de comparação da similaridade entre eles (conjunto de valores {0,25;0,5;0,75;1}).

Para as representações de textos geradas por meio da abordagem BoV foi utilizado o *script* BoV.py¹², tendo sido explorados tanto os modelos pré-treinados da Wikipédia quanto o modelo pré-treinado da técnica NASARI_{embed}, este último disponível apenas para o idioma inglês. Para a execução da abordagem BoV_Wiki foram aplicados os 6 modelos pré-treinados da Wikipédia; considerando as versões “raw”, “word” e “id” das coleções analisadas, foram geradas 3 versões de modelos pré-treinados para cada idioma. Já no segundo caso, em razão do modelo NASARI_{embed} ser formado por termos simples e compostos e não por *synset ids*, a abordagem BoV_NASARI foi executada apenas nas versões “raw” e “word” das coleções.

Diferente das outras abordagens de representação de textos avaliadas, cada uma executada com apenas 1 configuração de parâmetros, para cada abordagem CIRT.5 L e CIRT.5 E-L foram testadas 24 combinações dos parâmetros de contextos e de limiares de similaridade. Por conta dessa característica, o número de observações de performance das abordagens associadas à técnica CIRT.5 é consideravelmente maior do que para as demais abordagens, embora não seja um fator para enviesar os resultados dos experimentos realizados.

4.2.3 Algoritmos de classificação aplicados

Conforme abordado na Subseção 2.2.2, a classificação é uma das tarefas mais amplamente difundidas na área de IA, e pode ser realizada por meio de algoritmos baseados em distância como o KNN, em probabilidade como o NB, em procura como o C4.5 ou em otimização como o IMBHN. Para explorar essas diferentes classes de métodos, foram aplicados alguns dos principais algoritmos da literatura por meio da Geral¹³ (versão *MultiThreading* 20180104), uma ferramenta de acesso livre disponível na linguagem de programação Java.

¹¹ Esses modelos também foram utilizados nas representações de textos da abordagem BoV_Wiki.

¹² <<https://github.com/joao8tunes/BoV>>

¹³ <http://sites.labicc.icmc.usp.br/ragero/thesis/text_categorization_tool>

Na lista a seguir são apresentadas as configurações consideradas para a aplicação de cada algoritmo de classificação, ressaltando a quantidade de arquivos de resultado gerados em cada caso. Configurações similares também já foram utilizadas de forma efetiva em outros trabalhos associados ao LABIC, seja na exploração do uso de entidades nomeadas (SINOARA; ROSSI; REZENDE, 2016), de expressões de domínio (SCHEICHER; SINOARA; KOGA; REZENDE, 2016) ou de papéis semânticos (SINOARA, 2018) na tarefa de classificação (ROSSI; FALEIROS; LOPES; REZENDE, 2012).

- **NB** com parâmetros estimados automaticamente – **Gera 1 arquivo**;
- **J48** (baseado no algoritmo C4.5) com nível de confiança de 0,25 – **Gera 1 arquivo**;
- **KNN** com o conjunto de valores $\{1; 3; 5; 7; 9; 11; 13; 15; 17; 19; 25; 35; 45; 55\}$ e utilizando voto com e sem ponderação pela distância euclidiana e pela medida de similaridade do cosseno – **Gera 56 arquivos**;
- **IMBHN^C** e **IMBHN^R** utilizando o conjunto de valores $\{0,01; 0,05; 0,1; 0,5\}$ para taxa de correção de erros, com número máximo de 1.000 iterações e erro dos mínimos quadrados com critério de parada de 0,01 – **Geram 8 arquivos**.

Ao todo foram obtidos 66 arquivos de resultados para cada representação de textos gerada. Os resultados dessas classificações são apresentados na seção a seguir, tanto para as distribuições de Acurácia quanto para as distribuições de F1. Embora os algoritmos de classificação *Multi-layer Perceptron* e SVM sejam frequentemente empregados na literatura, eles não foram utilizados nesta dissertação devido ao tempo e ao custo de processamento inviáveis quando aplicados em representações de textos com alta esparsidade e dimensionalidade. Baseado em experimentos preliminares, em cenários como os piores casos da abordagem CIRT.5 L, por exemplo, a execução dos algoritmos *Multi-layer Perceptron* e SVM levou semanas ao invés de horas para classificar uma única representação de textos.

4.3 Avaliação das abordagens de representação de textos

Em todos os experimentos realizados nesta dissertação, a tarefa de classificação foi avaliada por meio das métricas Acurácia e F1 (utilizando média macro) apresentadas na Subsubseção 2.2.2.2, com ênfase na Acurácia e considerando o método de validação cruzada *k-fold* com $k = 10$, também denominado *10-fold cross validation*. O método *k-fold* divide o conjunto de documentos em k subconjuntos mutualmente exclusivos do mesmo tamanho. Assim, os modelos são sucessivamente criados de maneira alternada a partir de $k - 1$ subconjuntos e testados com o subconjunto restante. Após k iterações as métricas de avaliação da tarefa realizada são calculadas considerando a média obtida nas iterações, o que permite obter medidas mais confiáveis acerca das capacidades dos modelos avaliados (KOHAVI, 1995).

Os resultados obtidos na tarefa de classificação automática de textos estão divididos de acordo com o nível de complexidade semântica considerado em cada coleção. O desempenho das coleções são exibidos de acordo com as distribuições de valores de cada nível de complexidade: 1º nível; 2º nível; e 1º e 2º níveis. Essas distribuições são representadas por gráficos do tipo *boxplot* gerados por meio da linguagem de programação R¹⁴ (versão 3.5.1), formada por um conjunto de bibliotecas estatísticas de código aberto e acesso livre.

Os *boxplots* permitem representar distribuições por meio de diagramas de caixa, que dividem um conjunto de valores em partes proporcionais plotadas em uma reta de extremos. O conjunto de valores observados é dividido em regiões que delimitam as porcentagens de valores da distribuição nos intervalos (quartis) 0-25%, 25-50%, 50-75% e 75-100%. A mediana é representada pela linha interna das caixas plotadas, e os valores fora das retas (*whiskers*) são considerados pontos atípicos discrepantes da tendência dos dados (TUKEY, 1977).

Conforme ressaltado em Bussab e Morettin (2010, p. 47–50), os *boxplots* permitem identificar por meio das caixas plotadas onde estão localizados 50% dos valores mais prováveis das distribuições, que variam entre o 1º e o 3º quartis. No caso dos gráficos apresentados a seguir, os valores das distribuições de Acurácia e de F1 variam respectivamente entre 0 e 100 e entre 0 e 1, relativos à porcentagem da performance de classificação e à razão F1.

Em relação aos resultados da classificação isolada do 1º nível de complexidade semântica, apresentados na Figura 12, foram processadas as coleções de textos BS-4, SE e BBC. Considerando o número total de observações obtidas para cada abordagem de representação de textos tem-se que: para cada abordagem BoW, LDA e BoV_Wiki foram obtidas 594 observações; para cada abordagem CIRT.5 L e CIRT.5 E-L foram obtidas 14.256 observações; e para a abordagem BoV_NASARI foram obtidas 264 observações. Assim, nesses problemas de classificação os melhores resultados para cada abordagem de representação de textos avaliada foi o seguinte:

- **Coleção BS-4:** BoW – “raw”, “word” e “id” (100% e 1 – IMBHN^R e KNN); LDA – “raw” (99,6429% e 0,9957 – IMBHN^R e KNN); BoV_Wiki – “raw” (98,9286% e 0,9898 – KNN); CIRT.5 E-L – “raw”, contexto = 10 e limiar = 0,25 (98,5961% e 0,9870 – IMBHN^R); e CIRT.5 L – “raw”, contexto = 20 e limiar = 0,75 (89,3719% e 0,8883 – IMBHN^R);
- **Coleção SE:** BoW – “raw” (99,6326% e 0,9611 – KNN); BoV_Wiki – “id” (98,5276% e 0,9329 – KNN); LDA – “raw” (96,5673% e 0,9259 – KNN); CIRT.5 E-L – “id”, contexto = 20 e limiar = 0,25 (94,2307% e 0,8290 – IMBHN^R); CIRT.5 L – “word”, contexto = 20 e limiar = 0,75 (80,1220% e 0,5649 – IMBHN^R); e BoV_NASARI – “raw” (73,2415% e 0,5679% – KNN);
- **Coleção BBC:** BoW – “id” (98,0693% e 0,9809 – IMBHN^R); LDA – “raw” (96,5406% e 0,9654 – IMBHN^R); BoV_Wiki – “id” (96,2697% e 0,9631 – KNN); CIRT.5 E-L – “raw”,

¹⁴ <<https://cran.r-project.org>>

contexto = 20 e limiar = 0,25 (95,0119% e 0,9501 – IMBHN^R); BoV_NASARI – “word” (88,6297% e 0,8873 – KNN); e CIRT.5 L – “raw” e contexto = 1 (78,6505% e 0,7860 – IMBHN^R).

Figura 12 – Performance das abordagens de representação de textos: 1º nível.

(a) Distribuições de valores da métrica Acurácia.

(b) Distribuições de valores da métrica F1.

Fonte: Dados da pesquisa.

Em relação aos resultados da classificação isolada do 2º nível de complexidade semântica, ilustrados na Figura 13, e da classificação conjunta do 1º e 2º níveis de complexidade semântica, ilustrados na Figura 14, foram processadas apenas as coleções de textos BS-4 e SE, devido ao fato dos problemas de classificação da coleção BBC permearem somente o 1º nível de complexidade semântica. Considerando o número total de observações obtidas para cada abordagem de representação de textos tem-se que: para cada abordagem BoW, LDA e BoV_Wiki foram obtidas 396 observações; para cada abordagem CIRT.5 L e CIRT.5 E-L foram obtidas 9.504 observações; e para a abordagem BoV_NASARI foram obtidas 132 observações.

Figura 13 – Performance das abordagens de representação de textos: 2º nível.

(a) Distribuições de valores da métrica Acurácia.

(b) Distribuições de valores da métrica F1.

Fonte: Dados da pesquisa.

Os problemas de classificação associados à análise isolada do 2º nível de complexidade semântica são caracterizados pela atribuição de classes a partir de informações semânticas específicas associadas aos textos da coleção. Nesse cenário, os melhores resultados para cada abordagem de representação de textos avaliada foi o seguinte:

- **Coleção BS-4:** BoW – “word” (68,9163% e 0,6915 – IMBHN^R); BoV_Wiki – “word” (66,8719% e 0,6677 – KNN); LDA – “raw” (66,8227% e 0,6631 – IMBHN^R); CIRT.5 E-L – “id”, contexto = 20 e limiar = 0,5 (57,5862% e 0,5892 – IMBHN^R); e CIRT.5 L – “raw”, contexto = 20 e limiar = 0,75 (48,6207% e 0,4513 – IMBHN^R);
- **Coleção SE:** BoW – “raw” (83,6902% e 0,5720 – IMBHN^R); BoV_Wiki – “word” (82,9298% e 0,5464 – IMBHN^R); LDA – “raw” (77,7883% e 0,5017 – KNN); CIRT.5 E-L – “id”, contexto = 20 e limiar = 0,25 (76,0659% e 0,5108 – IMBHN^R); CIRT.5 L – “word”, contexto = 20 e limiar = 0,5 (68,8392% e 0,4067 – KNN); e BoV_NASARI – “word” (67,3592% e 0,3716% – KNN).

Já nos problemas de classificação associados à análise conjunta do 1º e 2º níveis de complexidade semântica, os melhores resultados para cada abordagem de representação de textos avaliada foi o seguinte:

Figura 14 – Performance das abordagens de representação de textos: 1º e 2º níveis.

(a) Distribuições de valores da métrica Acurácia.

(b) Distribuições de valores da métrica F1.

Fonte: Dados da pesquisa.

- **Coleção BS-4:** BoW – “raw” (65,7512% e 0,4994 – IMBHN^C); LDA – “raw” (65,7389% e 0,5006 – KNN); BoV_Wiki – “word” (65,1232% e 0,4615 – KNN); CIRT.5 E-L – “id”, contexto = 20 e limiar = 0,5 (55,8744% e 0,4089 – IMBHN^C); e CIRT.5 L – “raw”, contexto = 25 e limiar = 0,75 (42,0936% e 0,2133 – IMBHN^R);
- **Coleção SE:** BoW – “raw” (79,0108% e 0,4187 – IMBHN^R); BoV_Wiki – “word” (77,9148% e 0,4290 – KNN); LDA – “raw” (75,2123% e 0,4058 – IMBHN^R); CIRT.5 E-L – “word”, contexto = 25 e limiar = 0,25 (68,0819% e 0,2924 – IMBHN^R); CIRT.5 L – “word”, contexto = 25 e limiar = 0,75 (52,7718% e 0,1791 – IMBHN^R); e BoV_NASARI – “word” (46,7465% e 0,1797% – KNN).

Por meio das distribuições de resultados apresentadas pode-se extrair evidências a respeito do desempenho da tarefa de classificação automática de textos em diferentes níveis de complexidade semântica. Assim como apresentado na Figura 12, normalmente os problemas de classificação baseados em Tópico ou Categoria são mais fáceis de serem solucionados do que os problemas baseados em Semântica ou Polaridade ou em Tópico e Semântica ou Categoria e Polaridade. Nesse último caso, os problemas de classificação costumam ser mais complexos e podem apresentar resultados menos efetivos para algoritmos supervisionados de AM, conforme apresentado na Figura 13 e na Figura 14 e avaliado empiricamente em Sinoara (2018).

Para determinar o ranque de desempenho médio das representações de textos em razão das distribuições de Acurácia, foi aplicado¹⁵ o teste não paramétrico de Friedman (FRIEDMAN, 1940) com pós teste de Nemenyi (NEMENYI, 1963), considerando 95% de confiança para verificar a existência de diferenças estatísticas entre as performances das abordagens de representação de textos avaliadas. Conforme ressaltado em Demšar (2006), a aplicação desses testes é recomendável para a avaliação da significância estatística dos resultados de múltiplos classificadores em múltiplas coleções de dados. O autor ainda apresenta uma proposta de representação visual do ranque de desempenho médio calculado, denominada diagrama de distância crítica.

Nos diagramas de distância crítica apresentados na Figura 15, na Figura 16 e na Figura 17, as performances das abordagens de representação de textos estão dispostas de acordo com os níveis de complexidade semântica abordado nas classificações, com escala proporcional ao número de abordagens avaliadas em cada cenário. Em cada diagrama de distância crítica foram considerados os maiores valores de Acurácia de cada abordagem de representação de textos, relativos às 66 configurações de classificação aplicadas (vide o Apêndice B).

Em relação à geração de cada diagrama associado às versões “raw” ou “word” das coleções, foram utilizadas 396 observações relativas às performances de todas as abordagens de representação de textos avaliadas. Todavia, para a geração de cada diagrama associado à versão “id” das coleções foram utilizadas 330 observações, devido ao fato da abordagem BoV_NASARI utilizar um modelo incompatível com esse tipo de versão (vide a Subseção 4.2.2).

¹⁵ <<https://cran.r-project.org/web/packages/scmamp>>

Figura 15 – Ranque das abordagens de representação de textos: 1º nível.

Fonte: Dados da pesquisa.

Figura 16 – Ranque das abordagens de representação de textos: 2º nível.

Fonte: Dados da pesquisa.

Figura 17 – Ranque das abordagens de representação de textos: 1º e 2º níveis.

Fonte: Dados da pesquisa.

Os ranques de desempenho médio apresentados nos diagramas de distância crítica estão organizados em ordem crescente da esquerda para a direita, da performance mais efetiva para a menos efetiva em cada situação analisada. Quando as linhas verticais associadas às abordagens de representação de textos não são unidas por uma linha horizontal em negrito, diz-se que as diferenças de performance são estatisticamente significantes. Caso contrário, diz-se que as diferenças de performance das abordagens unidas não são estatisticamente significantes (DEMŠAR, 2006). Considerando os diferentes níveis de complexidade semântica e as diferentes versões das coleções, pôde-se extrair as seguintes informações sobre as abordagens de representação de textos avaliadas:

- **1º nível (Tópico ou Categoria)** – A abordagem tradicional de representação de textos BoW foi mais efetiva do que as demais abordagens em todas as 3 versões das coleções. Embora não tenham apresentado diferenças estatisticamente significantes, as duas abordagens mais efetivas nesse cenário foram respectivamente a BoW e a BoV_Wiki. Assim como na abordagem BoV_Wiki, aparentemente a performance da técnica CIRT.5 melhora em decorrência do nível de enriquecimento textual aplicado, já que de acordo com os ranques apresentados na Figura 15 houve um aumento da performance das abordagens CIRT.5 L e CIRT.5 E-L nas versões “word” e “id” das coleções, principalmente considerando a abordagem CIRT.5 E-L;

- **2º nível (Semântica ou Polaridade)** – A abordagem de representação de textos BoV_Wiki foi mais efetiva do que as demais abordagens em todas as 3 versões das coleções, embora não tenha apresentado diferenças estatisticamente significantes na versão “raw”. Assim como na abordagem BoV_Wiki, mais uma vez a performance da técnica CIRT.5 melhorou em decorrência do nível de enriquecimento textual aplicado. De acordo com os ranques apresentados na Figura 16, a performance da abordagem CIRT.5 E-L chega a superar a tradicional abordagem BoW na versão “word” das coleções, embora não tenha apresentado diferenças estatisticamente significantes;
- **1º e 2º níveis (Tópico e Semântica ou Categoria e Polaridade)** – A abordagem de representação de textos BoV_Wiki novamente foi mais efetiva do que as demais abordagens em todas as 3 versões das coleções, embora não tenha apresentado diferenças estatisticamente significantes. Na versão “raw” das coleções as abordagens mais efetivas foram respectivamente a BoV_Wiki, a BoW e a LDA, que não apresentaram diferenças estatisticamente significantes. De acordo com os ranques apresentados na Figura 17, outra vez houve melhora nas performances das abordagens BoV_Wiki e CIRT.5 considerando as versões “word” e “id” das coleções.

No que se refere aos algoritmos de classificação mais apropriados para as abordagens de representação de textos avaliadas, não foram identificadas evidências específicas para cada caso. Entretanto, de maneira geral os resultados mais expressivos em todas as abordagens avaliadas foram obtidos por meio dos algoritmos KNN, IMBHN^C e IMBHN^R, assim como salientado na descrição dos melhores resultados obtidos apresentada no início desta seção.

Baseado nos diagramas de distância crítica apresentados, os resultados mais expressivos da abordagem **BoW** foram obtidos no 1º nível de complexidade semântica por meio da versão “raw” das coleções, ilustrado na Figura 15. Nesse cenário de classificação o desempenho de algoritmos supervisionados de AM normalmente é efetivo considerando apenas a coocorrência dos termos, tratados como atributos independentes na técnica BoW (vide a Figura 12).

Os resultados mais expressivos da abordagem **LDA** foram obtidos no 2º nível de complexidade semântica por meio da versão “raw” das coleções (vide a Figura 16). Assim como retratado na Figura 12, na Figura 13 e na Figura 14, a efetividade da técnica LDA normalmente está associada à exploração de relacionamentos latentes entre os termos, que pode ser prejudicada por procedimentos de enriquecimento textual como nas versões “word” e “id” das coleções.

Nas abordagens similares **BoV_NASARI** e **BoV_Wiki** os resultados mais expressivos foram respectivamente no 1º e no 2º nível de complexidade semântica por meio da versão “word” das coleções. Ao contrário do que ocorre na abordagem LDA, aparentemente a performance das abordagens BoV_NASARI e BoV_Wiki melhora em decorrência do nível de enriquecimento textual aplicado, embora por motivo de incompatibilidade não tenha sido possível avaliar a performance do modelo *Embed* (Wiki) da técnica NASARI_{embed} na versão “id” das coleções.

Em relação às abordagens da técnica **CIRT.5**, os resultados mais expressivos foram obtidos no 2º nível de complexidade semântica por meio da versão “word” das coleções. Segundo o ranque das abordagens CIRT.5 L e CIRT.5 E-L em todos os diagramas de distância crítica apresentados, a performance da técnica CIRT.5 tende a melhorar quando considerada a aplicação de enriquecimentos textuais e a inclusão de textos de fontes externas de conhecimento no pré-treinamento dos modelos de linguagem. Essas características podem ser melhor observadas na abordagem CIRT.5 E-L aplicada nas versões “word” e “id” das coleções.

Na subseção a seguir são apresentadas as propriedades da técnica CIRT.5 evidenciando sua performance nos diferentes idiomas tratados. Segundo os resultados das correlações dos principais parâmetros da técnica proposta, há indícios de que os valores dos limiares de similaridade são inversamente proporcionais ao desempenho das representações de textos considerando as observações da métrica Acurácia.

4.4 Propriedades dos parâmetros da técnica CIRT.5

Conforme apresentado na Seção 3.3, a implementação da técnica CIRT.5 está associada a diversas variáveis como os tipos de enriquecimento textual aplicados, a delimitação dos escopos de interesse, a utilização de fontes externas de conhecimento e aos tipos de modelos de linguagem utilizados. Entretanto, os principais parâmetros dessa técnica são os tamanhos dos contextos extraídos e os limiares de similaridade entre contextos empregados como atributos na representação de textos. Para avaliar as relações entre esses parâmetros e os valores de Acurácia e de F1 observados, foram calculadas as correlações dos parâmetros das abordagens CIRT.5 L e CIRT.5 E-L para todas as coleções analisadas, conforme apresentado na Figura 18.

Utilizados para medir o grau e a direção com que duas variáveis numéricas tendem a mudar juntas, os coeficientes de correlação podem assumir valores entre -1 e 1. Quando o valor da correlação é igual a -1 tem-se uma correlação negativa perfeita entre as duas variáveis, ou seja, se o valor de uma variável aumenta o valor da outra diminui. Analogamente, quando o valor da correlação é igual a 1 tem-se uma correlação positiva perfeita entre as duas variáveis. Todavia, quando o valor da correlação é igual a 0 diz-se que as variáveis são independentes (BUSSAB; MORETTIN, 2010, p. 261–295).

Para verificar a normalidade das distribuições de Acurácia das abordagens CIRT.5 L e CIRT.5 E-L, suposta como pré-requisito para o cálculo de coeficientes de correlação como ρ de Pearson (PEARSON, 1901) e τ de Spearman (SPEARMAN, 1904), foi aplicado¹⁶ o teste de normalidade Shapiro-Wilk (SHAPIRO; WILK, 1965). Entretanto, de acordo com os resultados obtidos nenhuma das distribuições passou no teste de normalidade¹⁷.

¹⁶ <<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html>>

¹⁷ Todas as distribuições apresentaram valor de $p = 2,2 * 10^{-16}$ para amostras de 5.000 observações. Logo, como $p < 0,05$ então as distribuições não são normais à um nível de significância de 5%.

Embora as distribuições de Acurácia e de F1 das abordagens CIRT.5 L e CIRT.5 E-L não tenham passado no teste de normalidade Shapiro-Wilk, de acordo com o Teorema Limite Central, sabe-se que distribuições amostrais suficientemente grandes formadas por mais de 30 observações aproximam-se de uma distribuição normal (BUSSAB; MORETTIN, 2010, p. 277–281). Baseado nesse argumento, para analisar as propriedades dos parâmetros da técnica CIRT.5 foram calculados os coeficientes ρ de Pearson e τ de Spearman, já que as distribuições para os idiomas português (coleção BS-4) e inglês (coleções SE e BBC) apresentam respectivamente 28.512 e 38.016 observações de resultados.

Assim como ressaltado em Mukaka (2012), os métodos de Pearson e de Spearman podem ser considerados complementares por permitem analisar aspectos diferentes das correlações. Segundo o autor, os coeficientes ρ de Pearson são mais sensíveis à pontos atípicos (*outliers*) nos extremos das distribuições do que os coeficientes τ de Spearman, isso porque enquanto o primeiro permite avaliar relações lineares proporcionais, o segundo permite avaliar relações monotônicas não necessariamente proporcionais e/ou constantes.

Os coeficientes de correlação gerados a partir das abordagens CIRT.5 L e CIRT.5 E-L podem ser visualizados na Figura 18, em que pode ser comparada as diferenças de correlação para os idiomas português (PT) e inglês (EN).

Figura 18 – Correlações dos parâmetros da técnica CIRT.5: contextos e limiares.

(a) Coeficientes de correlação de Pearson.

(b) Coeficientes de correlação de Spearman.

Fonte: Dados da pesquisa.

Para o cálculo dos coeficientes de correlação também foi utilizada a linguagem de programação R¹⁸ e, tanto na Figura 18a quanto na Figura 18b as correlações foram geradas com um nível de confiança de 95%. Conforme ressaltado em Cohen (1992), a magnitude dos coeficientes de correlação pode ser caracterizada como pequena (entre 0,1 e 0,29), como média (entre 0,3 e 0,49) ou como grande (entre 0,5 e 1).

Em ambos os correlogramas da Figura 18, as variáveis em destaque “Contextos” e “Limiares” referem-se respectivamente aos tamanhos dos contextos e aos limiares de similaridade da técnica CIRT.5. As demais variáveis referem-se à performance da abordagem CIRT.5 na tarefa de classificação automática de textos, sendo que “Acurácia” e “F1” referem-se respectivamente aos valores das métricas Acurácia e F1, e “Construção” e “Classificação” referem-se respectivamente aos tempos gastos para a construção dos classificadores na etapa de treinamento, e para a tarefa de classificação em si na etapa de teste¹⁹. Por meio desses correlogramas pôde-se extrair as seguintes informações:

- **“Contextos”** – Aparentemente há uma pequena correlação positiva entre o tamanho dos contextos e o tempo de classificação das representações de textos do idioma português, já que em todos os correlogramas o par “Contextos-Classificação” apresentou valores entre 0,2 e 0,25; para o idioma inglês essa correlação só foi observada na abordagem CIRT.5 L. Na abordagem CIRT.5 E-L identificou-se uma pequena correlação negativa com os valores da Acurácia, já que o par “Contextos-Acurácia” apresentou valores entre -0,14 e -0,15;
- **“Limiares”** – Aparentemente há uma correlação negativa entre os limiares de similaridade e os valores de Acurácia e de F1, o tempo de construção dos classificadores e o tempo de classificação das representações de textos. No caso dos pares “Limiares-Acurácia” e “Limiares-F1”, a magnitude das correlações varia de pequena (entre -0,1 e -0,27) à média (entre -0,34 e -0,49) para ambos os idiomas; essas correlações tornam-se ainda mais evidentes na abordagem CIRT.5 E-L (vide a Figura 19). Em relação ao par “Limiares-Construção”, a magnitude das correlações também varia de pequena (entre -0,13 e -0,29) à média (entre -0,41 e -0,44) para ambos os idiomas, tornando-se ainda mais evidente na abordagem CIRT.5 E-L. Já no par “Limiares-Classificação” a magnitude das correlações varia de média (entre -0,27 e -0,47) à grande (entre -0,59 e -0,74) para ambos os idiomas;
- **“Acurácia”** – Aparentemente há uma correlação positiva entre a métrica Acurácia, a métrica F1 e os tempos de construção dos classificadores e da classificação das representações de textos. No caso do par “Acurácia-F1”, a magnitude das correlações é grande para ambos os idiomas, já que em todos os correlogramas os valores variaram entre 0,64 e 0,97. Em relação aos pares “Acurácia-Construção” e “Acurácia-Classificação”, possivelmente a

¹⁸ <<https://cran.r-project.org/web/packages/corrplot/>>

¹⁹ Os experimentos de classificação foram executados com os mesmos recursos e configurações de *hardware* e de *software*, não relatados nesta dissertação devido ao escopo do projeto.

magnitude das correlações varia de pequena (entre 0,19 e 0,26) à média (entre 0,3 e 0,43) para ambos os idiomas;

- **“F1”** – Aparentemente há uma correlação positiva entre a métrica F1 e os tempos de construção dos classificadores e da classificação das representações de textos. Nos pares “F1-Construção” e “F1-Classificação” possivelmente a magnitude das correlações varia de pequena (entre 0,11 e 0,25) à média (entre 0,3 e 0,45) para ambos os idiomas, embora só seja evidenciada por meio dos coeficientes de correlação de Spearman;
- **“Construção”** – Possivelmente há uma correlação positiva, de magnitude entre pequena e média, entre o tempo de construção dos classificadores e o tempo de classificação das representações de textos, já que as coeficientes variam entre -0,11 e 0,4 para o idioma português e entre -0,01 e 0,41 para o inglês.

Figura 19 – Performance das abordagens CIRT.5 em razão dos limiares de similaridade.

(a) Distribuições de valores da métrica Acurácia.

(b) Distribuições de valores da métrica F1.

Fonte: Dados da pesquisa.

Entre as propriedades dos parâmetros da técnica CIRT.5, a visualização da correlação entre a Acurácia e os limiares de similaridade também é perceptível por meio de *boxplots*. Conforme ilustrado na Figura 19, cujos gráficos foram gerados a partir de todos as 33.264 observações de resultados para cada abordagem CIRT.5 L e CIRT.5 E-L, quanto menores os limites mínimos de similaridade entre os contextos, maiores são os valores de Acurácia obtidos na tarefa de classificação. Embora tal propriedade tenha sido observada em ambas as abordagens, essa correlação é mais evidente nos resultados da abordagem CIRT.5 E-L, já que a mediana das distribuições de Acurácia e de F1 diminuem à medida que os limiares de similaridade aumentam.

Baseado nos correlogramas apresentados na Figura 18, quanto menores os limiares de similaridade maiores os tempos da tarefa de classificação automática de textos, pois as representações de textos geradas se tornam menos esparsas em razão do volume de similaridades consideradas entre os atributos. Entretanto, de acordo com os coeficientes de correlação calculados quanto maiores os tempos de classificação maiores os valores de Acurácia. Aparentemente os valores de Acurácia aumentam de maneira proporcional com os valores de F1, indicando que a técnica CIRT.5 apresenta a característica conjunta de efetividade e cobertura.

Apesar das abordagens CIRT.5 L e CIRT.5 E-L não terem se destacado como as mais efetivas, seus resultados permitem extrair informações importantes sobre o estudo da semântica no processo de MT. De acordo com os experimentos realizados, há evidências estatísticas de que representações de textos baseadas em contextos podem contribuir para uma melhora na performance de algoritmos supervisionados de AM. Aparentemente, atributos caracterizados por meio de conjuntos de termos correlacionados compartilham similaridades que permitem aumentar a eficiência de algoritmos de classificação, principalmente em problemas em que há a necessidade de se analisar os significados dos termos nos textos. Também foram obtidos indícios de que a utilização de fontes externas de conhecimento, tanto para o enriquecimento textual das coleções quanto para a identificação dos atributos das representações de textos, pode contribuir para um aumento significativo da performance de algoritmos de classificação.

4.5 Considerações finais

Neste capítulo foram apresentadas evidências promissoras a respeito da efetividade da técnica CIRT.5 na tarefa de classificação automática de textos. De acordo com os resultados obtidos, aparentemente há uma melhora na eficiência de algoritmos supervisionados de AM quando são consideradas informações semânticas e contextuais na geração de representações de textos. Os resultados mais expressivos das abordagens baseadas na técnica CIRT.5 foram justamente no cenário para o qual ela foi proposta, em problemas associados ao 2º nível de complexidade semântica. No capítulo seguinte são retomadas as motivações que levaram ao desenvolvimento desta dissertação e apresentadas as principais contribuições científicas realizadas, seguidas dos trabalhos e pesquisas pretendidas.

CONCLUSÕES E PERSPECTIVAS FUTURAS

Neste capítulo são apresentadas e discutidas as conclusões obtidas acerca do problema de aplicação e de enriquecimento de representações de textos em diferentes níveis de complexidade semântica. Primeiramente são retomadas as hipóteses e objetivos inicialmente definidos, ressaltando os resultados da avaliação empírica que auxiliaram nessas verificações; essas conclusões representam as principais contribuições desta dissertação. Em seguida são discutidas as aplicações e limitações da técnica de representação de textos proposta, evidenciando sua efetividade em diferentes tarefas e cenários de aplicação. Por fim, também são apresentados e discutidos os próximos trabalhos e pesquisas pretendidas.

5.1 Considerações iniciais

Assim como ressaltado na Seção 1.2, entre as principais motivações para o enriquecimento semântico de representações de textos destaca-se a possibilidade do aumento da performance de algoritmos supervisionados de AM em tarefas da MT. A importância desse tipo de solução baseada em enriquecimento semântico torna-se ainda mais evidente à medida que o volume e a variedade de documentos disponíveis em formato digital aumenta, principalmente em aplicações voltadas para a mineração e análise de coleções de textos (vide a Seção 2.2).

Na seção a seguir são apresentadas e discutidas as principais contribuições científicas obtidas por meio desta dissertação. Essas conclusões foram baseadas na avaliação empírica das abordagens de representação de textos avaliadas na tarefa de classificação automática de textos, exploradas em diferentes níveis de complexidade semântica e considerando diferentes versões das coleções de textos selecionadas. Conforme apresentado na Seção 4.3, há evidências estatísticas de que o enriquecimento semântico e contextual das representações de textos contribui significativamente para o aumento da performance de algoritmos supervisionados de AM, principalmente em tarefas da MT baseadas em informações semânticas dos textos.

5.2 Principais contribuições desta dissertação

O desenvolvimento dos assuntos tratados nesta dissertação foi baseado nas hipóteses e objetivos inicialmente definidos na Seção 1.3 e na Seção 1.4, respectivamente, fundamentados no problema de enriquecimento semântico e contextual de representações de textos para potencializar a performance de algoritmos supervisionados de AM em tarefas da MT. Nesse caso, os experimentos realizados para avaliar as abordagens de representação de textos foram baseados na tarefa de classificação automática de textos, apresentada na Subseção 2.2.2.

Baseado na análise dos resultados obtidos acerca do desempenho das abordagens de representação de textos avaliadas, pode-se extrair as seguintes conclusões acerca das duas hipóteses que fundamentaram o desenvolvimento desta dissertação:

- **Verificação da Hipótese 1** – As representações de textos baseadas em atributos caracterizados por conjuntos de termos correlacionados, como por exemplo os atributos caracterizados por contextos da técnica de representação de textos CIRT.5 (vide a Seção 3.3), podem contribuir para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos;
- **Verificação da Hipótese 2** – O enriquecimento semântico dos atributos das representações de textos, como por exemplo o enriquecimento textual baseado em REN e DLS aplicado nas versões “word” e “id” das coleções de textos selecionadas (vide a Seção 4.2), pode contribuir para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos.

Para evidenciar a aplicabilidade das abordagens de representação de textos avaliadas, antes é preciso considerar os diferentes níveis de complexidade semântica apresentados na Subseção 2.3.1. De modo geral, na tarefa de classificação automática de textos baseada no 1º nível de complexidade (classificação por meio de informações de tópicos ou de categorias) o desempenho mais efetivo foi da técnica BoW, que aparentemente é prejudicada por procedimentos de enriquecimento textual quando aplicada no 2º nível de complexidade semântica (classificação por meio de informações semânticas ou de polaridade). Analogamente, o desempenho da técnica LDA também parece ser prejudicado por procedimentos de enriquecimento textual, embora seus resultados mais expressivos tenham sido obtidos no 2º nível de complexidade semântica.

Na tarefa de classificação automática de textos baseada no 2º nível de complexidade semântica o desempenho mais efetivo foi da abordagem BoV_Wiki, que assim como na abordagem BoV_NASARI e nas abordagens da técnica CIRT.5, foi implementada por meio de modelos baseados em *word embeddings*. A abordagem BoV_Wiki também foi mais efetiva do que as demais abordagens considerando o cenário conjunto do 1º e 2º níveis de complexidade semântica, enquanto os resultados mais expressivos da abordagem BoV_NASARI foram obtidos no 1º nível de complexidade semântica.

De acordo com os resultados obtidos, a performance das abordagens baseadas em *word embeddings* melhora significativamente quando considerada a aplicação de procedimentos de enriquecimento textual. No caso da técnica CIRT.5, os resultados mais expressivos foram obtidos no 2º nível de complexidade semântica, chegando até mesmo a superar a tradicional técnica BoW. Nesse caso, aparentemente a inclusão de textos de fontes externas de conhecimento no pré-treinamento dos modelos de linguagem também melhora significativamente a performance da técnica CIRT.5. Nos resultados individuais das coleções de textos, além das diferenças de desempenho entre as abordagens CIRT.5 L e CIRT.5 E-L também foram identificadas diferenças de performance da técnica CIRT.5 entre os idiomas tratados, sendo que no geral o desempenho nas coleções de textos em inglês foi mais efetivo do que nas coleções de textos em português.

Considerando os melhores resultados em cada nível de complexidade semântica avaliado, os valores de Acurácia e de F1 da abordagem CIRT.5 E-L foram em média cerca de 12% e 17% superiores, respectivamente, à performance da abordagem CIRT.5 L em ambos os idiomas. Com exceção do 1º nível de complexidade semântica, em que a performance da técnica CIRT.5 foi superior para o idioma português, nos outros cenários avaliados os valores de Acurácia e de F1 da técnica CIRT.5 no idioma inglês foram em média cerca de 15% e 10% superiores, respectivamente, à performance no idioma português.

A diferença de performance entre as abordagens CIRT.5 L e CIRT.5 E-L aparentemente está relacionada ao uso efetivo de fontes externas de conhecimento no enriquecimento das representações de textos, e consequentemente no aumento da performance de algoritmos supervisionados de AM. Já a diferença de performance da técnica CIRT.5 entre os idiomas português e inglês pode estar relacionada às diferenças consideráveis entre os volumes das coleções de textos e das fontes externas utilizadas em cada idioma, isto é, nas diferenças em relação à quantidade de documentos de texto de cada fonte textual (vide o Apêndice A).

De acordo com os coeficientes das correlações paramétricas da técnica CIRT.5 apresentados na Seção 4.4, aparentemente há uma correspondência proporcional inversa entre os limiares de similaridade entre contextos e os valores das métricas Acurácia e F1. O desempenho das abordagens CIRT.5 L e CIRT.5 E-L indicam que quanto menor o limiar de similaridade maior é a performance da técnica CIRT.5, o que dá indícios de que representações de textos caracterizadas por atributos dependentes que compartilham similaridades por meio de termos correlacionados contribuem para uma melhora na performance de algoritmos supervisionados de AM na tarefa de classificação automática de textos.

Embora no geral tenho sido menos efetiva do que a abordagem BoV_Wiki na tarefa de classificação automática de textos, a técnica CIRT.5 possui a vantagem de produzir representações de textos mais interpretáveis independente do tipo de modelo de linguagem utilizado. Em contraste com os modelos baseados em *word embeddings*, por exemplo, ao invés de dimensões abstratas os atributos da técnica CIRT.5 são caracterizados por contextos formados por conjuntos de termos correlacionados que podem ser analisados por qualquer indivíduo.

Em relação ao **Objetivo 1 (teórico)** associado à contribuição para a fundamentação teórica do uso de fontes externas de conhecimento para o enriquecimento semântico e contextual de representações de textos, pode-se destacar as duas seguintes abordagens analisadas por meio da tarefa de classificação automática de textos:

- Aplicação de procedimentos de enriquecimento textual baseados em REN e em DLS por meio da rede semântica BabelNet;
- Aplicação de textos da Wikipédia tanto no pré-treinamento dos modelos da técnica CIRT.5 (abordagem CIRT.5 E-L) quanto na abordagem de representação de textos BoV_Wiki.

Em relação ao **Objetivo 2 (prático)** associado à proposta, desenvolvimento e avaliação de técnicas não supervisionadas de representação de textos que exploram informações semânticas e contextuais, pode-se destacar as seguintes abordagens de representação de textos, baseadas na técnica proposta CIRT.5, analisadas por meio da tarefa de classificação automática de textos:

- Abordagem CIRT.5 L com modelos treinados apenas a partir das coleções analisadas;
- Abordagem CIRT.5 E-L com modelos pré-treinados a partir de fontes externas de conhecimento (Wikipédia) e retreinados a partir das coleções analisadas.

Em relação ao **Objetivo 3 (aplicacional)** associado à aplicação das técnicas desenvolvidas em diferentes cenários de representação de textos, explorando possíveis variações de performance entre os idiomas inglês e português em diferentes bases de dados e comparando os resultados obtidos com soluções existentes na literatura, pode-se destacar as seguintes abordagens analisadas por meio da tarefa de classificação automática de textos:

- Avaliação das abordagens de representação de textos BoW, LDA, CIRT.5 L, CIRT.5 E-L, BoV_Wiki e BoV_NASARI em diferentes níveis de complexidade semântica;
- Utilização de bases de dados textuais dos idiomas português (BS-4) e inglês (SE e BBC);
- Preparação de diferentes versões de coleções de textos (“raw”, “word” e “id”) obtidas a partir das bases de dados selecionadas.

Além dos objetivos principais inicialmente definidos nesta dissertação, também foram alcançados **objetivos secundários** relacionados à implementação e disponibilização de ferramentas ou *scripts* para auxiliar na aplicação de procedimentos de enriquecimento textual e na geração de representações de textos. Assim como todo o conteúdo gerado por meio desta dissertação, os *scripts* listados a seguir também se encontram disponíveis no *site* do LABIC¹.

¹ <http://sites.labic.icmc.usp.br/MSc-Thesis_Antunes_2018>

- Em relação à aplicação de procedimentos de enriquecimento textual, à criação de modelos de linguagem baseados em *word embeddings* ou à extração de contextos textuais pode-se citar respectivamente os *scripts* `S-Enrich.py`, `Wiki2Model.py` e `Wiki2Model_S-Enrich.py` apresentados na Subseção 3.4.1 e o *script* `CIET.5_embed.py` apresentado na Subseção 3.4.2;
- Em relação às abordagens de representação de textos pode-se citar o *script* `CIRT.5_embed.py` utilizado nas abordagens CIRT.5 L e CIRT.5 E-L, também apresentado na Subseção 3.4.1, e os *scripts* `BoW.py`, utilizado na abordagem BoW, e `BoV.py`, utilizado nas abordagens BoV_Wiki e BoV_NASARI, apresentados na Subseção 4.2.2.

Na seção a seguir são discutidas as aplicações e limitações da técnica de representação de textos proposta. Além da aplicação principal da técnica CIRT.5 na representação de coleções de textos, também são discutidas as limitações relacionadas à aplicação dessa técnica na extração de contextos, isto é, da técnica de extração de contextos textuais CIET.5 (vide a Subseção 3.4.2).

5.3 Aplicações e limitações das técnicas propostas

Embora ambas as técnicas propostas nesta dissertação tenham apresentado resultados promissores em diferentes áreas de atuação, tanto a técnica CIRT.5 na área de MT quanto a técnica CIET.5 na área de Sistemas de Recomendação (SUNDERMANN; ANTUNES; DOMINGUES; REZENDE, 2018), suas aplicações foram exploradas considerando somente a delimitação de escopos baseada nas sentenças dos textos. Portanto, como a delimitação do escopo de interesse define o âmbito em que serão treinados os modelos de linguagem e consequentemente extraídos os contextos, é natural que textos com sentenças muito curtas ou com sentenças formadas por termos pouco representativos prejudiquem a qualidade do processo, além do fato de que o processo de tokenização da API NLTK apresentada na Subseção 3.4.1 pode estar sujeito a inconsistências.

Além da limitação exploratória relacionada à delimitação dos escopos de interesse, o processo de preparação de textos também impacta na qualidade dos procedimentos realizados por ambas as técnicas propostas. Em relação à técnica de representação de textos CIRT.5, por exemplo, o desempenho inferior na versão “id” em comparação com a versão “word” das coleções pode estar associado à escolha dos parâmetros e/ou das abordagens dos processos de REN e de DLS, já que foram considerados somente os valores padrão da API Babelfy (vide a Subseção 4.2.1).

No que se refere à aplicação de fontes externas de conhecimento para o pré-treinamento dos modelos de linguagem, em ambas as técnicas propostas também devem ser levadas em consideração tanto o volume quanto a qualidade dos textos. No caso da técnica de representação de textos CIRT.5, por exemplo, apesar de terem sido realizados os mesmos procedimentos de

preparação de textos para as coleções analisadas e para as fontes externas utilizadas, o volume de documentos da Wikipédia é consideravelmente maior do que o volume de documentos das coleções. Consequentemente, é possível que mesmo após o retreinamento a partir de cada coleção analisada, os modelos não sofram diferenças significativas, o que pode enviesar as representações de textos geradas com base no conteúdo das fontes externas de conhecimento utilizadas.

Na seção a seguir são apresentados e discutidos os temas de possíveis trabalhos futuros, que também englobam algumas das limitações das técnicas propostas nesta dissertação. Ainda que limitações técnicas como a qualidade do processo de tokenização ou a escolha dos parâmetros dos procedimentos de enriquecimento textual aplicados possa influenciar na qualidade das representações de textos obtidas, os trabalhos e pesquisas apresentados a seguir dizem respeito à aplicabilidade das técnicas propostas de modo geral.

5.4 Trabalhos e pesquisas pretendidas

Em relação às propriedades paramétricas das técnicas CIRT.5 e CIET.5, há uma série de variações que podem ser exploradas em trabalhos futuros, como por exemplo: a consideração de diferentes escopos de interesse baseados em parágrafos ou nos textos integrais dos documentos da coleção analisada; a consideração de diferentes tamanhos de contextos e de limiares de similaridade próximos de 0, menores do que 0,25; e a consideração de abordagens de atribuição de pesos no retreinamento dos modelos de linguagem a partir dos textos da coleção analisada.

No caso das aplicações das técnicas propostas nesta dissertação, também há uma série de possibilidades que podem ser explorados em trabalhos futuros, como por exemplo: a investigação do enriquecimento semântico e contextual das representações de textos em razão dos contextos extraídos por meio da técnica CIET.5; a investigação do aumento da performance de algoritmos de AM em razão da expansão dos atributos de representações de textos (*multi-view ensemble learning*) por meio da técnica CIRT.5; e a investigação da aplicação de diferentes modelos de linguagem na performance de ambas as técnicas propostas.

5.5 Considerações finais

Assim como abordado neste capítulo, há evidências favoráveis a respeito da performance e da aplicabilidade das técnicas propostas nesta dissertação. Portanto, as técnicas CIET.5 e CIRT.5 surgem respectivamente como alternativas efetivas para a extração de contextos textuais e para a representação de textos. Embora possibilitem a agregação conjunta de informações semânticas e contextuais por meio de modelos baseados em *word embeddings*, por exemplo, ambas as técnicas podem ser caracterizadas pela legibilidade das soluções obtidas, auxiliando tanto na melhora da performance de algoritmos supervisionados de AM em diferentes tarefas da MT quanto na interpretabilidade dessas soluções por especialistas do domínio analisado.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. A Survey of Text Classification Algorithms. In: **Mining Text Data**. [S.l.]: Springer, 2012. cap. 6, p. 163–222. Citado nas páginas 33 e 34.

_____. An Introduction to Text Mining. In: **Mining Text Data**. [S.l.]: Springer, 2012. cap. 1, p. 1–10. Citado nas páginas 21, 22, 23, 30, 46, 48, 51 e 52.

AGIRRE, E.; SOROA, A. Personalizing PageRank for Word Sense Disambiguation. In: **Conference of the European Chapter of the Association for Computational Linguistics (EACL)**. [S.l.: s.n.], 2009. p. 33–41. Citado na página 55.

AGIRRE, E.; STEVENSON, M. Knowledge Sources for WSD. In: AGIRRE, E.; EDMONDS, P. (Ed.). **Word Sense Disambiguation: Algorithms and Applications**. [S.l.]: Springer, 2007. cap. 8, p. 217–251. Citado na página 55.

AHO, A. V.; LAM, M. S.; SETHI, R.; ULLMAN, J. D. **Compilers: Principles, Techniques, and Tools**. [S.l.]: Addison-Wesley, 2006. Citado na página 46.

ALJABER, B.; MARTINEZ, D.; STOKES, N.; BAILEY, J. Improving MeSH Classification of Biomedical Articles Using Citation Contexts. **Journal of Biomedical Informatics**, v. 44, n. 5, p. 881–896, 2011. Citado na página 54.

ALMEIDA, T. A.; SILVA, T. P.; SANTOS, I.; HIDALGO, J. M. G. Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering. **Knowledge-Based Systems**, v. 108, p. 25–32, 2016. Citado na página 53.

ALTMAN, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992. Citado na página 34.

ANTHONY, E. M. Approach, Method and Technique. **English Language Teaching (ELT)**, v. 17, n. 2, p. 63–67, 1963. Citado na página 105.

BAKER, C. **Foundations of Bilingual Education and Bilingualism**. [S.l.]: Multilingual Matters, 2001. Citado na página 29.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't Count, Predict! A Systematic Comparison of Context-Counting Vs. Context-Predicting Semantic Vectors. In: **Annual Meeting of the Association for Computational Linguistics (ACL)**. [S.l.: s.n.], 2014. v. 1, p. 238–247. Citado nas páginas 48 e 55.

BARONI, M.; ZAMPARELLI, R. Nouns Are Vectors, Adjectives Are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In: **Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2010. p. 1183–1193. Citado na página 39.

BECHARA, E. **Moderna Gramática Portuguesa**. [S.l.]: Nova Fronteira, 2012. Citado na página 37.

- BEEL, J.; GIPP, B.; LANGER, S.; BREITINGER, C. Research-Paper Recommender Systems: A Literature Survey. **International Journal on Digital Libraries**, v. 17, n. 4, p. 305–338, 2016. Citado na página 32.
- BERTIN, M.; ATANASSOVA, I. Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis. **D-Lib Magazine**, v. 18, n. 7–8, 2012. Citado na página 53.
- BICK, E. **The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. [S.l.]: Aarhus University, 2000. Citado nas páginas 47 e 55.
- BIEMANN, C.; FARALLI, S.; PANCHENKO, A.; PONZETTO, S. P. A Framework for Enriching Lexical Semantic Resources with Distributional Semantics. **Natural Language Engineering**, v. 24, n. 2, p. 265–312, 2018. Citado na página 55.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **Journal of Machine Learning Research (JMLR)**, v. 3, p. 993–1022, 2003. Citado nas páginas 33, 45 e 48.
- BORDES, A.; GLOT, X.; WESTON, J.; BENGIO, Y. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. In: **International Conference on Artificial Intelligence and Statistics (AISTATS)**. [S.l.: s.n.], 2012. p. 127–135. Citado na página 54.
- BROWN, P. F.; PIETRA, V. J. D.; SOUZA, P. V. de; LAI, J. C.; MERCER, R. L. Class-Based N-gram Models of Natural Language. **Computational Linguistics**, v. 18, n. 4, p. 467–479, 1992. Citado na página 52.
- BUSSAB, W. de O.; MORETTIN, P. A. **Estatística Básica**. [S.l.]: Saraiva, 2010. Citado nas páginas 74, 82 e 83.
- CAMACHO-COLLADOS, J.; PILEHVAR, M. T.; NAVIGLI, R. NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities. **Artificial Intelligence**, v. 240, p. 36–64, 2016. Citado nas páginas 45, 48 e 49.
- CAMPOS, R.; MANGARAVITE, V.; PASQUALI, A.; JORGE, A. M.; NUNES, C.; JATOWT, A. A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In: **European Conference on Advances in Information Retrieval Research (ECIR)**. [S.l.: s.n.], 2018. p. 684–691. Citado na página 54.
- CANÇADO, M. **Manual de semântica: noções básicas e exercícios**. [S.l.]: UFMG, 2008. Citado na página 38.
- CASTELLVÍ, M. T. C. Terminology and Translation. In: GAMBIER, Y.; DOORSLAER, L. van (Ed.). **Handbook of Translation Studies**. [S.l.]: John Benjamins, 2010. v. 1, p. 356–365. Citado na página 105.
- CAVNAR, W. B.; TRENKLE, J. M. N-Gram-Based Text Categorization. **Annual Symposium on Document Analysis and Information Retrieval (SDAIR)**, v. 48113, n. 2, p. 161–175, 1994. Citado na página 52.
- CHANG, H.; YANG, M.; YANG, J. Learning a Structure Adaptive Dictionary for Sparse Representation Based Classification. **Neurocomputing**, v. 190, p. 124–131, 2016. Citado na página 53.

CHOMSKY, N. Three Models for the Description of Language. **IRE Transactions on Information Theory**, v. 2, n. 3, p. 113–124, 1956. Citado na página 46.

_____. On Certain Formal Properties of Grammars. **Information and Control**, v. 2, n. 2, p. 137–167, 1959. Citado na página 37.

Cisco Systems. **The Zettabyte Era: Trends and Analysis**. 2017. Cisco Systems web page. Citado na página 30.

COECKE, B.; SADRZADEH, M.; CLARK, S. Mathematical Foundations for a Compositional Distributional Model of Meaning. **Computing Research Repository (CoRR)**, 2010. Cornell University Library. Citado na página 39.

COHEN, J. A Power Primer. **Psychological Bulletin**, v. 112, n. 1, p. 155–159, 1992. Citado na página 84.

COLLOBERT, R.; WESTON, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2008. p. 160–167. Citado na página 40.

CONDORI, R. E. L.; PARDO, T. A. S. Experiments on Sentence Boundary Detection in User-Generated Web Content. In: **International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)**. [S.l.: s.n.], 2015. p. 227–237. Citado na página 62.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. Citado na página 34.

CRAIN, S. P.; ZHOU, K.; YANG, S.-H.; ZHA, H. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In: AGGARWAL, C. C.; ZHAI, C. (Ed.). **Mining Text Data**. [S.l.]: Springer, 2012. cap. 5, p. 129–161. Citado nas páginas 33 e 48.

DANG, E. K. F.; LUK, R. W. P.; ALLAN, J. Beyond Bag-of-Words: Bigram-Enhanced Context-Dependent Term Weights. **Journal of the Association for Information Science and Technology (JASIST)**, v. 65, n. 6, p. 1134–1148, 2014. Citado na página 54.

DEMŠAR, J. Statistical Comparisons of Classifiers Over Multiple Data Sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006. Citado nas páginas 78 e 80.

DEY, A. K. Understanding and Using Context. **Personal and Ubiquitous Computing**, v. 5, n. 1, p. 4–7, 2001. Citado na página 23.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. de A. Mineração de Textos. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicações**. [S.l.]: Manole, 2003. cap. 13, p. 337–364. Citado nas páginas 21, 22, 23, 30, 31, 36 e 52.

ECKERSLEY, C. E.; MACAULAY, M. **Brighter Grammar**. [S.l.]: Longman, 1987. v. 1. Citado na página 37.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. de. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. [S.l.]: LTC, 2011. Citado nas páginas 33, 34, 35 e 36.

- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996. Citado na página 22.
- FELLBAUM, C. (Ed.). **WordNet: An Electronic Lexical Database**. [S.l.]: MIT, 1998. Citado nas páginas 45 e 53.
- FILHO, P. B.; PARDO, T. A. S. NILC_USP: Aspect Extraction using Semantic Labels. In: **International Workshop on Semantic Evaluation (SemEval)**. [S.l.: s.n.], 2014. p. 433–436. Citado na página 47.
- FREITAS, A.; O’RIAIN, S.; CURRY, E.; SILVA, J. C. P. da; CARVALHO, D. S. Representing Texts as Contextualized Entity-Centric Linked Data Graphs. In: **International Workshop on Database and Expert Systems Applications (DEXA)**. [S.l.: s.n.], 2013. p. 133–137. Citado na página 45.
- FRIEDMAN, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. **Annals of Mathematical Statistics**, v. 11, n. 1, p. 86–92, 1940. Citado na página 78.
- FRIGG, R.; HARTMANN, S. Models in Science. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. [S.l.]: Metaphysics Research Lab, Stanford University, 2018. Citado na página 105.
- FUCHS, C.; BOERSMA, K.; ALBRECHTSLUND, A.; SANDOVAL, M. Introduction: Internet and Surveillance. In: **Internet and Surveillance: The Challenges of Web 2.0 and Social Media**. [S.l.]: Routledge, 2013. v. 16, cap. 1. Citado na página 30.
- FÜRNKRANZ, J. **A Study Using N-Grams Features for Text Categorization**. [S.l.], 1998. Citado na página 52.
- GABRILOVICH, E.; MARKOVITCH, S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In: **International Joint Conference on Artificial Intelligence (IJCAI)**. [S.l.: s.n.], 2007. p. 1606–1611. Citado na página 59.
- GARTEN, J.; HOOVER, J.; JOHNSON, K. M.; BOGHRATI, R.; ISKIWITCH, C.; DEHGHANI, M. Dictionaries and Distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis. **Behavior Research Methods**, v. 50, n. 1, p. 344–361, 2018. Citado na página 53.
- GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. **Transactions on Pattern Analysis and Machine Intelligence (PAMI)**, v. 6, n. 6, p. 721–741, 1984. Citado na página 71.
- GONZALEZ, M. A. I.; LIMA, V. L. S. de; LIMA, J. V. de. Tools for Nominalization: An Alternative for Lexical Normalization. In: **International Workshop on Computational Processing of the Portuguese Language (PROPOR)**. [S.l.: s.n.], 2006. p. 100–109. Citado na página 52.
- GOOSSEN, F.; IJNTEMA, W.; FRASINCAR, F.; HOGENBOOM, F.; KAYMAK, U. News Personalization Using the CF-IDF Semantic Recommender. In: **International Conference on Web Intelligence, Mining and Semantics (WIMS)**. [S.l.: s.n.], 2011. Citado na página 64.
- GREENE, D.; CUNNINGHAM, P. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2006. p. 377–384. Citado na página 68.

- GUTHRIE, D.; ALLISON, B.; LIU, W.; GUTHRIE, L.; WILKS, Y. A Closer Look at Skip-gram Modelling. In: **International Conference on Language Resources and Evaluation (LREC)**. [S.l.: s.n.], 2006. p. 1222–1225. Citado na página 52.
- HAN, B.; COOK, P.; BALDWIN, T. Lexical Normalization for Social Media Text. **Transactions on Intelligent Systems and Technology (TIST)**, v. 4, n. 1, p. 1–27, 2013. Citado na página 53.
- HARB, A.; HAJLAOUI, K. Enhanced Semantic Automatic Ontology Enrichment. In: **Conference on Intelligent Systems Design and Applications (ISDA)**. [S.l.: s.n.], 2010. p. 1113–1118. Citado na página 53.
- HARRIS, Z. S. Distributional Structure. **Word**, v. 10, n. 2–3, p. 146–162, 1954. Citado nas páginas 21, 23 e 39.
- HEARST, M. A. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. In: **Annual Conference of the University of Waterloo Centre for the New OED and Text Research**. [S.l.: s.n.], 1991. Citado na página 55.
- _____. Untangling Text Data Mining. In: **Annual Meeting of the Association for Computational Linguistics (ACL)**. [S.l.: s.n.], 1999. p. 3–10. Citado na página 30.
- HENNIG, L.; UMBRATH, W.; WETZKER, R. An Ontology-Based Approach to Text Summarization. In: **Joint International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (WI-IAT)**. [S.l.: s.n.], 2008. p. 291–294. Citado na página 47.
- HUANG, X.; ALLEVA, F.; HON, H. wuen; HWANG, M. yuh; ROSENFELD, R. The SPHINX-II Speech Recognition System: An Overview. **Computer, Speech & Language**, v. 7, n. 2, p. 137–148, 1993. Citado na página 52.
- IACOBACCI, I.; PILEHVAR, M. T.; NAVIGLI, R. SenseEmbed: Learning Sense Embeddings for Word and Relational Similarity. In: **Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing**. [S.l.: s.n.], 2015. p. 95–105. Citado nas páginas 55 e 57.
- _____. Embeddings for Word Sense Disambiguation: An Evaluation Study. In: **Annual Meeting of the Association for Computational Linguistics (ACL)**. [S.l.: s.n.], 2016. p. 897–907. Citado na página 55.
- Ji, M.; SUN, Y.; DANILEVSKY, M.; HAN, J.; GAO, J. Graph Regularized Transductive Classification on Heterogeneous Information Networks. In: **Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)**. [S.l.: s.n.], 2010. p. 570–586. Citado na página 22.
- JONES, K. S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972. Citado na página 32.
- _____. Index Term Weighting. **Information Storage and Retrieval**, v. 9, n. 11, p. 619–633, 1973. Citado na página 32.
- JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. [S.l.]: Prentice Hall, 2009. Citado nas páginas 24, 25, 46, 47, 52 e 57.

- KALCHBRENNER, N.; GREFFENSTETTE, E.; BLUNSON, P. A Convolutional Neural Network for Modelling Sentences. In: **Annual Meeting of the Association for Computational Linguistics (ACL)**. [S.l.: s.n.], 2014. v. 1, p. 655–665. Citado na página 40.
- KAMBHATLA, N. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. In: **Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions**. [S.l.: s.n.], 2004. Citado na página 47.
- KAO, A.; POTEET, S. R. **Natural Language Processing and Text Mining**. [S.l.]: Springer, 2007. Citado nas páginas 30 e 46.
- KARTSAKLIS, D. **Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras**. Tese (Doutorado) — Wolfson College, University of Oxford, 2014. Citado nas páginas 39 e 40.
- KIM, H. K.; KIM, H.; CHO, S. Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation. **Neurocomputing**, v. 266, p. 336–352, 2017. Citado na página 54.
- KITCHENHAM, B.; CHARTERS, S. **Guidelines for Performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007. Citado na página 42.
- KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: **International Joint Conference on Artificial Intelligence (IJCAI)**. [S.l.: s.n.], 1995. p. 1137–1145. Citado na página 73.
- KUHN, A.; DUCASSE, S.; GİRBA, T. Semantic Clustering: Identifying Topics in Source Code. **Information & Software Technology**, v. 49, n. 3, p. 230–243, 2007. Citado na página 45.
- LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An Introduction to Latent Semantic Analysis. **Discourse Processes**, Taylor & Francis, v. 25, p. 259–284, 1998. Citado na página 33.
- LE, Q. V.; MIKOLOV, T. Distributed Representations of Sentences and Documents. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2014. p. 1188–1196. Citado na página 54.
- LEBRET, R.; COLLOBERT, R. Rehabilitation of Count-Based Models for Word Vector Representations. In: **International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)**. [S.l.: s.n.], 2015. p. 417–429. Part I. Citado na página 45.
- LESER, U.; HAKENBERG, J. What Makes a Gene Name? Named Entity Recognition in the Biomedical Literature. **Briefings in Bioinformatics**, v. 6, n. 4, p. 357–369, 2005. Citado na página 55.
- LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of Massive Datasets**. [S.l.]: Cambridge University, 2014. Citado na página 32.
- LEVY, O.; GOLDBERG, Y. Linguistic Regularities in Sparse and Explicit Word Representations. In: **Conference on Computational Natural Language Learning (CoNLL)**. [S.l.: s.n.], 2014. p. 171–180. Citado na página 59.
- LI, C.; WANG, H.; ZHANG, Z.; SUN, A.; MA, Z. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In: **International Conference on Research and Development in Information Retrieval (SIGIR)**. [S.l.: s.n.], 2016. p. 165–174. Citado na página 53.

- LI, Y.; YANG, T. Guide to Big Data Applications. In: _____. [S.l.]: Springer, 2018. v. 26, cap. Word Embedding for Understanding Natural Language: A Survey, p. 83–104. Citado na página 54.
- LIN, D.; PANTEL, P. DIRT - Discovery of Inference Rules from Text. In: **International Conference on Knowledge Discovery and Data Mining (SIGKDD)**. [S.l.: s.n.], 2001. p. 323–328. Citado na página 39.
- LOVINS, J. B. Development of a Stemming Algorithm. **Mechanical Translation and Computational Linguistics**, v. 11, n. 1–2, p. 22–31, 1968. Citado na página 52.
- LUHN, H. P. The Automatic Creation of Literature Abstracts. **IBM Journal of Research and Development**, v. 2, n. 2, p. 159–165, 1958. Citado na página 32.
- MANCINI, M.; CAMACHO-COLLADOS, J.; IACOBACCI, I.; NAVIGLI, R. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In: **Conference on Computational Natural Language Learning (CoNLL)**. [S.l.: s.n.], 2017. p. 100–111. Citado nas páginas 55 e 57.
- MANNING, C. D.; RAGHAVAN, P.; SCHÜTZ, H. **Introduction to Information Retrieval**. [S.l.]: Cambridge University, 2008. Citado nas páginas 25 e 36.
- MARNEFFE, M.-C. de; MACCARTNEY, B.; MANNING, C. D. Generating Typed Dependency Parses from Phrase Structure Parses. In: **International Conference on Language Resources and Evaluation (LREC)**. [S.l.: s.n.], 2006. p. 449–454. Citado na página 55.
- 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: **International Conference on Engineering Design (ICED)**. [S.l.: s.n.], 2015. Citado na página 53.
- _____. Comparative Analysis of Methods and Tools Applicability for Product and IPSS Development Based on Text Mining Techniques. **Product: Management & Development**, v. 13, n. 2, p. 57–66, 2015. Citado na página 53.
- MCCALLUM, A. K. MALLET: A Machine Learning for Language Toolkit. Unpublished. 2002. Citado na página 71.
- MESNIL, G.; HE, X.; DENG, L.; BENGIO, Y. Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding. In: **Annual Conference of the International Speech Communication Association (INTERSPEECH)**. [S.l.: s.n.], 2013. p. 3771–3775. Citado na página 25.
- MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. **Computing Research Repository (CoRR)**, 2013. Cornell University Library. Citado nas páginas 25, 48, 54, 55 e 63.
- MIKOLOV, T.; LE, Q. V.; SUTSKEVER, I. Exploiting Similarities among Languages for Machine Translation. **Computing Research Repository (CoRR)**, 2013. Cornell University Library. Citado nas páginas 25 e 55.

- MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. In: **Annual Conference on Neural Information Processing Systems (NIPS)**. [S.l.: s.n.], 2013. p. 3111–3119. Citado nas páginas 25, 54, 55, 61 e 63.
- MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. J. Introduction to WordNet: An On-line Lexical Database. **International Journal of Lexicography**, v. 3, n. 4, p. 235–244, 1990. Citado na página 45.
- MITCHELL, J.; LAPATA, M. Composition in Distributional Models of Semantics. **Cognitive Science**, v. 34, n. 8, p. 1388–1429, 2010. Citado na página 39.
- MIZZARO, S.; PAVAN, M.; SCAGNETTO, I.; VALENTI, M. Short Text Categorization Exploiting Contextual Enrichment and External Knowledge. In: **International Workshop on Social Media Retrieval and Analysis (SoMeRA)**. [S.l.: s.n.], 2014. p. 57–62. Citado nas páginas 45 e 54.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicações**. [S.l.]: Manole, 2003. cap. 4, p. 89–114. Citado nas páginas 33 e 34.
- MORO, A.; CECCONI, F.; NAVIGLI, R. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In: **International Semantic Web Conference (ISWC)**. [S.l.: s.n.], 2014. p. 25–28. Citado na página 62.
- MORO, A.; RAGANATO, A.; NAVIGLI, R. Entity Linking Meets Word Sense Disambiguation: A Unified Approach. **Transactions of the Association for Computational Linguistics (TACL)**, v. 2, p. 231–244, 2014. Citado na página 57.
- MUKAKA, M. M. Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research. **Malawi Medical Journal**, v. 24, n. 3, p. 69–71, 2012. Citado na página 83.
- NAVIGLI, R.; PONZETTO, S. P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. **Artificial Intelligence**, v. 193, p. 217–250, 2012. Citado nas páginas 45 e 54.
- NEMENYI, P. **Distribution-Free Multiple Comparisons**. Tese (Doutorado) — Princeton University, 1963. Citado na página 78.
- NEWMAN, M. E. J. **Networks: An Introduction**. [S.l.]: Oxford University, 2010. Citado na página 22.
- NOURASHRAFEDDIN, S.; MILIOS, E. E.; ARNOLD, D. V. An Ensemble Approach for Text Document Clustering Using Wikipedia Concepts. In: **Symposium on Document Engineering (DocEng)**. [S.l.: s.n.], 2014. p. 107–116. Citado na página 22.
- NÓBREGA, F. A. A.; PARDO, T. A. S. Desambiguação Lexical de Sentido com uso de informação multidocumento por meio de redes de co-ocorrência. In: **Brazilian Symposium in Information and Human Language Technology (STIL)**. [S.l.: s.n.], 2013. p. 138–147. Citado na página 55.
- OLIVEIRA, R. P. de. **Semântica Formal: uma breve introdução**. [S.l.]: Mercado de Letras, 2001. Citado na página 23.

PEARSON, K. On Lines and Planes of Closest Fit to Systems of Points in Space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, 1901. Citado na página 82.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global Vectors for Word Representation. In: **Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 55.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic Mapping Studies in Software Engineering. In: **International Conference on Evaluation and Assessment in Software Engineering (EASE)**. [S.l.: s.n.], 2008. p. 68–77. Citado na página 42.

PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; MANANDHAR, S.; ANDROUTSOPOULOS, I. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: **International Workshop on Semantic Evaluation (SemEval)**. [S.l.: s.n.], 2015. p. 486–495. Citado na página 68.

POSSENTI, S. **Por que (não) ensinar gramática na escola**. [S.l.]: Mercado de Letras, 1996. Citado na página 37.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. [S.l.]: Morgan Kaufmann, 1993. Citado na página 35.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: **Language Resources and Evaluation Conference (LREC)**. [S.l.: s.n.], 2010. (Workshop on New Challenges for NLP Frameworks), p. 45–50. Citado na página 62.

RIEMER, N. **Introducing Semantics**. [S.l.]: Cambridge University, 2010. Citado na página 47.

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, 2015. Citado na página 35.

ROSSI, R. G.; FALEIROS, T. de P.; LOPES, A. de A.; REZENDE, S. O. Inductive Model Generation for Text Categorization Using a Bipartite Heterogeneous Network. In: **International Conference on Data Mining (ICDM)**. [S.l.: s.n.], 2012. p. 1086–1091. Citado na página 73.

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 (JCST)**, v. 29, n. 3, p. 361–375, 2014. Citado nas páginas 22 e 35.

ROSSI, R. G.; REZENDE, S. O. Building a Topic Hierarchy Using the Bag-of-Related-Words Representation. In: **Symposium on Document Engineering (DocEng)**. [S.l.: s.n.], 2011. p. 195–204. Citado na página 53.

SALTON, G. **Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**. [S.l.]: Addison-Wesley, 1989. Citado na página 105.

SALTON, G.; BUCKLEY, C.; ALLAN, J. Automatic Structuring of Text Files. **Electronic Publishing**, v. 5, n. 1, p. 1–17, 1992. Citado na página 105.

SALTON, G.; WONG, A.; YANG, C.-S. A Vector Space Model for Automatic Indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, 1975. Citado nas páginas 22 e 39.

- SAUSSURE, F. de. **Course in General Linguistics**. [S.l.]: McGraw-Hill, 1966. Translated by Wade Baskin. Citado nas páginas 29 e 37.
- SCHEICHER, R. B.; SINOARA, R. A.; KOGA, N. J.; REZENDE, S. O. Uso de expressões do domínio na classificação automática de documentos. In: **Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)**. [S.l.: s.n.], 2016. p. 625–636. Citado nas páginas 55 e 73.
- SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: **International Conference on New Methods in Language Processing (NeMLaP)**. [S.l.: s.n.], 1994. Citado na página 52.
- SCHMIDHUBER, J. Deep Learning in Neural Networks: An Overview. **Neural Networks**, v. 61, p. 85–117, 2015. Citado na página 45.
- SCHULZ, A.; RISTOSKI, P.; PAULHEIM, H. I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs. In: **Extended Semantic Web Conference (ESWC)**. [S.l.: s.n.], 2013. p. 22–33. Citado na página 22.
- SCHÜTZE, H.; PEDERSEN, J. A Vector Model for Syntagmatic and Paradigmatic Relatedness. In: **Annual Conference of the UW Centre for the New OED and Text Research**. [S.l.: s.n.], 1993. p. 104–113. Citado na página 38.
- SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys (CSUR)**, v. 34, n. 1, p. 1–47, 2002. Citado na página 22.
- SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3–4, p. 591–611, 1965. Citado na página 82.
- SHEHATA, S. A WordNet-Based Semantic Model for Enhancing Text Clustering. In: **International Conference on Data Mining (ICDM)**. [S.l.: s.n.], 2009. p. 477–482. Citado na página 22.
- SINOARA, R. A. **Aspectos semânticos na representação de textos para classificação automática**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018. Citado nas páginas 73 e 78.
- SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text Mining and Semantics: A Systematic Mapping Study. **Journal of the Brazilian Computer Society (JBCS)**, v. 23, n. 1, p. 1–20, 2017. Citado nas páginas 23, 24, 25, 40, 41, 42, 43, 44, 45 e 52.
- SINOARA, R. A.; REZENDE, S. O. **Best sports: A Portuguese Collection of Documents for Semantics-Concerned Text Mining Research**. [S.l.], 2018. Citado na página 68.
- SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. Semantic Role-Based Representations in Text Classification. In: **International Conference on Pattern Recognition (ICPR)**. [S.l.: s.n.], 2016. p. 2313–2318. Citado nas páginas 55 e 73.
- SINOARA, R. A.; SCHEICHER, R. B.; REZENDE, S. O. Evaluation of Latent Dirichlet Allocation for Document Organization in Different Levels of Semantic Complexity. In: **Symposium on Computational Intelligence and Data Mining (CIDM)**. [S.l.: s.n.], 2017. p. 1–8. Citado na página 41.

- SINOARA, R. A.; SUNDERMANN, C. V.; MARCACINI, R. M.; DOMINGUES, M. A.; REZENDE, S. O. Named Entities as Privileged Information for Hierarchical Text Clustering. In: **International Database Engineering & Applications Symposium (IDEAS)**. [S.l.: s.n.], 2014. p. 57–66. Citado na página 55.
- SOCHER, R.; HUVAL, B.; MANNING, C. D.; NG, A. Y. Semantic Compositionality through Recursive Matrix-Vector Spaces. In: **Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**. [S.l.: s.n.], 2012. p. 1201–1211. Citado na página 40.
- SOHRAB, M. G.; MIWA, M.; SASAKI, Y. Centroid-Means-Embedding: An Approach to Infusing Word Embeddings into Features for Text Classification. In: **Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)**. [S.l.: s.n.], 2015. p. 289–300. Part I. Citado nas páginas 45, 54 e 70.
- SOKOLOVA, M.; LAPALME, G. A Systematic Analysis of Performance Measures for Classification Tasks. **Information Processing and Management**, v. 45, n. 4, p. 427–437, 2009. Citado na página 36.
- SPEARMAN, C. The Proof and Measurement of Association Between Two Things. **The American Journal of Psychology**, v. 15, n. 1, p. 72–101, 1904. Citado na página 82.
- SRIVASTAVA, A. N.; SAHAMI, M. **Text Mining: Classification, Clustering, and Applications**. [S.l.]: CRC, 2009. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). Citado nas páginas 21, 22 e 30.
- SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. YAGO: A Large Ontology from Wikipedia and WordNet. **Journal of Web Semantics**, v. 6, n. 3, p. 203–217, 2008. Citado na página 54.
- SUNDERMANN, C. V.; ANTUNES, J.; DOMINGUES, M. A.; REZENDE, S. O. Exploration of Word Embedding Model to Improve Context-Aware Recommender Systems. In: **International Conference on Web Intelligence (WI)**. [S.l.: s.n.], 2018. Citado nas páginas 61, 65 e 91.
- TAMILIN, A.; MAGNINI, B.; SERAFINI, L.; GIRARDI, C.; JOSEPH, M.; ZANOLI, R. Context-Driven Semantic Enrichment of Italian News Archive. In: **Extended Semantic Web Conference (ESWC)**. [S.l.: s.n.], 2010. p. 364–378. Citado nas páginas 54 e 55.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2005. Citado na página 22.
- TSUJII, J.; ANANIADOU, S. Thesaurus or Logical Ontology, Which One Do We Need for Text Mining? **Language Resources and Evaluation**, v. 39, n. 1, p. 77–90, 2005. Citado na página 53.
- TUKEY, J. W. **Exploratory Data Analysis**. [S.l.]: Addison-Wesley, 1977. Citado na página 74.
- TURNEY, P. D. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In: **International Conference on Computational Linguistics (COLING)**. [S.l.: s.n.], 2008. p. 905–912. Citado na página 39.
- TURNEY, P. D.; PANTEL, P. From Frequency to Meaning: Vector Space Models of Semantics. **Journal of Artificial Intelligence Research (JAIR)**, v. 37, p. 141–188, 2010. Citado nas páginas 39 e 47.

- VAPNIK, V. N.; CHERVONENKIS, A. Y. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. **Theory of Probability and its Applications**, v. 16, n. 2, p. 264–280, 1971. Citado na página 35.
- VASQUES, D. G.; ZAMBON, A. C.; BAIOCO, G. B.; MARTINS, P. S. An Approach to Knowledge Acquisition Based on Verbal Semantics. In: **Hawaii International Conference on System Sciences (HICSS)**. [S.l.: s.n.], 2016. p. 4144–4153. Citado na página 55.
- WANG, P.; XU, B.; XU, J.; TIAN, G.; LIU, C.-L.; HAO, H. Semantic Expansion Using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification. **Neurocomputing**, v. 174, p. 806–814, 2016. Citado na página 45.
- WANG, X.; SUN, J.-T.; CHEN, Z.; ZHAI, C. Latent Semantic Analysis for Multiple-Type Inter-related Data Objects. In: **Annual International Conference on Research and Development in Information Retrieval (SIGIR)**. [S.l.: s.n.], 2006. p. 236–243. Citado na página 53.
- WEISS, S. M.; INDURKHYA, N.; ZHANG, T.; DAMERAU, F. **Text Mining: Predictive Methods for Analyzing Unstructured Information**. [S.l.]: Springer, 2010. Citado nas páginas 22, 31 e 32.
- WIMALASURIYA, D. C.; DOU, D. Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches. **Journal of Information Science**, v. 36, n. 3, p. 306–323, 2010. Citado na página 53.
- YAMAKAWA, H.; JING; FELDMAN, A. Semantic Enrichment of Text Representation with Wikipedia for Text Classification. In: **International Conference on Systems, Man and Cybernetics (SMC)**. [S.l.: s.n.], 2010. p. 4333–4340. Citado na página 53.
- ZIPF, G. K. Human Behavior and the Principle of Least Effort. **Addison-Wesley**, 1949. Citado na página 32.

GLOSSÁRIO

Atributo: Terminologia utilizada nesta dissertação para se referir às variáveis/características extraídas de um conjunto de documentos. A relação entre atributo e termo é hierárquica, considerando que um atributo pode ser formado por um conjunto não nulo de termos. Por exemplo, no caso da técnica de representação de textos CIRT.5 os atributos são os contextos (conjunto de termos) extraídos dos documentos de texto (SALTON, 1989).

Modelo: Terminologia utilizada nesta dissertação para se referir às representações matemáticas de simplificações da realidade, que permitem a exploração de relacionamentos entre conceitos mentais e experimentais. Na tarefa de classificação de documentos os modelos podem se referir às configurações de variáveis utilizadas para categorizar um novo exemplo, enquanto na representação de textos os modelos podem se referir aos espaços geométricos explorados por diferentes abordagens (FRIGG; HARTMANN, 2018).

Método: Terminologia utilizada nesta dissertação para se referir aos procedimentos específicos que devem ser efetuados para realizar algum tipo de tarefa. A relação entre técnica, método e abordagem é hierárquica, o que permite definir que as técnicas (como fazer) – implementadas por meio de algoritmos (forma como é feito) – executam um método (o que fazer) que é consistente com uma abordagem (o porquê fazer) (ANTHONY, 1963).

Termo: Terminologia utilizada nesta dissertação para se referir às unidades lexicais de um texto, formado por um conjunto de caracteres, utilizadas para designar conceitos em um discurso especializado. Um termo pode ser simples (p. ex., conta e guarda) ou composto (p. ex., conta corrente e guarda-chuva) em relação aos *tokens* que o formam, e designar/representar conceitos como substantivos, adjetivos, sintagmas, siglas, acrônimos, abreviaturas, fórmulas e símbolos (CASTELLVÍ, 2010).

Texto: Terminologia utilizada nesta dissertação para se referir à qualquer tipo de dado textual não estruturado (p. ex., notícias, *e-mails*, *reviews* e páginas da *Web*), em comparação com os dados textuais armazenados em bancos de dados. Assim, textos e documentos de texto são considerados sinônimos que se referem à qualquer conjunto ou coleção de dados não estruturados formados por termos (SALTON; BUCKLEY; ALLAN, 1992).

FONTES DE TEXTOS UTILIZADAS

Neste apêndice são apresentadas as propriedades dos recursos textuais utilizados para avaliar as abordagens de representação de textos. Além das coleções de textos preparadas para a tarefa de classificação automática de textos, formadas por documentos rotulados, também são apresentadas as principais propriedades dos textos extraídos da Wikipédia e utilizados como fonte externa de conhecimento em duas das abordagens de representação de textos avaliadas.

Por motivo de padronização, tanto as coleções de textos baseadas nas bases de dados selecionadas quanto os textos extraídos da Wikipédia foram preparados considerando os mesmo processos de enriquecimento textual. Em ambos os casos todos os textos foram inicialmente delimitados pelas suas sentenças, consideradas os escopos de interesse, e posteriormente tokenizados para separar termos de não termos (p. ex., pontuações). Esses procedimentos de delimitação de sentenças e de tokenização caracterizam a versão “raw” dos textos, que passam a representar as sentenças por linhas e os termos separados por espaços em branco.

Baseado na versão “raw”, foram aplicados diferentes procedimentos de enriquecimento textual nas versões “word” e “id” dos textos. Na versão “word” os termos dos textos passaram a caracterizar conceitos ou entidades nomeadas por meio de termos simples ou compostos separados pelo caractere “_”. Já na versão “id” os termos dos textos passaram a caracterizar conceitos ou entidades nomeadas desambiguadas por meio de *synset ids*. Essas duas conversões textuais permitem avaliar as abordagens de representação de textos sob diferentes cenários de aplicação, propiciando uma análise mais refinada de suas performances.

Nas seções a seguir são respectivamente apresentadas as principais propriedades das coleções de textos e dos textos extraídos da Wikipédia. Na Seção A.1, além das propriedades quantitativas dos textos são apresentadas as distribuições de cada coleção em razão do número de classes para cada nível de complexidade semântica abordado. Na Seção A.2 são apresentadas as propriedades quantitativas dos textos e dos modelos pré-treinados da Wikipédia.

A.1 Coleções de textos

Conforme abordado na Subseção 4.2.1, para o desenvolvimento desta dissertação foram selecionadas 3 bases de dados textuais, a BEST *sports* v1.1 do idioma português e a SE-ABSA15 e BBC do idioma inglês. Embora cada uma dessas bases de dados tenha originado 3 versões de coleções de textos diferentes, elas diferem sobretudo nos níveis de complexidade semântica que permitem explorar por meio da tarefa de classificação automática de textos.

Em relação às coleções de textos originadas da base de dados BEST *sports* v1.1, os documentos de texto foram selecionados em razão das 4 classes majoritárias de tópicos esportivos “ESP_FUTEBOL” (FUT), “ESP_TENIS” (TEN), “ESP_FORMULA_1” (F1) e “ESP_MOTOVELOCIDADE” (MOT). Essa seleção foi realizada com o objetivo de reduzir o desbalanceamento da base de dados original formada por 66 classes relacionadas à tópicos esportivos.

Considerando as classes “BRvenceu” (ven), “BRnaoVenceu” (nven), “BRnaoCitado” (ncit) e “Ruido” (ruid) relacionadas à informações semânticas a respeito do desempenho de atletas brasileiros (BR) nas competições esportivas alvo das notícias, a base de dados **BS-4** possui 15 classes relacionadas à informações de tópico e semântica. Portanto, as versões da coleção BS-4 são formadas por 283 documentos de texto distribuídos entre as 15 classes de tópico e semântica assim como ilustrado na Figura 20.

Figura 20 – Distribuição de classes da base de dados BS-4.

Fonte: Dados da pesquisa.

Em razão dos processos de enriquecimento textual aplicado nas versões “word” e “id” das coleções de textos, o número de sentenças e de termos difere em cada caso. A versão “raw” da coleção é formada por 3.081 sentenças e 61.990 termos, enquanto as versões “word” e “id”

são formadas por 2.797 sentenças e 24.642 termos. Como é possível observar por meio desses valores, o processo de enriquecimento textual pode modificar as coleções de textos implicando em diferentes números de sentenças e de termos para cada versão considerada. Considerando as características das três versões de textos, é possível afirmar que a base de dados BS-4 possui em média cerca de 10 sentenças e 131 termos por documento.

Em relação às coleções de textos originadas da base de dados SE-ABSA15, os documentos de texto foram selecionados em razão dos 3 níveis de notas considerados mais relevantes nas avaliações de produtos e serviços, já que além das 3 classes relacionadas às categorias “Laptops” (Lap), “Hotels” (Hot) e “Restaurants” (Res), os documentos também são rotulados por número de estrelas de 0 à 5. Considerando que o número de estrelas geralmente representa o nível de satisfação dos usuários em relação aos itens avaliados, foram considerados somente os documentos rotulados com pelo menos 1 estrela, já que a ausência de nota (0 estrelas) pode estar associada à possíveis ruídos das avaliações.

Por motivo de conveniência, os documentos rotulados por até 2 estrelas passaram a ser caracterizados pela classe “negative” (neg), enquanto os documentos rotulados por 3 estrelas passaram a ser caracterizados pela classe “neutral” (neut) e os demais documentos passaram a ser caracterizados pela classe “positive” (pos). Portanto, as versões da coleção **SE** são formadas por 815 documentos de texto distribuídos em 9 classes de categoria e polaridade assim como ilustrado na Figura 21. A versão “raw” da coleção é formada por 4.896 sentenças e 70.660 termos, enquanto as versões “word” e “id” são formadas por 3.918 sentenças e 23.264 termos. Considerando as características das três versões de textos, é possível afirmar que a base de dados SE possui em média cerca de 5 sentenças e 48 termos por documento.

Figura 21 – Distribuição de classes da base de dados SE.

Fonte: Dados da pesquisa.

Em relação às coleções de textos originadas da base de dados BBC, foram selecionados todos os 2.225 documentos de texto da base de dados original. Diferente das outras bases de dados utilizadas, os documentos de texto da BBC permitem explorar somente o 1º nível de complexidade semântica, já que são rotulados por meio dos 5 tópicos temáticos “business” (bus), “entertainment” (ent), “politics” (pol), “sport” (spo) e “tech” (tec).

Em relação às bases de dados BS-4 e SE, os documentos de texto da base de dados BBC possuem uma distribuição mais homogênea em razão das classes, assim como ilustrado na Figura 22. A versão “raw” da coleção é formada por 44.055 sentenças e 975.157 termos, enquanto as versões “word” e “id” são respectivamente formadas por 41.715 sentenças e 364.897 termos e por 41.696 sentenças e 364.878 termos. Considerando as características das três versões de textos, é possível afirmar que a base de dados BBC possui em média cerca de 19 sentenças e 255 termos por documento.

Figura 22 – Distribuição de classes da base de dados BBC.

Fonte: Dados da pesquisa.

A.2 Fontes externas de conhecimento

Conforme abordado na Subseção 4.2.2, os textos da Wikipédia foram utilizados como fonte externa de conhecimento tanto na abordagem CIRT.5 E-L da técnica de representação de textos proposta CIRT.5, quanto na abordagem BoV_Wiki inspirada em modelos baseados em *word embeddings*. Em ambos os casos os *backups* textuais da Wikipédia foram preparados e utilizados isoladamente para cada idioma tratado e para cada versão de textos avaliada, o que significa que todos os retreinamentos dos modelos utilizados foram realizados de maneira não consistente ou não incremental.

No idioma português, a versão “raw” dos textos extraídos e preparados é formada por 11.188.036 sentenças e 267.556.478 termos, enquanto as versões “word” e “id” são respectivamente formadas por 10.936.987 sentenças e 97.042.660 termos e por 10.916.928 sentenças e 97.022.601 termos. Já no idioma inglês, a versão “raw” dos textos extraídos e preparados é formada por 105.854.452 sentenças e 2.393.204.905 de termos, enquanto as versões “word” e “id” são respectivamente formadas por 104.788.075 sentenças e 838.854.133 termos e por 104.503.643 sentenças e 838.569.701 termos.

Em razão do *backup* da Wikipédia para o idioma inglês ser consideravelmente maior do que para o idioma português (cerca de 12 vezes maior), os modelos pré-treinados a partir desses textos também têm diferenças expressivas. No caso do idioma português, os modelos das versões “raw”, “word” e “id” são respectivamente formados por 331.369, 293.862 e 203.635 termos únicos, enquanto no idioma inglês essas versões resultam em modelos respectivamente formados por 1.391.032, 2.014.141 e 1.407.953 termos únicos. É importante ressaltar que tanto para o idioma português quanto para o inglês os modelos foram pré-treinados considerando corte de frequência mínima de 10 ocorrências para cada termo, ou seja, esses valores de termos únicos referem-se aos termos únicos que ocorreram pelo menos 10 vezes nos textos.

MELHORES RESULTADOS OBTIDOS

Neste apêndice são apresentados os resultados obtidos na tarefa de classificação automática de textos realizada, relativos às performances de cada abordagem de representação de textos avaliada (vide a Subseção 4.2.2). Em razão dos extensos resultados obtidos, são enfatizados os melhores valores de Acurácia de cada abordagem em cada nível de complexidade semântica explorado (vide a Subseção 2.3.1). As tabelas apresentadas também levam em conta as diferentes versões das coleções de textos, e os resultados são selecionados de acordo com cada configuração dos algoritmos de classificação utilizados.

Conforme apresentado na Subseção 4.2.3, a partir dos 5 algoritmos de classificação utilizados são gerados 66 arquivos de resultado relativos à cada configuração considerada. Por motivo de conveniência esses valores máximos de Acurácia são organizados em tuplas associadas às configurações de cada algoritmo. No caso dos algoritmos NB e J48, o conteúdo das tuplas não difere pois foi considerada uma única configuração para cada um; na tupla do algoritmo J48, o valor da tupla refere-se ao nível de confiança considerado. Em relação ao algoritmo KNN os valores das tuplas referem-se respectivamente às métrica de distância, às ponderações por distância e aos valores de K considerados. Já no caso dos algoritmos IMBHN^C e IMBHN^R , os valores das tuplas referem-se respectivamente às taxas de correção de erros e ao critério de parada do erro dos mínimos quadrados considerados.

As tabelas apresentadas a seguir referem-se aos valores máximos de Acurácia utilizados para mensurar a performance, por meio de diagramas de distância crítica, das abordagens de representação de textos avaliadas (vide a Seção 4.3). Em razão da extensão dessas distribuições de Acurácia, os resultados de cada cenário são apresentados por meio de duas tabelas de valores, onde a tabela de continuação é identificada pela abreviatura “cont.”.

Tabela 3 – Acurácia máxima na versão “raw” das coleções: 1º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	98,5837	92,9310	81,9704	93,9778	97,1552	75,3739
J48; 0,25	96,0837	94,6798	74,1133	91,1823	88,2218	68,4538
IMBHN ^C ; 0,01; 0,01	99,2980	98,2143	84,3966	96,4901	97,2945	81,3491
IMBHN ^C ; 0,05; 0,01	99,2980	98,2143	84,3966	96,4901	97,2945	81,3491
IMBHN ^C ; 0,1; 0,01	99,2980	98,2143	84,3966	96,4901	97,2945	81,3491
IMBHN ^C ; 0,5; 0,01	99,2980	98,2143	84,3966	96,4901	97,2945	81,3491
IMBHN ^R ; 0,01; 0,01	100,0000	99,6429	89,3719	98,5961	95,2153	80,6316
IMBHN ^R ; 0,05; 0,01	100,0000	99,6429	89,3719	98,5961	95,2153	80,6316
IMBHN ^R ; 0,1; 0,01	100,0000	99,6429	89,3719	98,5961	95,2153	80,6316
IMBHN ^R ; 0,5; 0,01	100,0000	99,6429	89,3719	98,5961	95,2153	80,6316
KNN; euclidiana; não; 1	98,2266	97,8695	71,4286	87,6355	98,2143	81,3940
KNN; euclidiana; não; 3	98,2389	97,1552	74,5690	87,3399	98,2143	79,8216
KNN; euclidiana; não; 5	98,2389	96,0961	73,8300	85,2217	97,8695	80,6306
KNN; euclidiana; não; 7	96,4532	95,7512	73,5222	83,8300	97,8571	80,8550
KNN; euclidiana; não; 9	96,7980	95,3941	72,4631	84,4704	97,5000	80,9001
KNN; euclidiana; não; 11	96,7980	95,7512	72,1182	82,7340	97,8695	80,7211
KNN; euclidiana; não; 13	96,1084	95,0369	70,7020	82,0567	97,1552	80,0917
KNN; euclidiana; não; 15	96,4532	93,6207	69,9877	81,6872	97,1675	79,8663
KNN; euclidiana; não; 17	97,1675	94,3350	69,3103	81,6995	96,4532	80,2711
KNN; euclidiana; não; 19	97,1675	94,3473	69,2980	82,0320	96,4655	79,7317
KNN; euclidiana; não; 25	94,6675	91,9555	68,9778	80,9729	95,7512	79,4168
KNN; euclidiana; não; 35	93,2635	91,1457	66,8596	80,9729	95,7635	78,7434
KNN; euclidiana; não; 45	90,7882	90,2923	65,4830	79,5813	94,4722	78,2047
KNN; euclidiana; não; 55	90,4310	89,9772	65,9312	78,5099	94,2924	77,1260
KNN; euclidiana; sim; 1	98,2266	97,8695	71,4286	87,6355	98,2143	81,3940
KNN; euclidiana; sim; 3	97,8818	97,5000	73,4975	87,3399	98,2143	82,1587
KNN; euclidiana; sim; 5	98,5961	97,1552	73,5099	85,2217	97,8695	83,0564
KNN; euclidiana; sim; 7	97,1552	96,4655	72,8202	83,8300	97,8571	82,6979
KNN; euclidiana; sim; 9	96,7980	96,4655	72,4631	84,4704	98,2143	83,2814
KNN; euclidiana; sim; 11	97,1552	95,7512	72,0443	82,7340	97,8695	82,8330
KNN; euclidiana; sim; 13	96,7980	96,4532	71,0591	82,0567	97,8695	82,4282
KNN; euclidiana; sim; 15	96,4532	94,3350	70,7020	81,6872	97,5246	82,1575
KNN; euclidiana; sim; 17	96,8103	95,0493	70,3571	81,6995	96,8103	82,5173

Fonte: Dados da pesquisa.

Tabela 4 – Acurácia máxima na versão “raw” das coleções: 1º nível (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	97,1675	95,0493	69,6798	82,0320	96,1084	81,9331
KNN; euclidiana; sim; 25	95,3818	93,0776	69,3350	80,9729	95,7512	81,7083
KNN; euclidiana; sim; 35	94,3227	92,7183	67,9079	81,3300	95,7635	80,8104
KNN; euclidiana; sim; 45	92,2044	92,0892	67,5494	79,5813	95,2365	80,4961
KNN; euclidiana; sim; 55	91,8473	91,9994	67,7275	78,5099	95,0117	79,5528
KNN; cosseno; não; 1	98,5837	99,6429	83,0296	92,9433	98,9286	80,3626
KNN; cosseno; não; 3	99,6552	98,9409	84,7906	93,6453	97,1755	79,7330
KNN; cosseno; não; 5	99,6552	97,5123	86,4901	94,0025	97,1755	79,7307
KNN; cosseno; não; 7	99,2980	96,0837	83,7562	92,2291	97,5414	79,8665
KNN; cosseno; não; 9	99,2980	96,0837	84,4458	93,2882	97,5444	79,5512
KNN; cosseno; não; 11	98,9409	95,5812	84,0517	93,6330	97,2990	79,3272
KNN; cosseno; não; 13	99,2980	95,5812	84,4212	94,3473	97,1771	79,1025
KNN; cosseno; não; 15	99,2857	95,7046	81,9335	92,9310	97,1771	78,9230
KNN; cosseno; não; 17	99,6429	96,0837	82,2906	92,4936	97,0551	79,1930
KNN; cosseno; não; 19	99,6429	96,7980	80,8867	92,2414	96,6847	78,8779
KNN; cosseno; não; 25	99,6429	96,7980	80,5296	92,3131	96,4393	78,1610
KNN; cosseno; não; 35	99,6429	95,9515	78,7685	92,1329	95,5827	77,4409
KNN; cosseno; não; 45	98,9286	96,2014	80,5296	91,4582	94,8479	76,3615
KNN; cosseno; não; 55	99,2857	94,8494	80,5172	91,0098	94,3571	75,2838
KNN; cosseno; sim; 1	98,5837	99,6429	83,0296	92,9433	98,9286	80,3626
KNN; cosseno; sim; 3	99,2980	99,2857	84,7906	93,6453	97,2990	80,4973
KNN; cosseno; sim; 5	99,6552	98,5837	86,4901	94,3473	97,1755	80,4052
KNN; cosseno; sim; 7	99,6552	97,5123	84,0271	93,2568	97,4179	80,3606
KNN; cosseno; sim; 9	99,6552	97,5123	85,1355	93,9901	97,4210	80,1808
KNN; cosseno; sim; 11	99,2980	97,1552	84,4704	94,0148	97,2990	79,5980
KNN; cosseno; sim; 13	99,2980	96,7980	84,4089	95,0493	97,1771	79,5966
KNN; cosseno; sim; 15	99,2980	96,4409	82,3522	94,3596	97,1771	79,2827
KNN; cosseno; sim; 17	99,3857	96,4409	82,3399	94,3719	97,0551	79,3284
KNN; cosseno; sim; 19	99,6429	96,7980	82,6478	94,0025	96,6847	79,1478
KNN; cosseno; sim; 25	99,6429	97,1552	82,6232	93,9769	96,4393	78,2051
KNN; cosseno; sim; 35	99,6429	96,7980	80,1970	93,7525	95,7061	77,4407
KNN; cosseno; sim; 45	99,6429	96,5673	81,2685	93,8420	94,8479	76,4964
KNN; cosseno; sim; 55	99,6429	96,3219	80,8498	93,6169	94,4791	75,6436

Fonte: Dados da pesquisa.

Tabela 5 – Acurácia máxima na versão “word” das coleções: 1º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	97,8941	25,3695	64,8544	91,5271	97,4225	85,9799
J48; 0,25	93,6576	58,1632	69,0997	92,5739	90,4216	78,4297
IMBHN ^C ; 0,01; 0,01	99,6429	48,5908	77,7793	95,3941	97,6649	88,0909
IMBHN ^C ; 0,05; 0,01	99,6429	48,5908	77,7793	95,3941	97,6649	88,0909
IMBHN ^C ; 0,1; 0,01	99,6429	48,5908	77,7793	95,3941	97,6649	88,0909
IMBHN ^C ; 0,5; 0,01	99,6429	48,5908	77,7793	95,3941	97,6649	88,0909
IMBHN ^R ; 0,01; 0,01	97,6647	45,5134	80,1220	96,0961	100,0000	88,3606
IMBHN ^R ; 0,05; 0,01	100,0000	45,5134	80,1220	96,0961	95,8296	88,3606
IMBHN ^R ; 0,1; 0,01	100,0000	45,5134	80,1220	96,0961	95,8296	88,3606
IMBHN ^R ; 0,5; 0,01	100,0000	45,5134	80,1220	96,0961	95,8296	88,3606
KNN; euclidiana; não; 1	94,6429	46,8564	64,8519	89,7537	96,4363	87,9102
KNN; euclidiana; não; 3	95,0123	48,4673	66,3911	89,3596	97,2975	86,8769
KNN; euclidiana; não; 5	96,0837	51,8910	67,8531	88,6576	97,0476	87,2817
KNN; euclidiana; não; 7	96,4409	51,2827	67,8711	88,3128	97,1755	86,9670
KNN; euclidiana; não; 9	96,7980	51,8925	67,1229	87,5985	97,0521	87,3718
KNN; euclidiana; não; 11	96,0837	52,9991	67,6182	86,5640	97,0521	87,1921
KNN; euclidiana; não; 13	95,7389	53,2355	66,7615	86,8966	97,1786	86,8337
KNN; euclidiana; não; 15	95,7266	52,8726	66,2752	86,5887	97,1771	87,0121
KNN; euclidiana; não; 17	95,0369	53,2385	66,8669	85,8867	97,1771	86,6521
KNN; euclidiana; não; 19	95,3818	52,8741	67,4947	86,2315	96,8097	86,6974
KNN; euclidiana; não; 25	93,9655	53,7293	67,7356	84,8399	96,5628	87,0571
KNN; euclidiana; não; 35	90,1108	53,7293	65,7814	83,8054	95,2168	86,6972
KNN; euclidiana; não; 45	86,5640	53,7293	65,7739	79,9138	94,6071	86,3366
KNN; euclidiana; não; 55	80,5911	53,7293	66,5086	80,2463	94,4726	86,2918
KNN; euclidiana; sim; 1	94,6429	46,8564	64,8519	89,7537	96,4363	87,9102
KNN; euclidiana; sim; 3	96,7857	49,5694	67,9871	89,3596	97,1740	88,2245
KNN; euclidiana; sim; 5	96,0837	51,7675	68,9672	88,6576	97,0476	88,4943
KNN; euclidiana; sim; 7	96,4409	51,1623	67,9840	88,3128	97,1755	88,3141
KNN; euclidiana; sim; 9	96,7980	51,6471	68,2294	88,3005	97,0521	88,5400
KNN; euclidiana; sim; 11	96,4409	53,0006	68,5998	86,5640	97,0521	88,5398
KNN; euclidiana; sim; 13	96,7980	52,8696	67,4842	86,8966	97,1786	88,4957
KNN; euclidiana; sim; 15	96,4409	52,5053	66,8774	86,5887	97,1771	88,4501
KNN; euclidiana; sim; 17	96,7980	53,1165	67,8561	85,8867	97,1771	88,4947

Fonte: Dados da pesquisa.

Tabela 6 – Acurácia máxima na versão “word” das coleções: 1º nível (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	96,0961	52,8741	67,9795	86,2315	96,9316	88,3152
KNN; euclidiana; sim; 25	95,7389	53,7293	67,4917	84,8399	96,5628	88,6297
KNN; euclidiana; sim; 35	91,5271	53,7293	66,0193	83,8054	95,7046	88,0903
KNN; euclidiana; sim; 45	89,3842	53,7293	66,0193	80,6281	94,9684	87,8201
KNN; euclidiana; sim; 55	83,0542	53,7293	66,2647	80,6034	94,9669	87,9548
KNN; cosseno; não; 1	100,0000	42,6950	69,8437	92,9310	96,6832	88,0011
KNN; cosseno; não; 3	100,0000	46,2587	73,5200	93,9655	98,1572	87,0565
KNN; cosseno; não; 5	100,0000	51,7766	75,9636	92,9187	98,0337	86,7875
KNN; cosseno; não; 7	98,9409	51,1563	77,4330	92,2167	98,0337	86,3380
KNN; cosseno; não; 9	99,2980	50,7889	77,3050	92,8941	98,0337	86,8784
KNN; cosseno; não; 11	98,5837	49,1990	75,5977	91,8350	97,7898	86,4297
KNN; cosseno; não; 13	98,9409	48,8347	75,7167	92,1798	97,9118	86,3853
KNN; cosseno; não; 15	98,5837	48,4628	76,1999	93,6084	97,6664	86,1148
KNN; cosseno; não; 17	98,5837	50,7859	75,4622	92,5493	97,6649	85,9352
KNN; cosseno; não; 19	98,9286	48,8211	76,3279	93,5961	97,1740	86,1597
KNN; cosseno; não; 25	99,2857	50,4170	77,6754	94,6798	97,1755	85,9334
KNN; cosseno; não; 35	98,9286	52,6272	75,7046	93,9778	96,6877	85,7088
KNN; cosseno; não; 45	98,5714	52,6272	77,4390	94,3350	95,3372	85,4389
KNN; cosseno; não; 55	98,5714	53,3559	76,6998	94,6675	94,7215	85,1238
KNN; cosseno; sim; 1	100,0000	42,6950	69,8437	92,9310	96,6832	87,9560
KNN; cosseno; sim; 3	100,0000	48,0970	71,9030	93,9655	98,1572	87,5060
KNN; cosseno; sim; 5	100,0000	50,7949	73,5095	93,9901	98,0337	87,1010
KNN; cosseno; sim; 7	100,0000	50,1716	75,5947	93,2759	98,0337	86,8780
KNN; cosseno; sim; 9	100,0000	49,0741	75,4667	93,2512	97,9103	86,9232
KNN; cosseno; sim; 11	100,0000	47,8425	75,5992	92,5393	97,9118	86,3845
KNN; cosseno; sim; 13	100,0000	46,8609	75,7182	92,8818	97,9118	86,2950
KNN; cosseno; sim; 15	100,0000	48,0834	76,3264	93,2512	97,6664	85,9797
KNN; cosseno; sim; 17	99,6429	49,1915	75,5902	94,3227	97,6649	86,0700
KNN; cosseno; sim; 19	99,6429	48,0894	75,5917	94,6675	97,1740	86,1599
KNN; cosseno; sim; 25	99,6429	48,4523	75,4652	94,3103	97,1755	86,0237
KNN; cosseno; sim; 35	99,2857	49,8133	75,4577	94,3227	96,6877	85,7536
KNN; cosseno; sim; 45	99,2857	50,7980	75,5887	93,6207	95,3372	85,3941
KNN; cosseno; sim; 55	99,2857	52,3924	75,2153	95,0246	94,7215	85,1238

Fonte: Dados da pesquisa.

Tabela 7 – Acurácia máxima na versão “id” das coleções: 1º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	97,5616	28,5961	69,2734	93,6330	98,1602	
J48; 0,25	94,7044	59,7651	66,9979	92,5246	89,8163	
IMBHN ^C ; 0,01; 0,01	99,2980	47,0988	78,3345	97,5123	97,7913	
IMBHN ^C ; 0,05; 0,01	99,2980	47,0988	78,3345	97,5123	97,7913	
IMBHN ^C ; 0,1; 0,01	99,2980	47,0988	78,3345	97,5123	97,7913	
IMBHN ^C ; 0,5; 0,01	99,2980	47,0988	78,3345	97,5123	97,7913	
IMBHN ^R ; 0,01; 0,01	100,0000	49,0726	79,1626	98,2266	95,4607	
IMBHN ^R ; 0,05; 0,01	100,0000	49,0726	79,1626	98,2266	95,4607	
IMBHN ^R ; 0,1; 0,01	100,0000	49,0726	79,1626	98,2266	95,4607	
IMBHN ^R ; 0,5; 0,01	100,0000	49,0726	79,1626	98,2266	95,4607	
KNN; euclidiana; não; 1	94,2980	46,7359	66,2007	89,7783	97,7928	
KNN; euclidiana; não; 3	96,7980	48,7142	64,7608	89,7783	97,9148	
KNN; euclidiana; não; 5	96,8103	51,4047	65,6595	90,1232	98,0352	
KNN; euclidiana; não; 7	96,7980	52,1454	66,5138	90,1355	98,0367	
KNN; euclidiana; não; 9	96,0837	53,3725	67,3690	89,1010	98,5276	
KNN; euclidiana; não; 11	96,4532	51,0343	67,9536	88,7562	98,1557	
KNN; euclidiana; não; 13	96,8227	52,3818	67,6389	88,0419	97,9103	
KNN; euclidiana; não; 15	95,7512	53,6119	66,9194	88,7438	98,1572	
KNN; euclidiana; não; 17	94,3350	53,3710	66,6960	87,3276	98,2806	
KNN; euclidiana; não; 19	94,6921	53,8543	66,6921	86,2562	98,0352	
KNN; euclidiana; não; 25	93,2389	53,4839	65,7959	86,2562	97,6679	
KNN; euclidiana; não; 35	89,7044	53,6058	65,3960	85,2217	96,1954	
KNN; euclidiana; não; 45	86,1823	53,7293	65,1551	84,4951	95,0903	
KNN; euclidiana; não; 55	83,0172	53,7293	64,5825	82,7340	94,9669	
KNN; euclidiana; sim; 1	94,2980	46,7359	66,2007	89,7783	97,7928	
KNN; euclidiana; sim; 3	96,0837	48,9581	67,1908	89,7783	97,9148	
KNN; euclidiana; sim; 5	97,5123	51,1593	67,6371	90,1232	97,9118	
KNN; euclidiana; sim; 7	97,5123	52,0205	68,8959	90,1355	98,0367	
KNN; euclidiana; sim; 9	97,1552	53,2460	69,5255	89,1010	98,4041	
KNN; euclidiana; sim; 11	97,8695	51,0328	70,3363	88,7562	98,0322	
KNN; euclidiana; sim; 13	97,5246	52,1364	70,1101	88,0419	97,9103	
KNN; euclidiana; sim; 15	96,1084	52,8756	69,2567	88,7438	97,9103	
KNN; euclidiana; sim; 17	96,4655	53,1241	69,1649	87,3276	98,1572	

Fonte: Dados da pesquisa.

Tabela 8 – Acurácia máxima na versão “id” das coleções: 1º nível (cont).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	96,4532	53,6073	68,9848	86,2562	98,0352	
KNN; euclidiana; sim; 25	95,0246	53,7293	68,1786	86,2562	97,7898	
KNN; euclidiana; sim; 35	90,7635	53,4839	67,2775	85,2217	96,3174	
KNN; euclidiana; sim; 45	87,9433	53,7293	67,3668	84,8522	95,5827	
KNN; euclidiana; sim; 55	85,1232	53,7293	67,0983	83,4483	95,0574	
KNN; cosseno; não; 1	99,6552	46,1397	73,0322	93,3128	97,7913	
KNN; cosseno; não; 3	98,9409	47,6137	72,1329	94,6798	98,2791	
KNN; cosseno; não; 5	99,2980	49,2081	73,3036	96,1084	98,2791	
KNN; cosseno; não; 7	98,5837	49,9563	73,4483	96,1207	98,0367	
KNN; cosseno; não; 9	98,6465	50,8010	73,7975	95,7512	97,9118	
KNN; cosseno; não; 11	98,5837	50,5480	73,8866	96,1084	97,6679	
KNN; cosseno; não; 13	98,5837	51,5191	73,5222	95,7759	97,6664	
KNN; cosseno; não; 15	98,9409	52,6227	74,2241	96,1330	97,7898	
KNN; cosseno; não; 17	99,2857	52,9931	73,5222	96,4778	97,5429	
KNN; cosseno; não; 19	99,6429	54,9593	73,8547	96,1207	97,6649	
KNN; cosseno; não; 25	98,9286	53,7383	72,4048	95,4310	97,1755	
KNN; cosseno; não; 35	99,2857	53,2445	72,1791	95,0493	96,3159	
KNN; cosseno; não; 45	98,9286	52,9871	72,2688	95,0739	95,2108	
KNN; cosseno; não; 55	98,9286	52,3728	72,2686	94,0025	94,8419	
KNN; cosseno; sim; 1	99,6552	46,1397	73,0322	93,3128	97,7913	
KNN; cosseno; sim; 3	99,2980	48,9657	74,6041	95,3941	98,1557	
KNN; cosseno; sim; 5	99,6552	48,9627	74,5599	96,0961	98,2791	
KNN; cosseno; sim; 7	99,2980	48,9717	75,5665	96,4532	97,9133	
KNN; cosseno; sim; 9	99,2980	50,4366	75,9729	96,1084	97,9118	
KNN; cosseno; sim; 11	99,2980	49,5679	76,6626	96,4655	97,6679	
KNN; cosseno; sim; 13	99,6552	49,9307	76,3177	96,1330	97,6664	
KNN; cosseno; sim; 15	99,2980	50,2996	77,3768	96,4901	97,7898	
KNN; cosseno; sim; 17	99,6429	50,4261	75,9729	96,4778	97,5429	
KNN; cosseno; sim; 19	99,6429	51,4107	75,9606	96,1207	97,6649	
KNN; cosseno; sim; 25	99,6429	50,6700	74,5690	95,7759	97,1755	
KNN; cosseno; sim; 35	99,6429	51,4077	73,1897	95,0739	96,3159	
KNN; cosseno; sim; 45	99,2857	52,1484	73,2560	95,4310	95,2108	
KNN; cosseno; sim; 55	99,2857	51,1653	73,5246	95,4187	94,8419	

Fonte: Dados da pesquisa.

Tabela 9 – Acurácia máxima na versão “raw” das coleções: 2º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	67,6016	45,7693	63,1813	63,5516	66,2436	33,9837
J48; 0,25	72,2824	67,1168	66,7465	66,8714	65,6519	61,7254
IMBHN ^C ; 0,01; 0,01	79,5242	68,8497	59,2578	72,6438	76,4363	50,5450
IMBHN ^C ; 0,05; 0,01	79,5242	68,8497	59,2578	72,6438	76,4363	50,5450
IMBHN ^C ; 0,1; 0,01	79,5242	68,8497	59,2578	72,6438	76,4363	50,5450
IMBHN ^C ; 0,5; 0,01	79,5242	68,8497	59,2578	72,6438	76,4363	50,5450
IMBHN ^R ; 0,01; 0,01	83,6902	76,6847	62,8184	74,3421	81,7043	59,6251
IMBHN ^R ; 0,05; 0,01	83,6902	76,6847	62,8184	74,3421	81,7043	59,6251
IMBHN ^R ; 0,1; 0,01	83,6902	76,6847	62,8184	74,3421	81,7043	59,6251
IMBHN ^R ; 0,5; 0,01	83,6902	76,6847	62,8184	74,3421	81,7043	59,6251
KNN; euclidiana; não; 1	56,5517	69,0861	65,2740	65,6429	67,2373	60,8687
KNN; euclidiana; não; 3	63,8151	71,0403	66,3821	66,7465	71,4152	64,4294
KNN; euclidiana; não; 5	63,6932	71,9045	67,0024	66,7465	73,3830	64,7877
KNN; euclidiana; não; 7	63,6932	74,7200	68,2174	66,7465	74,6131	63,4447
KNN; euclidiana; não; 9	63,5712	74,5995	68,2219	66,7465	74,1147	64,0650
KNN; euclidiana; não; 11	63,8181	74,3601	68,3424	66,7465	75,3433	63,8181
KNN; euclidiana; não; 13	63,9416	73,6164	68,4658	66,7465	76,8217	63,5727
KNN; euclidiana; não; 15	63,9416	73,4899	68,2174	66,7465	76,6908	63,8076
KNN; euclidiana; não; 17	63,8181	73,0006	68,2189	66,7465	76,3234	62,9509
KNN; euclidiana; não; 19	63,8181	71,5327	68,2204	66,7465	75,5887	63,4417
KNN; euclidiana; não; 25	63,8181	70,0602	67,9735	66,7465	75,7046	63,8061
KNN; euclidiana; não; 35	63,3333	68,3424	68,3424	66,7480	75,4592	64,7877
KNN; euclidiana; não; 45	63,3333	67,1153	67,9780	66,7465	76,3129	65,2815
KNN; euclidiana; não; 55	63,3348	66,8699	67,9750	67,2343	74,8389	66,1367
KNN; euclidiana; sim; 1	56,5517	69,0861	65,2740	65,6429	67,2373	60,8687
KNN; euclidiana; sim; 3	64,6748	71,5282	66,6230	66,6245	71,5402	63,8151
KNN; euclidiana; sim; 5	64,7983	73,2535	67,1229	66,6245	73,8768	65,1581
KNN; euclidiana; sim; 7	64,7983	75,0888	68,2174	66,7465	75,1069	64,1825
KNN; euclidiana; sim; 9	64,6763	74,8464	68,4658	66,7465	74,6101	64,5528
KNN; euclidiana; sim; 11	64,9232	73,6239	68,4643	66,7480	75,4682	64,5483
KNN; euclidiana; sim; 13	65,0467	73,1256	68,4658	66,7465	76,7013	64,5544
KNN; euclidiana; sim; 15	65,0467	73,1241	68,4643	66,7465	76,6923	64,7952
KNN; euclidiana; sim; 17	64,9232	72,8787	68,4658	66,7465	76,2014	64,1855

Fonte: Dados da pesquisa.

Tabela 10 – Acurácia máxima na versão “raw” das coleções: 2º nível (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	64,9232	71,6546	68,4658	66,8684	75,5887	64,0605
KNN; euclidiana; sim; 25	64,9232	69,9368	68,0939	66,7465	75,9530	64,5483
KNN; euclidiana; sim; 35	64,0681	68,2189	68,2189	66,7465	75,7061	64,6718
KNN; euclidiana; sim; 45	63,9461	67,1153	67,9735	66,8699	76,4363	65,2846
KNN; euclidiana; sim; 55	63,9476	67,1153	67,9735	67,2343	75,2078	65,8958
KNN; cosseno; não; 1	68,9702	71,7706	59,7636	65,3869	69,5694	48,0924
KNN; cosseno; não; 3	76,0750	76,3144	64,1719	69,1930	73,3770	58,4207
KNN; cosseno; não; 5	77,1740	77,3005	65,8913	70,5465	76,5613	59,1554
KNN; cosseno; não; 7	77,9118	77,4164	65,7663	72,8847	76,8037	60,0256
KNN; cosseno; não; 9	77,5399	77,1725	67,1033	71,6591	76,6862	59,3978
KNN; cosseno; não; 11	77,6664	77,4210	66,9828	71,5266	75,4652	58,9145
KNN; cosseno; não; 13	78,4026	77,4195	66,5086	71,2797	75,8341	58,2957
KNN; cosseno; não; 15	77,4149	77,1725	67,1108	71,1683	75,9515	58,4131
KNN; cosseno; não; 17	77,9073	77,6649	67,2343	70,8025	75,9515	59,0304
KNN; cosseno; não; 19	77,9088	77,7883	67,1168	70,4291	76,5658	63,5637
KNN; cosseno; não; 25	78,0337	77,2990	67,2403	69,5724	76,5658	64,1749
KNN; cosseno; não; 35	77,4210	76,4408	66,9934	68,7067	75,7061	64,9067
KNN; cosseno; não; 45	76,5568	75,9530	66,9919	68,2174	74,8494	65,8883
KNN; cosseno; não; 55	75,8220	75,9470	66,9919	68,0909	74,1147	65,7648
KNN; cosseno; sim; 1	68,9702	71,7706	59,8856	65,3869	69,5694	48,2144
KNN; cosseno; sim; 3	75,3403	75,7001	65,0376	69,3240	73,3755	58,2927
KNN; cosseno; sim; 5	76,6847	77,0491	67,2298	71,0373	76,5613	60,0151
KNN; cosseno; sim; 7	77,9103	77,5384	66,7465	72,7642	76,9271	60,8808
KNN; cosseno; sim; 9	78,0307	77,5384	67,6122	72,3924	76,6847	60,0120
KNN; cosseno; sim; 11	78,0367	77,1755	67,4872	72,1379	75,4622	59,2833
KNN; cosseno; sim; 13	78,0352	76,6832	67,2418	71,7690	76,1999	58,5411
KNN; cosseno; sim; 15	78,1542	76,3144	67,6001	71,7721	75,8281	59,1494
KNN; cosseno; sim; 17	78,0292	77,1740	68,2174	71,2903	76,3189	59,3978
KNN; cosseno; sim; 19	78,1542	76,9301	67,8516	70,6700	76,9316	59,0304
KNN; cosseno; sim; 25	77,9088	77,1755	67,4827	70,3026	76,4438	59,2728
KNN; cosseno; sim; 35	77,5414	76,5613	67,2358	69,6899	75,3388	60,0030
KNN; cosseno; sim; 45	77,4149	76,3159	67,2388	69,4460	74,6025	60,4908
KNN; cosseno; sim; 55	77,1740	76,4378	67,2388	69,2005	74,1132	60,9831

Fonte: Dados da pesquisa.

Tabela 11 – Acurácia máxima na versão “word” das coleções: 2º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	64,5212	28,9409	66,1412	66,1412	69,7983	35,7016
J48; 0,25	72,6408	66,7495	66,9904	67,5956	70,4351	61,8428
IMBHN ^C ; 0,01; 0,01	76,1834	52,1530	62,0897	69,6688	78,7609	48,8332
IMBHN ^C ; 0,05; 0,01	76,1834	52,1530	62,0897	69,6688	78,7609	48,8332
IMBHN ^C ; 0,1; 0,01	76,1834	52,1530	62,0897	69,6688	78,7609	48,8332
IMBHN ^C ; 0,5; 0,01	76,1834	52,1530	62,0897	69,6688	78,7609	48,8332
IMBHN ^R ; 0,01; 0,01	81,3475	57,9178	65,8898	75,0723	82,9298	51,5387
IMBHN ^R ; 0,05; 0,01	81,3475	57,9178	65,8898	75,0723	82,9298	51,5387
IMBHN ^R ; 0,1; 0,01	81,3475	57,9178	65,8898	75,0723	82,9298	51,5387
IMBHN ^R ; 0,5; 0,01	81,3475	57,943	65,8898	75,0723	82,9298	51,5387
KNN; euclidiana; não; 1	57,6108	55,3448	67,4737	68,0985	69,3180	59,5152
KNN; euclidiana; não; 3	50,5542	56,0795	66,7435	67,1123	75,0933	62,0837
KNN; euclidiana; não; 5	48,4113	55,2168	67,1183	66,7465	76,0795	64,6612
KNN; euclidiana; não; 7	67,2388	54,3631	67,1168	66,7465	76,5718	65,5239
KNN; euclidiana; não; 9	67,6092	56,6923	68,8392	66,7465	76,8157	65,4005
KNN; euclidiana; não; 11	65,0301	55,8205	68,4658	66,7465	76,5658	65,8898
KNN; euclidiana; não; 13	64,0455	58,6480	68,5878	66,7465	76,8097	66,3806
KNN; euclidiana; não; 15	64,6612	60,9816	68,5893	66,7465	76,9332	66,5071
KNN; euclidiana; não; 17	63,8001	61,5974	68,4658	66,8714	76,8067	66,6290
KNN; euclidiana; não; 19	63,4282	61,1096	68,3424	66,7495	78,0292	66,0193
KNN; euclidiana; não; 25	63,1828	66,1307	68,7112	67,1153	77,6603	66,3821
KNN; euclidiana; não; 35	66,7480	66,1322	68,4658	67,2373	78,8904	66,9889
KNN; euclidiana; não; 45	66,7465	66,6230	68,2204	67,4721	79,0093	65,7633
KNN; euclidiana; não; 55	66,7465	66,7465	68,3424	68,2099	79,5032	66,6245
KNN; euclidiana; sim; 1	57,6108	55,3448	67,4737	68,0985	69,3180	59,5152
KNN; euclidiana; sim; 3	52,3276	56,2030	66,4995	66,6230	74,8494	62,3276
KNN; euclidiana; sim; 5	50,8867	56,9362	67,2403	66,8699	76,2045	65,2755
KNN; euclidiana; sim; 7	66,9919	58,1677	67,3622	66,7465	76,9392	65,6459
KNN; euclidiana; sim; 9	67,2388	58,8994	68,7157	66,7465	76,8157	65,3990
KNN; euclidiana; sim; 11	65,3960	59,3857	68,5878	66,7465	76,4469	66,1352
KNN; euclidiana; sim; 13	65,5164	60,9846	68,4658	66,7465	76,4423	66,2587
KNN; euclidiana; sim; 15	66,2556	62,0837	68,3424	66,7465	76,6877	66,2602
KNN; euclidiana; sim; 17	66,1352	62,5775	68,3424	66,7465	76,5628	66,6260

Fonte: Dados da pesquisa.

Tabela 12 – Acurácia máxima na versão “word” das coleções: 2º nível (cont).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	66,0102	62,2147	68,4658	66,7465	77,9103	66,1382
KNN; euclidiana; sim; 25	64,2879	65,8883	68,7112	67,3547	77,6634	66,8744
KNN; euclidiana; sim; 35	66,1352	66,1322	68,3439	67,2373	78,7669	66,8684
KNN; euclidiana; sim; 45	66,7465	66,6230	68,2204	67,8485	79,1313	66,8699
KNN; euclidiana; sim; 55	66,7465	66,7465	68,3424	68,8241	79,7471	66,6245
KNN; cosseno; não; 1	68,0985	56,7013	62,2087	64,5408	73,1286	49,3210
KNN; cosseno; não; 3	75,0843	59,3948	66,5056	71,2843	76,6877	50,3011
KNN; cosseno; não; 5	78,2791	60,0166	66,3746	71,8985	79,1463	49,6763
KNN; cosseno; não; 7	77,6573	60,3779	66,3806	72,9991	78,6435	50,9184
KNN; cosseno; não; 9	78,6375	61,9723	67,2553	73,2490	78,7730	51,5312
KNN; cosseno; não; 11	77,7793	62,8320	67,2463	72,7492	78,4071	51,4002
KNN; cosseno; não; 13	78,2791	63,8076	67,8666	72,6227	78,8949	51,7736
KNN; cosseno; não; 15	78,2761	62,5821	68,3499	72,5053	78,2776	51,7796
KNN; cosseno; não; 17	78,4026	63,4387	68,3484	72,8741	78,4026	52,2659
KNN; cosseno; não; 19	77,6694	64,1734	68,4568	73,3634	77,9133	65,5239
KNN; cosseno; não; 25	78,1632	64,0485	68,2234	72,8756	78,0322	66,7450
KNN; cosseno; não; 35	76,3174	65,8898	67,2403	71,8925	78,0352	67,1138
KNN; cosseno; não; 45	75,9485	67,1153	67,1153	71,6471	77,0521	67,3592
KNN; cosseno; não; 55	75,0918	66,7465	66,9919	71,5221	76,9241	66,5026
KNN; cosseno; sim; 1	68,0985	56,7013	62,3306	64,5408	73,1286	48,1015
KNN; cosseno; sim; 3	74,2246	59,1478	66,6245	70,4261	76,9316	49,3255
KNN; cosseno; sim; 5	78,4026	60,8762	67,3547	72,5143	79,2683	49,3165
KNN; cosseno; sim; 7	78,1497	60,9877	67,7281	73,1271	78,6435	50,3132
KNN; cosseno; sim; 9	78,2716	62,2192	67,6001	74,1072	79,0184	50,6790
KNN; cosseno; sim; 11	77,7808	63,0744	68,2264	73,3634	78,4056	50,4246
KNN; cosseno; sim; 13	78,3996	64,0530	68,5727	73,7308	78,5260	51,0419
KNN; cosseno; sim; 15	78,4011	63,9310	68,5893	73,3634	78,1542	51,1698
KNN; cosseno; sim; 17	78,1572	64,4249	68,4613	73,1226	77,7883	51,1683
KNN; cosseno; sim; 19	78,0352	64,2999	68,7157	73,2415	77,5429	50,6760
KNN; cosseno; sim; 25	78,5276	64,1765	68,7067	73,3665	77,9103	51,8985
KNN; cosseno; sim; 35	76,9286	64,6673	68,2159	72,9961	78,2836	52,3908
KNN; cosseno; sim; 45	76,9301	64,9112	67,8485	72,5053	76,9301	52,5128
KNN; cosseno; sim; 55	76,8067	65,1566	67,7281	72,1379	76,9256	51,5342

Fonte: Dados da pesquisa.

Tabela 13 – Acurácia máxima na versão “id” das coleções: 2º nível.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	62,6950	23,3128	65,2740	68,7142	68,0864	
J48; 0,25	69,2036	67,1168	66,8714	68,3439	72,3969	
IMBHN ^C ; 0,01; 0,01	78,2716	51,7811	61,0810	72,8907	74,7260	
IMBHN ^C ; 0,05; 0,01	78,2716	51,7811	61,0810	72,8907	74,7260	
IMBHN ^C ; 0,1; 0,01	78,2716	51,7811	61,0810	72,8907	74,7260	
IMBHN ^C ; 0,5; 0,01	78,2716	51,7811	61,0810	72,8907	74,7260	
IMBHN ^R ; 0,01; 0,01	83,1828	58,0458	63,0744	76,0659	81,9512	
IMBHN ^R ; 0,05; 0,01	83,1828	58,0458	63,0744	76,0659	81,9512	
IMBHN ^R ; 0,1; 0,01	83,1828	58,0458	63,0744	76,0659	81,9512	
IMBHN ^R ; 0,5; 0,01	83,1828	58,0458	63,1963	76,0659	81,9512	
KNN; euclidiana; não; 1	53,6453	54,1147	67,3517	67,7236	70,4261	
KNN; euclidiana; não; 3	46,6626	58,5321	66,9889	66,9904	73,1376	
KNN; euclidiana; não; 5	50,8867	62,4586	66,8699	66,7465	73,0126	
KNN; euclidiana; não; 7	48,4360	62,8169	66,8699	66,7465	72,8907	
KNN; euclidiana; não; 9	47,6970	63,9295	66,7465	66,7465	73,1376	
KNN; euclidiana; não; 11	45,9729	64,2894	67,6031	66,7465	72,8907	
KNN; euclidiana; não; 13	44,5544	64,5363	67,9750	67,1123	72,8907	
KNN; euclidiana; não; 15	53,3845	66,0102	67,6137	66,7465	72,8907	
KNN; euclidiana; não; 17	61,2331	66,3806	68,1015	66,8684	72,8922	
KNN; euclidiana; não; 19	63,6856	66,3776	67,8516	66,8654	72,5263	
KNN; euclidiana; não; 25	67,2433	66,3791	68,3439	67,2358	72,1605	
KNN; euclidiana; não; 35	67,1168	66,7465	68,2219	66,9904	72,5279	
KNN; euclidiana; não; 45	66,8684	66,7465	68,4658	66,8684	72,2824	
KNN; euclidiana; não; 55	66,7465	66,7465	68,2219	66,8684	71,6682	
KNN; euclidiana; sim; 1	53,6453	54,1147	67,3517	67,7236	70,4261	
KNN; euclidiana; sim; 3	53,0296	58,8979	66,9889	66,9889	73,0142	
KNN; euclidiana; sim; 5	54,7783	61,4740	66,7465	66,7374	73,0126	
KNN; euclidiana; sim; 7	52,6847	62,5715	66,7465	66,6245	72,8907	
KNN; euclidiana; sim; 9	49,1133	63,9295	66,7465	66,7465	73,0142	
KNN; euclidiana; sim; 11	47,7463	64,1689	67,2328	66,7465	72,7672	
KNN; euclidiana; sim; 13	45,2463	64,6582	67,7281	66,9904	72,7672	
KNN; euclidiana; sim; 15	53,8768	65,8898	67,4917	66,9919	72,7672	
KNN; euclidiana; sim; 17	61,3565	66,2572	68,1015	66,8684	72,7687	

Fonte: Dados da pesquisa.

Tabela 14 – Acurácia máxima na versão “id” das coleções: 2º nível (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	63,8091	66,2556	67,7281	66,8669	72,4029	
KNN; euclidiana; sim; 25	67,4887	66,5011	68,3454	67,3592	72,0370	
KNN; euclidiana; sim; 35	67,1168	66,7465	68,3424	67,1138	72,4044	
KNN; euclidiana; sim; 45	66,8684	66,7465	68,2234	66,7465	72,1590	
KNN; euclidiana; sim; 55	66,7465	66,7465	68,3439	66,7465	71,5447	
KNN; cosseno; não; 1	70,9139	52,1439	58,4222	64,6522	68,7112	
KNN; cosseno; não; 3	74,5905	58,1662	61,9633	67,2448	73,1316	
KNN; cosseno; não; 5	74,9563	58,5291	63,5336	69,8254	73,9898	
KNN; cosseno; não; 7	76,4363	61,7148	65,7648	71,1728	75,3403	
KNN; cosseno; não; 9	77,1710	63,4417	65,8928	71,9030	76,2030	
KNN; cosseno; não; 11	77,0521	63,3062	66,7480	71,7811	76,4559	
KNN; cosseno; não; 13	77,4164	62,9449	67,3622	72,8892	76,4529	
KNN; cosseno; não; 15	76,8037	63,1873	66,5026	73,9898	77,3126	
KNN; cosseno; não; 17	76,3144	63,5547	66,8774	73,8693	76,7013	
KNN; cosseno; não; 19	76,1894	63,8016	67,6107	73,4989	78,1768	
KNN; cosseno; não; 25	76,3144	64,2939	67,4797	74,1072	77,5610	
KNN; cosseno; não; 35	76,6802	65,0331	67,3577	72,5098	77,0641	
KNN; cosseno; não; 45	75,9425	66,0132	66,8699	72,2659	76,5778	
KNN; cosseno; não; 55	75,3267	66,2556	67,3622	71,9015	76,0795	
KNN; cosseno; sim; 1	70,9139	52,1439	58,4222	64,6522	68,7112	
KNN; cosseno; sim; 3	74,7185	58,6585	62,2117	67,1259	73,1301	
KNN; cosseno; sim; 5	75,6986	59,2653	64,0394	70,3162	73,9898	
KNN; cosseno; sim; 7	76,9301	62,2057	66,2541	71,0539	75,5872	
KNN; cosseno; sim; 9	77,2975	62,5805	65,8898	71,6546	75,7151	
KNN; cosseno; sim; 11	76,9301	63,8001	66,6245	72,5158	76,2135	
KNN; cosseno; sim; 13	77,4179	63,8031	66,9889	73,1301	76,4559	
KNN; cosseno; sim; 15	77,4195	63,0668	67,2328	73,6164	77,3141	
KNN; cosseno; sim; 17	77,2975	63,1903	67,2433	73,7413	76,8248	
KNN; cosseno; sim; 19	76,9256	63,6811	67,1108	73,6209	78,0533	
KNN; cosseno; sim; 25	77,1740	63,8016	67,4872	74,5980	77,1921	
KNN; cosseno; sim; 35	76,6802	64,1689	67,3577	72,7582	76,6938	
KNN; cosseno; sim; 45	76,4303	64,7817	67,3577	72,7567	76,9452	
KNN; cosseno; sim; 55	76,5522	65,0256	67,3607	72,3924	76,0810	

Fonte: Dados da pesquisa.

Tabela 15 – Acurácia máxima na versão “raw” das coleções: 1º e 2º níveis.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	68,0894	42,7463	37,4261	55,4487	65,7814	21,5884
J48; 0,25	66,1503	59,8780	35,5706	56,2014	55,5842	41,2180
IMBHN ^C ; 0,01; 0,01	75,2243	68,7112	42,0690	67,2207	74,7230	38,8934
IMBHN ^C ; 0,05; 0,01	75,2243	68,7112	42,0690	67,2207	74,7230	38,8934
IMBHN ^C ; 0,1; 0,01	75,2243	68,7112	42,0690	67,2207	74,7230	38,8934
IMBHN ^C ; 0,5; 0,01	75,2243	68,7112	42,0690	67,2207	74,7230	38,8934
IMBHN ^R ; 0,01; 0,01	79,0108	75,2123	46,6155	67,2373	66,7525	43,9326
IMBHN ^R ; 0,05; 0,01	79,0108	75,2123	46,6155	67,2373	66,7525	43,9326
IMBHN ^R ; 0,1; 0,01	79,0108	75,2123	46,6155	67,2373	66,7525	43,9326
IMBHN ^R ; 0,5; 0,01	79,0108	75,2123	46,6155	67,2373	66,7525	43,9326
KNN; euclidiana; não; 1	55,8374	59,0108	34,4821	43,8300	63,9175	43,5697
KNN; euclidiana; não; 3	51,2685	59,7425	33,8528	41,2436	68,3409	44,3059
KNN; euclidiana; não; 5	52,6724	63,3092	35,2168	43,9401	71,0464	46,2632
KNN; euclidiana; não; 7	54,4212	66,7465	37,0566	45,4095	71,9030	44,6643
KNN; euclidiana; não; 9	51,2685	67,2388	37,6558	44,3044	71,8970	44,5498
KNN; euclidiana; não; 11	47,7586	67,6152	38,7775	45,0346	73,2535	45,2846
KNN; euclidiana; não; 13	49,1256	67,1244	37,6769	43,1963	74,1162	45,0346
KNN; euclidiana; não; 15	45,9360	66,8729	38,2957	43,6932	73,9883	45,0286
KNN; euclidiana; não; 17	45,9236	65,6489	38,6646	42,5911	73,0096	44,6567
KNN; euclidiana; não; 19	44,1626	64,6673	40,4983	42,3427	73,0066	44,4128
KNN; euclidiana; não; 25	41,7118	62,0882	38,6691	41,9648	73,1256	44,1614
KNN; euclidiana; não; 35	41,0222	57,9268	40,3598	41,9708	72,1424	45,7618
KNN; euclidiana; não; 45	38,8793	56,4589	39,8841	40,9892	71,6516	45,5164
KNN; euclidiana; não; 55	37,4631	55,3508	40,3689	40,3764	70,2951	44,7802
KNN; euclidiana; sim; 1	55,8374	59,0108	34,4821	43,8300	63,9175	43,5697
KNN; euclidiana; sim; 3	57,2906	62,1996	35,3312	42,8350	68,4658	45,0452
KNN; euclidiana; sim; 5	57,2660	64,5363	35,2138	45,2906	71,2978	46,6366
KNN; euclidiana; sim; 7	56,2069	67,3668	38,0382	47,9886	72,7672	45,6489
KNN; euclidiana; sim; 9	54,7906	67,8531	38,6450	47,9855	72,5128	45,5330
KNN; euclidiana; sim; 11	50,2094	68,2294	38,4146	46,7570	73,7459	46,0178
KNN; euclidiana; sim; 13	49,8522	67,8621	38,0292	45,5285	74,2382	46,0163
KNN; euclidiana; sim; 15	49,1379	67,7356	39,4068	45,8988	74,3556	46,0132
KNN; euclidiana; sim; 17	47,7217	66,5101	39,5333	44,4354	72,8862	45,3960

Fonte: Dados da pesquisa.

Tabela 16 – Acurácia máxima na versão “raw” das coleções: 1º e 2º níveis (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	46,2931	65,6459	40,0286	43,3243	73,4974	44,5378
KNN; euclidiana; sim; 25	43,1281	62,8275	39,2818	42,3366	73,6164	44,5348
KNN; euclidiana; sim; 35	41,7241	58,7880	40,2620	43,3273	72,6332	45,7678
KNN; euclidiana; sim; 45	39,2365	57,6889	40,4938	41,7284	72,0205	45,5194
KNN; euclidiana; sim; 55	38,1773	56,2105	40,2469	40,8657	70,2981	45,0286
KNN; cosseno; não; 1	65,5285	67,9720	37,1831	57,9027	66,8684	34,9714
KNN; cosseno; não; 3	72,5173	71,1638	36,3054	58,8859	70,6700	37,3020
KNN; cosseno; não; 5	75,8296	73,4974	38,7699	61,8383	74,2291	42,9554
KNN; cosseno; não; 7	76,9347	73,7413	38,6495	64,2969	75,5796	43,1993
KNN; cosseno; não; 9	77,2930	73,3710	38,8874	63,1903	74,4806	43,4372
KNN; cosseno; não; 11	77,1740	73,6164	39,7395	63,6856	73,9913	41,9663
KNN; cosseno; não; 13	77,9118	73,4929	40,6022	63,1978	74,4851	42,3291
KNN; cosseno; não; 15	76,9271	73,4944	40,7212	62,8245	73,9883	42,4466
KNN; cosseno; não; 17	77,6649	74,7185	40,3202	62,4616	74,7185	43,5547
KNN; cosseno; não; 19	78,0307	74,7200	40,1189	62,4601	74,4776	43,5516
KNN; cosseno; não; 25	77,5429	73,9837	40,9651	61,4740	74,1132	44,5348
KNN; cosseno; não; 35	76,3144	74,1057	41,4694	60,8642	72,1469	45,6384
KNN; cosseno; não; 45	76,3083	73,7413	41,1006	61,2285	71,4137	45,8838
KNN; cosseno; não; 55	75,9410	72,5113	42,0837	61,2301	69,8223	45,1475
KNN; cosseno; sim; 1	65,5285	67,9720	37,1059	57,9027	66,8684	32,8832
KNN; cosseno; sim; 3	73,2490	71,4092	39,2241	61,0945	71,2843	35,8235
KNN; cosseno; sim; 5	75,2153	72,8771	39,9631	62,6965	74,8419	41,8488
KNN; cosseno; sim; 7	77,4195	73,6149	42,0813	64,5438	76,0690	41,7239
KNN; cosseno; sim; 9	77,1725	73,9837	40,3079	64,6612	75,0888	41,8383
KNN; cosseno; sim; 11	77,5474	73,8663	41,0345	64,1719	74,5995	41,3580
KNN; cosseno; sim; 13	77,6694	73,6179	40,6404	63,5592	75,0964	40,3689
KNN; cosseno; sim; 15	77,7883	73,4959	40,9696	63,8061	74,4776	40,2424
KNN; cosseno; sim; 17	77,9088	74,2337	41,4664	63,3198	74,9654	40,3674
KNN; cosseno; sim; 19	78,2776	74,3556	41,2135	62,9524	75,0873	40,7377
KNN; cosseno; sim; 25	77,6649	74,7215	42,3156	62,9464	74,3601	40,7377
KNN; cosseno; sim; 35	76,5613	73,7368	42,8154	62,2117	72,6423	41,5914
KNN; cosseno; sim; 45	76,5583	74,2291	42,2011	62,7040	71,6591	41,4679
KNN; cosseno; sim; 55	76,8067	74,3511	43,4342	62,4586	70,0662	41,4694

Fonte: Dados da pesquisa.

Tabela 17 – Acurácia máxima na versão “word” das coleções: 1º e 2º níveis.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	65,5119	11,8925	40,2695	57,1620	71,4107	28,3439
J48; 0,25	65,1596	36,6877	40,8808	54,5905	60,6128	39,7561
IMBHN ^C ; 0,01; 0,01	72,1424	27,7401	49,5815	66,3595	76,5613	31,0313
IMBHN ^C ; 0,05; 0,01	72,1424	27,7401	49,5815	66,3595	76,5613	31,0313
IMBHN ^C ; 0,1; 0,01	72,1424	27,7401	49,5815	66,3595	76,5613	31,0313
IMBHN ^C ; 0,5; 0,01	72,1424	27,7401	49,5815	66,3595	76,5613	31,0313
IMBHN ^R ; 0,01; 0,01	76,9226	29,0756	52,7718	68,0819	67,2298	31,0419
IMBHN ^R ; 0,05; 0,01	76,9226	29,0756	52,7718	68,0819	67,2298	31,0419
IMBHN ^R ; 0,1; 0,01	76,9226	29,0756	52,7718	68,0819	67,2298	31,0419
IMBHN ^R ; 0,5; 0,01	76,9226	29,0756	52,7718	68,0819	67,2298	31,0419
KNN; euclidiana; não; 1	56,1823	27,7236	38,8949	46,3177	66,8609	41,2331
KNN; euclidiana; não; 3	48,7562	23,3092	39,0319	41,0468	72,5128	41,3490
KNN; euclidiana; não; 5	48,3990	24,2894	40,9907	43,8300	73,7459	45,1551
KNN; euclidiana; não; 7	49,7906	22,0837	41,5959	41,7241	74,3601	45,5224
KNN; euclidiana; não; 9	47,3153	23,4357	42,5760	41,9648	74,2352	45,8868
KNN; euclidiana; não; 11	46,9828	24,6507	42,7040	44,4249	74,5995	46,3746
KNN; euclidiana; não; 13	46,6502	25,0286	42,9494	45,0286	75,3342	46,3791
KNN; euclidiana; não; 15	46,2931	23,5547	43,4342	45,5224	74,6010	46,2587
KNN; euclidiana; não; 17	46,6502	24,2879	43,1873	46,8759	75,5781	46,0132
KNN; euclidiana; não; 19	45,9360	25,2740	43,9250	47,7326	76,7992	46,1382
KNN; euclidiana; não; 25	41,6872	29,8118	43,5607	45,3899	75,6986	45,4035
KNN; euclidiana; não; 35	39,9261	32,6347	43,3198	45,2605	76,7992	46,2541
KNN; euclidiana; não; 45	38,4975	32,6393	43,0789	43,5516	75,9425	44,4173
KNN; euclidiana; não; 55	35,7143	32,8847	43,1978	44,0455	75,6956	45,7648
KNN; euclidiana; sim; 1	56,1823	27,7236	38,8949	46,3177	66,8609	41,2331
KNN; euclidiana; sim; 3	52,3276	27,1138	40,3749	43,8916	72,3924	42,4541
KNN; euclidiana; sim; 5	50,5419	27,4767	42,9464	43,5547	74,4836	45,2755
KNN; euclidiana; sim; 7	51,2438	24,0500	43,3138	44,1644	75,0918	45,7678
KNN; euclidiana; sim; 9	48,7562	25,5224	43,3092	43,9235	74,8494	45,3975
KNN; euclidiana; sim; 11	48,4113	26,4950	43,6871	45,2830	75,0949	46,5026
KNN; euclidiana; sim; 13	47,7463	26,7450	44,0545	46,2587	75,0918	46,1352
KNN; euclidiana; sim; 15	47,7094	26,2496	43,8091	46,8759	74,7230	46,3791
KNN; euclidiana; sim; 17	48,4113	26,9979	43,5562	48,2294	75,3327	46,3791

Fonte: Dados da pesquisa.

Tabela 18 – Acurácia máxima na versão “word” das coleções: 1º e 2º níveis (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	47,3522	27,4812	44,2894	48,7112	76,9286	46,5041
KNN; euclidiana; sim; 25	45,2217	30,1807	44,1749	46,7404	75,5766	46,5071
KNN; euclidiana; sim; 35	42,0443	33,0021	43,4447	46,0012	76,9226	46,5011
KNN; euclidiana; sim; 45	42,0320	32,6393	43,3168	44,2894	76,5553	46,6275
KNN; euclidiana; sim; 55	37,7956	32,8847	42,8335	43,6826	75,8205	46,7465
KNN; cosseno; não; 1	62,6108	24,7952	41,4709	55,9500	70,6700	38,0292
KNN; cosseno; não; 3	70,4201	24,1689	42,3397	60,6083	74,4761	37,5384
KNN; cosseno; não; 5	74,4806	28,6013	47,3698	62,5655	77,5474	38,0202
KNN; cosseno; não; 7	76,1909	27,2403	48,5848	65,5194	76,8022	39,1313
KNN; cosseno; não; 9	78,0247	27,2328	49,9443	64,7802	76,9316	39,0184
KNN; cosseno; não; 11	77,5324	27,1108	48,2415	65,8823	76,5643	35,9455
KNN; cosseno; não; 13	77,7838	28,0924	49,9262	65,6323	77,1755	39,2547
KNN; cosseno; não; 15	76,3159	27,6061	50,4426	63,9295	76,4363	39,1388
KNN; cosseno; não; 17	76,6862	27,8500	48,9702	64,8976	76,1894	44,0455
KNN; cosseno; não; 19	76,3189	27,1093	49,9443	64,1629	75,9470	43,9265
KNN; cosseno; não; 25	76,5658	27,3713	49,8088	64,2879	75,4562	44,2909
KNN; cosseno; não; 35	74,9669	30,9289	49,4625	62,9344	74,8464	45,3914
KNN; cosseno; não; 45	74,1087	33,0096	49,6914	62,6905	72,8802	45,2710
KNN; cosseno; não; 55	72,7567	33,0066	50,4276	62,3231	72,8756	44,5348
KNN; cosseno; sim; 1	62,6108	24,7952	40,1189	55,7046	70,6700	28,9566
KNN; cosseno; sim; 3	69,6868	26,3866	42,8350	60,6067	74,9639	29,4444
KNN; cosseno; sim; 5	74,4791	28,6043	46,7495	63,4282	77,9148	29,8058
KNN; cosseno; sim; 7	76,3114	27,2373	48,7248	64,7862	77,2915	30,5540
KNN; cosseno; sim; 9	76,9241	28,4658	48,9627	64,9051	77,1755	30,9229
KNN; cosseno; sim; 11	76,9226	28,3409	49,5769	65,5164	76,6847	30,4231
KNN; cosseno; sim; 13	77,0461	28,5787	49,1945	65,8793	77,0521	31,1608
KNN; cosseno; sim; 15	77,1710	27,2298	49,0936	65,0211	76,5583	31,1653
KNN; cosseno; sim; 17	77,2945	27,6016	49,0906	65,6399	76,0659	31,1638
KNN; cosseno; sim; 19	76,8052	26,8579	49,4535	65,5179	75,8235	30,6760
KNN; cosseno; sim; 25	76,6862	26,7419	50,1746	65,2725	75,5781	31,0388
KNN; cosseno; sim; 35	75,3388	26,9813	49,7124	64,1644	75,0933	31,7736
KNN; cosseno; sim; 45	75,8235	26,9874	50,0542	63,7955	73,1241	31,8970
KNN; cosseno; sim; 55	75,4592	27,6001	50,6745	64,0394	72,8756	31,2843

Fonte: Dados da pesquisa.

Tabela 19 – Acurácia máxima na versão “id” das coleções: 1º e 2º níveis.

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
NB	64,4053	11,3054	34,3511	61,2195	71,6546	
J48; 0,25	64,6673	39,5227	40,1325	54,3556	60,9756	
IMBHN ^C ; 0,01; 0,01	73,8648	24,2984	43,7970	67,2313	76,8037	
IMBHN ^C ; 0,05; 0,01	73,8648	24,2984	43,7970	67,2313	76,8037	
IMBHN ^C ; 0,1; 0,01	73,8648	24,2984	43,7970	67,2313	76,8037	
IMBHN ^C ; 0,5; 0,01	73,8648	24,2984	43,7970	67,2313	76,8037	
IMBHN ^R ; 0,01; 0,01	77,7838	29,2036	51,1668	66,9949	67,8470	
IMBHN ^R ; 0,05; 0,01	77,7838	29,2036	51,1668	66,9949	67,8470	
IMBHN ^R ; 0,1; 0,01	77,7838	29,2036	51,1668	66,9949	67,8470	
IMBHN ^R ; 0,5; 0,01	77,7838	29,2036	51,1668	66,9949	67,8470	
KNN; euclidiana; não; 1	52,5862	25,8838	39,2788	45,2586	69,0756	
KNN; euclidiana; não; 3	45,9113	26,9813	38,0367	43,4729	71,3023	
KNN; euclidiana; não; 5	52,9926	28,8241	37,4300	42,3291	73,6269	
KNN; euclidiana; não; 7	51,5887	28,9476	40,8702	42,3336	75,5902	
KNN; euclidiana; não; 9	49,4335	30,6730	41,3731	44,8013	76,5733	
KNN; euclidiana; não; 11	47,3645	30,6715	40,5134	46,8790	76,3264	
KNN; euclidiana; não; 13	45,6034	30,4291	41,2466	46,2662	75,3403	
KNN; euclidiana; não; 15	44,1995	31,7781	41,7389	46,3881	77,0611	
KNN; euclidiana; não; 17	41,7241	31,1593	41,8473	46,3896	75,8311	
KNN; euclidiana; não; 19	43,1404	32,9931	41,3565	45,5269	75,9606	
KNN; euclidiana; não; 25	44,5074	32,2674	43,6826	45,4095	76,6968	
KNN; euclidiana; não; 35	41,7241	32,7642	44,3014	44,7983	76,0825	
KNN; euclidiana; não; 45	39,9507	32,8847	42,4631	44,9082	74,4851	
KNN; euclidiana; não; 55	37,1429	32,8847	43,0819	44,4204	73,8723	
KNN; euclidiana; sim; 1	52,5862	25,8838	39,2788	45,2586	69,0756	
KNN; euclidiana; sim; 3	51,9458	26,8564	39,2758	45,2463	72,0370	
KNN; euclidiana; sim; 5	56,8966	28,5742	39,1524	43,1527	74,3601	
KNN; euclidiana; sim; 7	54,4458	28,8211	42,8380	43,8136	76,3264	
KNN; euclidiana; sim; 9	49,8276	29,4429	42,8485	45,9048	76,5718	
KNN; euclidiana; sim; 11	47,7340	30,6655	42,4752	47,6197	76,4469	
KNN; euclidiana; sim; 13	45,9483	30,0602	41,8609	47,6167	76,0750	
KNN; euclidiana; sim; 15	44,5443	31,4122	42,1078	47,6167	77,5504	
KNN; euclidiana; sim; 17	45,6281	31,2843	41,8594	47,1213	76,4438	

Fonte: Dados da pesquisa.

Tabela 20 – Acurácia máxima na versão “id” das coleções: 1º e 2º níveis (cont.).

Configuração	BoW	LDA	CIRT.5 L	CIRT.5 E-L	BoV_Wiki	BoV_NASARI
KNN; euclidiana; sim; 19	47,7463	32,2599	41,7299	46,9994	76,5733	
KNN; euclidiana; sim; 25	46,6502	31,6531	43,1933	45,6504	76,6953	
KNN; euclidiana; sim; 35	43,4606	32,8847	44,1765	46,3957	76,3279	
KNN; euclidiana; sim; 45	42,0443	32,6408	42,5881	46,2617	75,0979	
KNN; euclidiana; sim; 55	39,2365	33,1286	43,6962	45,8973	74,2412	
KNN; cosseno; não; 1	65,5134	24,1704	34,8284	57,1665	67,1168	
KNN; cosseno; não; 3	68,2174	27,3622	35,8311	57,4375	71,0449	
KNN; cosseno; não; 5	71,0358	27,3683	39,3812	60,6323	72,0295	
KNN; cosseno; não; 7	75,5812	29,2051	41,0915	62,2313	72,6408	
KNN; cosseno; não; 9	76,0720	30,7995	41,9587	61,9693	73,7504	
KNN; cosseno; não; 11	76,1969	29,5709	44,0425	65,5269	73,7534	
KNN; cosseno; não; 13	75,9515	31,2767	44,5363	65,2800	74,4866	
KNN; cosseno; não; 15	75,7076	31,2858	46,3806	65,0437	74,8555	
KNN; cosseno; não; 17	74,9684	32,1409	45,1460	64,9127	74,6131	
KNN; cosseno; não; 19	76,5613	32,1439	45,2725	66,3851	76,2105	
KNN; cosseno; não; 25	76,9256	33,4974	44,6612	63,5652	75,3523	
KNN; cosseno; não; 35	76,1894	32,6438	43,1918	63,9310	74,6116	
KNN; cosseno; não; 45	74,8434	31,6606	43,4417	64,0500	73,1451	
KNN; cosseno; não; 55	74,4731	31,2873	43,4342	63,9295	72,6468	
KNN; cosseno; sim; 1	65,3899	24,1704	34,3797	56,7992	67,1168	
KNN; cosseno; sim; 3	69,8118	27,6167	37,9298	59,2653	71,0403	
KNN; cosseno; sim; 5	72,3878	26,2587	40,3689	61,6110	72,1515	
KNN; cosseno; sim; 7	75,5812	27,6107	42,6950	62,7115	73,3785	
KNN; cosseno; sim; 9	76,0705	29,2096	43,0653	62,8305	73,3845	
KNN; cosseno; sim; 11	76,3219	29,0816	44,4219	66,0223	74,3707	
KNN; cosseno; sim; 13	76,1954	29,2051	45,1581	65,1566	74,3677	
KNN; cosseno; sim; 15	76,0705	29,5724	45,1566	65,1626	75,1039	
KNN; cosseno; sim; 17	76,1939	30,8025	45,2725	64,6703	75,2273	
KNN; cosseno; sim; 19	76,1924	30,3117	45,8898	65,8958	76,0870	
KNN; cosseno; sim; 25	76,4393	31,4152	44,9036	63,7955	74,9834	
KNN; cosseno; sim; 35	76,4288	31,4137	43,9310	63,6856	74,4881	
KNN; cosseno; sim; 45	75,4517	30,1822	43,3213	63,4402	73,1436	
KNN; cosseno; sim; 55	75,9425	30,0647	44,1704	64,1734	72,4014	

Fonte: Dados da pesquisa.