UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Aspectos semânticos na representação de textos para classificação automática

Roberta Akemi Sinoara

Tese de Doutorado 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: _____

Roberta Akemi Sinoara

Aspectos semânticos na representação de textos para classificação automática

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora 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
Junho de 2018

Esse trabalho foi desenvolvido com o apoio da FAPESP: processos no. 2013/14757-6 e 2016/07620-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade dos autores e não necessariamente refletem a visão da FAPESP.

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

S617a Sinoara, Roberta Akemi
Aspectos semânticos na representação de textos
para classificação automática / Roberta Akemi
Sinoara; orientador Solange Oliveira Rezende. --
São Carlos, 2018.
209 p.

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

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

Roberta Akemi Sinoara

Semantic aspects in the representation of texts for automatic classification

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate 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
June 2018

This work was supported by grants #2013/14757-6 and #2016/07620-2, São Paulo Research Foundation (FAPESP). The opinions, hypotheses and conclusions or recommendations expressed in this material are the authors' own and do not necessarily reflect the views of FAPESP.

AGRADECIMENTOS

Agradeço por todas as oportunidades que tive e a todas as pessoas que conheci.

Agradeço à minha família, em especial, meus pais Eliane e Mario, meus irmãos Fabio e Guilherme, e meu marido Nabil. Por todo amor, dedicação, apoio e compreensão!

Agradeço à minha orientadora Solange Rezende, amiga e incentivadora. Pela orientação acadêmica, profissional e pessoal, pela atenção, e por sempre procurar entender as características individuais de cada um de seus alunos!

Agradeço à Professora Maria Carolina Monard e aos amigos do LABIC, aqueles da minha primeira passagem pela pesquisa. Por deixarem uma marca especial na minha vida, que me fez querer voltar. E agradeço a todos os novos amigos do LABIC, aqueles que conheci nos últimos anos, pela troca de conhecimento e experiências, pelo companheirismo, e pelas conversas descontraídas na hora do café. Agradeço a gentil ajuda no início do doutorado, as parcerias, os trabalhos conjuntos e as revisões de Bruno Nogueira, Fabiano Fernandes, Marcos Domingues, Rafael Giusti, Ricardo Marcacini, Antonio Parmezan, Renan de Padua, Jorge Valverde-Rebaza, João Antunes, Ricardo Scheicher e Vitor Tonon. Agradecimentos especiais ao Rafael Rossi, pela parceria, várias dicas e toda a ajuda durante o desenvolvimento deste trabalho. E às meninas, Camila Sundermann, Ivone Matsuno e Jessica de Souza, pela amizade, apoio, carinho, carinhas e tudo mais! ;) :* :D

Agradeço ao Professor Roberto Navigli, da Sapienza - Università di Roma, e aos integrantes do seu grupo Linguistic Computing Laboratory, Alessandro Raganato, Claudio Delli Bovi, Francesco Cecconi, Ignacio Iacobacci, Tommaso Pasini e Valentina Pyatkin, que me receberam tão bem e tornaram minha visita muito proveitosa. Agradecimento especial ao José Camacho Collados, pela atenção e pelo trabalho conjunto. *Grazie, ragazzi! I wish you a lot of pasticcini on your journey!*

Agradeço a todos que contribuíram, direta ou indiretamente, para o desenvolvimento deste trabalho. À Universidade de São Paulo, aos professores e funcionários do ICMC-USP, ao Anandsing Dwarkasing, à Martina Vizzani, aos membros das bancas do exame de qualificação e da defesa, e aos pesquisadores que disponibilizaram os resultados de seus trabalhos possibilitando o desenvolvimento deste.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que por meio dos processos 2013/14757-6 e 2016/07620-2, apoiou financeiramente a realização desta pesquisa e a sua divulgação em conferências científicas internacionais.

Agradeço também aos *amici brasiliani*, Bruna Brogin, Hieda Adriana, Rafael Garcia, Robson Vieira e Thyago Nepomuceno, minha família em Roma! Espero que nossos caminhos voltem a se cruzar, desse ou do outro lado do oceano! ;)

Por fim, agradeço aos desconhecidos Malcolm Gladwell, Paul Hewson, David Evans, Larry Mullen Jr. e Adam Clayton, cujas obras me libertaram e me motivaram.

RESUMO

SINOARA, R. A. **Aspectos semânticos na representação de textos para classificação automática**. 2018. 209 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Dada a grande quantidade e diversidade de dados textuais sendo criados diariamente, as aplicações do processo de Mineração de Textos são inúmeras e variadas. Nesse processo, a qualidade da solução final depende, em parte, do modelo de representação de textos adotado. Por se tratar de textos em língua natural, relações sintáticas e semânticas influenciam o seu significado. No entanto, modelos tradicionais de representação de textos se limitam às palavras, não sendo possível diferenciar documentos que possuem o mesmo vocabulário, mas que apresentam visões diferentes sobre um mesmo assunto. Nesse contexto, este trabalho foi motivado pela diversidade das aplicações da tarefa de classificação automática de textos, pelo potencial das representações no modelo espaço-vetorial e pela lacuna referente ao tratamento da semântica inerente aos dados em língua natural. O seu desenvolvimento teve o propósito geral de avançar as pesquisas da área de Mineração de Textos em relação à incorporação de aspectos semânticos na representação de coleções de documentos. Um mapeamento sistemático da literatura da área foi realizado e os problemas de classificação foram categorizados em relação à complexidade semântica envolvida. Aspectos semânticos foram abordados com a proposta, bem como o desenvolvimento e a avaliação de sete modelos de representação de textos: (i) *gBoED*, modelo que incorpora a semântica obtida por meio de conhecimento do domínio; (ii) *Uni-based*, modelo que incorpora a semântica por meio da desambiguação lexical de sentidos e hiperônimos de conceitos; (iii) *SR-based Terms* e *SR-based Sentences*, modelos que incorporam a semântica por meio de anotações de papéis semânticos; (iv) *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, modelos que incorporam a semântica por meio de desambiguação lexical de sentidos e *embeddings* de palavras e conceitos. Representações de coleções de documentos geradas com os modelos propostos e outros da literatura foram analisadas e avaliadas na classificação automática de textos, considerando *datasets* de diferentes níveis de complexidade semântica. As propostas *gBoED*, *Uni-based*, *SR-based Terms* e *SR-based Sentences* apresentam atributos mais expressivos e possibilitam uma melhor interpretação da representação dos documentos. Já as propostas *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec* incorporam, de maneira latente, a semântica obtida de *embeddings* geradas a partir de uma grande quantidade de documentos externos. Essa propriedade tem um impacto positivo na performance de classificação.

Palavras-chave: Representação de textos, Semântica, Classificação de textos, Mineração de Textos.

ABSTRACT

SINOARA, R. A. **Semantic aspects in the representation of texts for automatic classification.** 2018. 209 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Text Mining applications are numerous and varied since a huge amount of textual data are created daily. The quality of the final solution of a Text Mining process depends, among other factors, on the adopted text representation model. Despite the fact that syntactic and semantic relations influence natural language meaning, traditional text representation models are limited to words. The use of such models does not allow the differentiation of documents that use the same vocabulary but present different ideas about the same subject. The motivation of this work relies on the diversity of text classification applications, the potential of vector space model representations and the challenge of dealing with text semantics. Having the general purpose of advance the field of semantic representation of documents, we first conducted a systematic mapping study of semantics-concerned Text Mining studies and we categorized classification problems according to their semantic complexity. Then, we approached semantic aspects of texts through the proposal, analysis, and evaluation of seven text representation models: (i) *gBoED*, which incorporates text semantics by the use of domain expressions; (ii) *Uni-based*, which takes advantage of word sense disambiguation and hypernym relations; (iii) *SR-based Terms* and *SR-based Sentences*, which make use of semantic role labels; (iv) *NASARI_{docs}*, *Babel2Vec* and *NASARI+Babel2Vec*, which take advantage of word sense disambiguation and embeddings of words and senses. We analyzed the expressiveness and interpretability of the proposed text representation models and evaluated their classification performance against different literature models. While the proposed models *gBoED*, *Uni-based*, *SR-based Terms* and *SR-based Sentences* have improved expressiveness, the proposals *NASARI_{docs}*, *Babel2Vec* and *NASARI+Babel2Vec* are latently enriched by the embeddings' semantics, obtained from the large training corpus. This property has a positive impact on text classification performance.

Keywords: Text representation, Semantics, Text classification, Text mining.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo ilustrativo de uma <i>bag-of-words</i> dos documentos $D1$, $D2$, $D3$ e $D4$	27
Figura 2 – Processo de Mineração de Textos	39
Figura 3 – Ilustração da representação de uma coleção de N documentos como uma matriz documento-termo	41
Figura 4 – Esquema ilustrativo da classificação automática de textos por meio de aprendizado indutivo supervisionado	45
Figura 5 – Matriz de confusão para a classe c_i	47
Figura 6 – Formulário para extração de informações durante a condução do mapeamento sistemático da literatura.	55
Figura 7 – Resultados do mapeamento dos estudos de mineração de textos que consideram a semântica.	56
Figura 8 – Exemplo 1: Documentos ilustrativos	58
Figura 9 – Exemplo 2: Documentos ilustrativos	59
Figura 10 – Largura de silhueta dos cenários das coleções <i>BBC</i> e <i>SE-ABSA 15</i>	68
Figura 11 – Largura de silhueta dos cenários da coleção <i>BEST sports</i>	69
Figura 12 – Exemplo 3: Documentos ilustrativos	75
Figura 13 – Representações dos documentos $D1$, $D2$ e $D3$ da Figura 12	76
Figura 14 – Esquema da representação de coleção de documentos <i>gBoED</i>	78
Figura 15 – Construção da representação <i>Uni-based</i>	83
Figura 16 – Estruturas de predador e argumentos identificadas na sentença “ <i>It is rare that hotels in this class serve hot meals, yet they do!</i> ”.	92
Figura 17 – Exemplo 4: Documentos ilustrativos	94
Figura 18 – Exemplo 4: Documentos da Figura 17 com anotações de classes morfossintáticas e papéis semânticos.	94
Figura 19 – Representações dos documentos $D1$, $D2$ e $D3$ da Figura 17	95
Figura 20 – <i>Box plots</i> das acurácias de classificação, por representação utilizada, para cada <i>dataset</i>	99
Figura 21 – Diagrama de diferença crítica entre as representações avaliadas	102
Figura 22 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>SR-based Terms</i> , utilizando a estratégia SC.	104
Figura 23 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>SR-based Terms</i> , utilizando a estratégia WSC.	105

Figura 24 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>SR-based Terms</i> , utilizando a estratégia MC.	106
Figura 25 – Diagrama de diferença crítica entre as representações avaliadas	108
Figura 26 – Processo de construção das representações <i>NASARI_{docs}</i> , <i>Babel2Vec</i> e <i>NA-SARI+Babel2Vec</i> - Etapa 1	117
Figura 27 – Processo de construção das representações <i>NASARI_{docs}</i> , <i>Babel2Vec</i> e <i>NA-SARI+Babel2Vec</i> - Etapa 2	118
Figura 28 – Relações entre similaridades atribuídas por anotadores e similaridades calculadas por distância de vetores	130
Figura 29 – <i>Box plots</i> das acurácias de classificação, por representação utilizada, para cada <i>dataset</i>	133
Figura 30 – Relação entre $F1^\mu$ e $F1^M$ por representação utilizada	134
Figura 31 – Diagrama de diferença crítica entre as representações avaliadas, considerando a medida Acurácia	138
Figura 32 – Diagrama de diferença crítica entre as representações avaliadas, considerando a medida $F1^M$	138
Figura 33 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>NA-SARI+Babel2Vec</i> , utilizando a estratégia SC	140
Figura 34 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>NA-SARI+Babel2Vec</i> , utilizando a estratégia WSC	141
Figura 35 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>NA-SARI+Babel2Vec</i> , utilizando a estratégia MC	142
Figura 36 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>Babel2Vec</i> , utilizando a estratégia SC	143
Figura 37 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>Babel2Vec</i> , utilizando a estratégia WSC	144
Figura 38 – <i>Box plots</i> das acurácias de cada <i>dataset</i> para a combinação de BOW e <i>Babel2Vec</i> , utilizando a estratégia MC	145
Figura 39 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e <i>NASARI+Babel2Vec</i> - Acurácia	147
Figura 40 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e <i>NASARI+Babel2Vec</i> - $F1^M$	148
Figura 41 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e <i>Babel2Vec</i> - Acurácia	150
Figura 42 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e <i>Babel2Vec</i> - $F1^M$	151
Figura 43 – Diagrama de diferença crítica entre as melhores as representações e combinações avaliadas - Acurácia	152

Figura 44 – Diagrama de diferença crítica entre as melhores as representações e combinações avaliadas - $F1^M$	152
Figura 45 – Guia para rotulação da coleção <i>BS-Top4</i>	180
Figura 46 – Configuração do processo de rotulação da coleção <i>BS-Top4</i> na ferramenta RotuLABIC	181
Figura 47 – <i>Box plots</i> para $F1^M$ por representação utilizada, para cada <i>dataset</i>	200
Figura 48 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e NA-SARI+Babel2Vec, utilizando a estratégia SC	203
Figura 49 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e NA-SARI+Babel2Vec, utilizando a estratégia WSC	204
Figura 50 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e NA-SARI+Babel2Vec, utilizando a estratégia MC	205
Figura 51 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e Babel2Vec, utilizando a estratégia SC	206
Figura 52 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e Babel2Vec, utilizando a estratégia WSC	207
Figura 53 – <i>Box plots</i> para $F1^M$ de cada <i>dataset</i> para a combinação de BOW e Babel2Vec, utilizando a estratégia MC	208

LISTA DE QUADROS

Quadro 1 – Problemas tratados em coleções de textos de <i>benchmarking</i> apresentadas por Rossi, Marcacini e Rezende (2013)	62
Quadro 2 – Descrição das coleções de textos e <i>datasets</i>	67
Quadro 3 – Conjunto de etiquetas morfossintáticas do Penn Treebank.	92
Quadro 4 – Descrições dos <i>synsets</i> identificados na sentença “ <i>The Toshiba Net book operates very well.</i> ”	119
Quadro 5 – <i>Synsets</i> identificados no <i>Documento A</i>	124
Quadro 6 – As 5 palavras ou <i>synsets</i> mais próximos às representações do <i>Documento A</i>	125
Quadro 7 – As 5 palavras ou <i>synsets</i> mais próximos às representações do <i>Documento B</i>	127
Quadro 8 – Comparação da desambiguação para o documento em português	197
Quadro 9 – Comparação da desambiguação para o documento em inglês	199

LISTA DE TABELAS

Tabela 1 – Melhores acurácias para as coleções <i>BS-Top4</i> e <i>SE-ABSA15</i>	80
Tabela 2 – Melhores acurácias para as coleções <i>BS-Top4</i> e <i>SE-ABSA15</i>	84
Tabela 3 – Características dos conjuntos de atributos das representações propostas	96
Tabela 4 – Correlação das medidas de performance de classificação utilizando as abordagens <i>micro-averaging</i> (acurácia) e <i>macro-averaging</i>	100
Tabela 5 – Melhores acurácias para as coleções <i>BS-Top4</i> e <i>SE-ABSA15</i>	101
Tabela 6 – Similaridade entre as representações de documentos com e sem MCS	128
Tabela 7 – Correlação entre similaridades atribuídas por anotadores e similaridades calculadas por distância de vetores	129
Tabela 8 – Correlação de Pearson entre $F1^u$ e $F1^M$	135
Tabela 9 – Melhores acurácias para as coleções <i>BBC</i> , <i>SE-ABSA15</i> e <i>BS-Top4</i>	136
Tabela 10 – Melhores acurácias para as coleções <i>BBC</i> , <i>SE-ABSA15</i> e <i>BS-Top4</i> para combinação de classificadores	146
Tabela 11 – Distribuição dos documentos da coleção <i>BEST sports - Top 4</i>	183
Tabela 12 – Maiores valores de $F1^M$ para as coleções <i>BBC</i> , <i>SE-ABSA15</i> e <i>BS-Top4</i>	201
Tabela 13 – Melhores $F1^M$ para as coleções <i>BBC</i> , <i>SE-ABSA15</i> e <i>BS-Top4</i> para combinação de classificadores	209

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
BoED	<i>bag-of-expressions-of-domain</i>
BOW	<i>bag-of-words</i>
IMBHN	<i>Inductive Model based on Bipartite Heterogeneous Networks</i>
k-NN	<i>k-Nearest Neighbors</i>
LDA	<i>Latent Dirichlet Allocation</i>
MC	<i>Most Confident</i>
MCS	<i>Most common sense</i>
MNB	<i>Multinomial Naive Bayes</i>
MT	Mineração de Textos
NB	<i>Naive Bayes</i>
PLN	Processamento de Língua Natural
PLSI	<i>Probabilistic Latent Semantic Indexing</i>
SC	<i>Sum of Confidences</i>
SVM	<i>Support Vector Machine</i>
TF	<i>term frequency</i>
TF-IDF	<i>term frequency - inverse document frequency</i>
WSC	<i>Weighted Sum of Confidences</i>

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Contextualização	25
1.2	Motivação e lacunas	28
1.3	Questões de pesquisa e objetivos	30
1.4	Principais resultados	31
1.5	Organização do texto	33
2	FUNDAMENTOS E DEFINIÇÃO DO PROBLEMA	37
2.1	Considerações iniciais	37
2.2	Mineração de textos	38
2.2.1	<i>Pré-processamento e representação de documentos</i>	40
2.2.2	<i>Classificação automática de textos</i>	44
2.3	Análise semântica	48
2.4	Mapeamento da semântica na Mineração de Textos	52
2.4.1	<i>Protocolo do mapeamento</i>	53
2.4.2	<i>Resultados do mapeamento</i>	54
2.5	Definição do problema: níveis de complexidade semântica na mine- ração de textos	58
2.5.1	<i>Níveis de complexidade semântica</i>	58
2.5.2	<i>Coleções de documentos e datasets utilizados</i>	64
2.5.3	<i>Abordagens adotadas em direção à incorporação de aspectos se- mânticos na representação de textos</i>	67
2.6	Considerações finais	70
3	EXPLORAÇÕES INICIAIS: REPRESENTAÇÕES BASEADAS EM CONHECIMENTO DO DOMÍNIO E HIPERÔNIMOS	73
3.1	Considerações iniciais	73
3.2	Representação baseada em expressões do domínio	74
3.2.1	<i>Modelo proposto para representação de coleções de documentos</i>	77
3.2.2	<i>Avaliação experimental - representação gBoED</i>	78
3.2.2.1	<i>Datasets, representações de textos e configuração experimental</i>	78
3.2.2.2	<i>Resultados - gBoED</i>	79
3.3	Representação baseada em hiperônimos	79

3.3.1	<i>Modelo proposto para representação de coleções de documentos</i>	82
3.3.2	<i>Avaliação experimental - representação Uni-based</i>	83
3.3.2.1	<i>Datasets, representações de textos e configuração experimental</i>	83
3.3.2.2	<i>Resultados - Uni-based</i>	83
3.4	<i>Considerações finais</i>	85
4	REPRESENTAÇÕES BASEADAS EM PAPÉIS SEMÂNTICOS	87
4.1	<i>Considerações iniciais</i>	87
4.2	<i>Trabalhos relacionados</i>	89
4.3	<i>Modelos propostos para representação de coleções de documentos</i>	91
4.4	<i>Análise das representações <i>SR-based Terms</i> e <i>SR-based Sentences</i></i>	95
4.5	<i>Avaliação experimental - representações baseadas em papéis semânticos</i>	97
4.5.1	<i>Datasets, representações de textos e configuração experimental</i>	97
4.5.2	<i>Resultados - papéis semânticos</i>	98
4.6	<i>Considerações finais</i>	107
5	REPRESENTAÇÕES BASEADAS EM EMBEDDINGS	111
5.1	<i>Considerações iniciais</i>	111
5.2	<i>Trabalhos relacionados</i>	112
5.3	<i>Recursos linguísticos e ferramentas utilizadas</i>	115
5.4	<i>Modelos propostos para representação de coleções de documentos</i>	116
5.5	<i>Análise das representações <i>NASARI+Babel2Vec</i>, <i>NASARI_{docs}</i> e <i>Babel2Vec</i></i>	120
5.5.1	<i>Proximidade a palavras e conceitos</i>	122
5.5.2	<i>Similaridade textual</i>	128
5.6	<i>Avaliação experimental - representações baseadas em embeddings</i>	131
5.6.1	<i>Datasets, representações de textos e configuração experimental</i>	131
5.6.2	<i>Resultados - embeddings</i>	132
5.7	<i>Considerações finais</i>	149
6	CONCLUSÕES	155
6.1	<i>Contribuições científicas</i>	155
6.2	<i>Publicações</i>	160
6.3	<i>Limitações e trabalhos futuros</i>	163
	REFERÊNCIAS	167
APÊNDICE A	PROCESSO DE ROTULAÇÃO DA COLEÇÃO <i>BEST SPORTS - TOP 4</i>	179

APÊNDICE B	CONFIGURAÇÃO EXPERIMENTAL	185
B.1	Geração das representações	185
B.1.1	<i>Representações de comparação (baselines)</i>	186
B.1.2	<i>Representações propostas</i>	188
B.2	Algoritmos de aprendizado supervisionado	191
B.3	Combinação de classificadores	193
B.4	Critérios de avaliação	194
APÊNDICE C	ANÁLISES E RESULTADOS EXTRAS	195
C.1	Análise do parâmetro <i>Matching Type</i> do sistema Babelfy	195
C.2	Resultados complementares do Capítulo 5 - Representações baseadas em <i>embeddings</i>	198

INTRODUÇÃO

1.1 Contextualização

Os avanços e a disseminação do uso de tecnologias de informação e comunicação têm causado um aumento expressivo na geração e armazenamento de dados em formato digital. Um levantamento do International Data Corporation realizado em 2014 ([TURNER *et al.*, 2014](#)) indicava que o universo digital estava dobrando de tamanho a cada dois anos e em 2020 alcançará 44 zettabytes (44 trilhões de gigabytes). Projeções realizadas pela mesma corporação em 2017 indicam que em 2025 serão gerados 165 zettabytes, sendo dez vezes maior do que o volume gerado em 2016, que foi de 16,1 zettabytes ([REINSEL; GANTZ; RYDNING, 2017](#)). Uma parte desse universo é composta por dados não estruturados, como os documentos textuais gerados internamente nas empresas, revisões e comentários sobre produtos e serviços em páginas da *web*, e *posts* em redes sociais. Considerando a análise desses dados, [Reinsel, Gantz e Rydning \(2017\)](#) apontam como uma tendência futura, que irá mudar o cenário da análise de dados, a aplicação de técnicas de Inteligência Artificial, como Aprendizado de Máquina (AM) e Processamento de Língua Natural (PLN).

Tendo em vista esse cenário, as técnicas de Mineração de Textos (MT) tornam-se essenciais para apoio à extração de conhecimento dos textos. Visando a identificação de padrões em textos escritos em língua natural, o processo de Mineração de Textos pode ser visto como uma sequência de etapas genéricas, que devem ser instanciadas de acordo com os dados disponíveis e o conhecimento que se espera obter ([AGGARWAL; ZHAI, 2012a; REZENDE, 2003](#)). Por exemplo, a partir de um conjunto de documentos rotulados pode-se empregar técnicas de classificação de textos com o objetivo de obter classificadores que relacionem novos documentos ao conjunto de classes previamente estabelecido ([SEBASTIANI, 2002](#)). Já em aplicações que têm como objetivo uma organização da informação textual, porém sem conhecimento prévio sobre as classes existentes, pode-se utilizar técnicas de agrupamento de dados ([AGGARWAL; ZHAI, 2012b](#)).

Para que os algoritmos de AM possam ser utilizados na Mineração de Textos, os dados não estruturados sofrem uma série de transformações para se obter uma representação estruturada dos mesmos. Os textos são normalmente representados no modelo espaço-vetorial (*vector space model*), formando uma matriz atributo-valor (também chamada de matriz documento-termo), na qual cada linha representa um documento e cada coluna representa um termo presente na coleção de documentos. As palavras são comumente utilizadas como termos, dando origem à representação chamada de *bag-of-words* (BOW). Esse formato é simples e possibilita o uso direto de métodos de AM, porém pode trazer grandes desafios para o processo de Mineração de Textos.

Um dos desafios da Mineração de Textos refere-se à esparsidade e à dimensionalidade da representação dos dados textuais. Uma coleção de documentos pode conter milhares de palavras enquanto que um de seus documentos pode ser formado por apenas um conjunto pequeno dessas palavras. Esse fato pode, por exemplo, tornar o processo de análise muito custoso computacionalmente, ou até mesmo inviável, além de afetar negativamente o resultado de alguns algoritmos utilizados para extração de conhecimento. Os problemas de alta dimensionalidade e esparsidade têm sido tratados no pré-processamento dos dados com atividades como remoção de *stopwords*, normalização de termos (como radicalização e lematização), uso de termos compostos (n-gramas) em substituição a termos simples, e aplicação de técnicas para extração ou seleção dos atributos mais relevantes (CONRADO *et al.*, 2014; ZHENG; AN; XU, 2013; NOGUEIRA; REZENDE, 2009; EBECKEN; LOPES; COSTA, 2003). Além de diminuir a dimensionalidade e/ou esparsidade da representação dos textos, essas atividades também podem levar ao aumento da representatividade dos atributos. No entanto, ainda existe um outro grande desafio relacionado com a semântica dos dados textuais.

Uma representação dos textos como um conjunto não ordenado de palavras, como a representação *bag-of-words*, ignora propriedades relacionadas à sintaxe e às relações semânticas existentes em textos em língua natural, como voz ativa e passiva, termos sinônimos e hiperônimos. Isso tem como consequência a perda de parte das informações contidas nos textos. O seguinte exemplo ilustra esse fato. Considere uma coleção formada por quatro documentos, *D1*, *D2*, *D3* e *D4*, apresentados a seguir.

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

Uma matriz que representa essa coleção de documentos, considerando o formato *bag-of-words* é apresentada na Figura 1.

Figura 1 – Exemplo ilustrativo de uma *bag-of-words* dos documentos *D1*, *D2*, *D3* e *D4*

	empr	alf	adquirir	bet	foi	compr
<i>D1</i>	2	1	1	1	0	0
<i>D2</i>	2	1	1	1	0	0
<i>D3</i>	2	1	1	1	1	0
<i>D4</i>	2	1	0	1	0	1

Fonte: Elaborada pelo autor.

Nota – Neste exemplo foi considerada a remoção de *stopwords*, os termos foram radicalizados e a frequência dos termos nos documentos foi usada como peso dos atributos.

Nesse exemplo simples é possível notar duas limitações da representação utilizada. Considerando a *bag-of-words*, os documentos *D1* e *D2* possuem a mesma representação e são considerados iguais, apesar de apresentarem sentidos opostos. Analisando as sentenças apenas por suas palavras, não é possível diferenciá-las. Porém, ao se considerar a sintaxe das sentenças, pode-se perceber que elas são opostas. As sentenças de *D1* e *D2* possuem o mesmo verbo e o sujeito de uma é o objeto da outra. Analisando um pouco mais a fundo, considerando a semântica das sentenças por meio dos papéis semânticos, pode-se identificar quais são os agentes de cada sentença. Assim, considerando os papéis semânticos, é possível perceber que *D1* é igual a *D3*, apesar de terem sujeitos e objetos opostos. Já considerando as relações semânticas entre as palavras (como sinonímia, por exemplo), pode-se perceber que *D4* expressa o mesmo evento de *D1* e *D3*.

Dependendo da aplicação, o tratamento adequado de informações semânticas dos textos pode levar a resultados mais adequados (AGGARWAL; ZHAI, 2012a). Jain (2010) afirma que a representação dos dados é um dos fatores que mais impacta na qualidade do resultado obtido. Apesar de Jain (2010) tratar o caso específico da tarefa de agrupamento de dados, o mesmo pode ser generalizado para as demais tarefas. Em se tratando de dados textuais, relações semânticas têm impacto sobre o significado do conteúdo dos documentos e podem servir para, por exemplo, diferenciar documentos que utilizam a mesmo vocabulário mas que apresentam ideias diferentes sobre um mesmo assunto.

Uma revisão da literatura realizada sobre estudos de Mineração de Textos que consideram aspectos semânticos revela que o uso de representações mais ricas é o foco de muitos estudos (SINOARA; ANTUNES; REZENDE, 2017). Grande parte dos trabalhos concentra-se na proposta e/ou uso de atributos mais elaborados para representar os documentos no modelo espaço-vetorial. Alguns trabalhos realizados na direção do uso da semântica na Mineração de Textos utilizam técnicas de modelagem de tópicos (*topic modeling*), como *Probabilistic Latent Semantic Indexing* (PLSI) e *Latent Dirichlet Allocation* (LDA), para obter atributos semanticamente mais ricos (AGGARWAL; ZHAI, 2012a; LU; MEI; ZHAI, 2011; LIU *et al.*, 2011; BLEI; NG; JORDAN, 2003). Tais atributos formam um espaço de semântica latente (*latent semantic space*), que é um espaço de vetores de dimensão fixa e normalmente baixa, no qual formas

alternativas de se expressar determinado conceito são projetadas para uma representação comum. Assim, tais representações lidam com a semântica dos textos de forma latente e reduzem ruídos causados por sinonímia e polissemia.

Ainda considerando a semântica latente, outra abordagem que apresenta resultados promissores é o uso de *embeddings*. *Word embeddings* têm ganhado atenção em pesquisas da área de semântica lexical (*lexical semantics*). Um destaque nessa área foi a proposta dos modelos de aprendizado *Continuous Bag-of-Words* e *Skip-gram* (MIKOLOV *et al.*, 2013a). A partir deste trabalho, vários outros foram desenvolvidos para estender essas abordagens e/ou aplicá-las a diferentes problemas. Inspirado pelos trabalhos relacionados a representações vetoriais de palavras, Le e Mikolov (2014) propuseram uma abordagem para aprendizado de vetores para itens linguísticos maiores, como sentenças ou parágrafos. Tal abordagem foi denominada *Paragraph Vector* e, posteriormente, *Doc2Vec* e apresentou resultados competitivos com métodos da literatura.

Além da semântica latente, o uso de conceitos com base em fontes externas de conhecimento, como a WordNet ou Wikipedia, e a aplicação de métodos de Processamento de Língua Natural são outras abordagens exploradas para enriquecer a representação de textos (SINOARA; ROSSI; REZENDE, 2016; KIM; HONG; CHANG, 2015; SINOARA *et al.*, 2014; SPANAKIS; SIOLAS; STAFYLOPATIS, 2012; HU *et al.*, 2009). Técnicas de PLN podem auxiliar a Mineração de Textos em relação ao tratamento da semântica, visto que essa área estuda tarefas de análise semântica que tratam a semântica de maneira mais profunda do que técnicas como PLSI e LDA. Esse é o caso da anotação de papéis semânticos, tarefa na qual documentos textuais são processados para identificar e rotular os papéis semânticos de cada termo de uma sentença. Outros exemplos de tarefas são o reconhecimento de entidades nomeadas, a desambiguação lexical dos sentidos e a resolução de correferências, tarefas que contribuem para uma melhor definição do conteúdo dos textos.

1.2 Motivação e lacunas

Apesar dos diversos trabalhos realizados em busca de alternativas para representação de textos, o impacto da semântica na representação de coleções de documentos têm sido pouco explorado na literatura. Já no contexto da representação de palavras por meio de *word embeddings*, Levy e Goldberg (2014) discutem brevemente os diferentes aspectos pelos quais pode-se verificar a similaridade entre palavras. Os autores distinguem a similaridade tópica (*topical similarity*) da similaridade funcional (*functional similarity*). Para ilustrar a diferença entre os dois tipos de similaridade, considere a palavra “Turing”¹. Considerando a similaridade tópica, palavras similares podem ser aquelas relacionadas a Turing e ao seu trabalho, como “determinístico”, “computabilidade”, “estado-finito”. Já considerando a similaridade funcional, palavras similares

¹ Exemplo baseado nos resultados obtidos por Levy e Goldberg (2014).

podem ser nomes de outros cientistas, como “Pauling” e “Hamming”. [Levy e Goldberg \(2014\)](#) variam o contexto utilizado para gerar *word embeddings*, obtendo diferentes vetores para as palavras. Dependendo do contexto utilizado, os vetores favorecem a identificação de um dos tipos de similaridade abordada: similaridade tópica ou funcional. Com o trabalho de [Levy e Goldberg \(2014\)](#) pode-se perceber a importância em se ter representações adequadas para se atingir os resultados esperados na tarefa de identificar palavras semelhantes (*word similarity task*). No entanto, apesar da relevância do tema, o impacto da variação de representações de acordo com os resultados esperados não tem sido explorado no contexto da representação de documentos e de tarefas como classificação e agrupamento de textos.

Por outro lado, no contexto da Mineração de Textos, [Aggarwal e Zhai \(2012a\)](#) chama a atenção para o problema da semântica dos textos e para a necessidade de se desenvolver representações mais ricas que a tradicional *bag-of-words* para problemas mais complexos. Por considerar que o entendimento de informação textual seja fundamental para a MT, os autores apresentam, como uma direção futura promissora, a representação de documentos além da *bag-of-words*, com o uso de métodos escaláveis e robustos de entendimento de língua natural. Apesar de várias frentes de pesquisa terem o foco em modelos alternativos de representação de textos, como as redes, o modelo espaço-vetorial ainda apresenta grande potencial e a vantagem de poder ser utilizado diretamente em algoritmos tradicionais de Aprendizado de Máquina. [Turney e Pantel \(2010\)](#) estabelecem o potencial desse modelo e a possibilidade de seu uso para problemas mais complexos com o desenvolvimento de novos métodos. Os autores acreditam que os limites do modelo espaço-vetorial ainda não foram atingidos e que limitações do modelo na representação da semântica dos textos podem ser superados em pesquisas futuras.

Considerando-se a representação do significado de textos em língua natural, [Widdows \(2004\)](#) apresenta o poder de modelos matemáticos, porém também afirma que toda abstração matemática usada como modelo linguístico contém simplificações. Segundo o autor, o desafio é encontrar simplificações inteligentes que possibilitem que as características principais do evento modelado sejam aparentes. Em cenários reais, existe uma grande gama de possibilidades de aplicação da MT, dada a grande variedade e diversidade de textos disponíveis. Nesse contexto, neste trabalho foram explorados alguns aspectos semânticos na representação de coleções de documentos. Considerando a observação de [Widdows \(2004\)](#) sobre as simplificações, neste trabalho buscou-se explorar diferentes abordagens em direção a representações semanticamente enriquecidas, visando tornar aparente aos algoritmos de AM diferentes padrões semânticos das coleções de documentos.

Para a avaliação da aplicação das diferentes representações de documentos no processo de MT, utilizou-se da tarefa de classificação de textos, visto a grande aplicabilidade desta tarefa em problemas reais. Segundo uma pesquisa realizada entre profissionais que trabalham com análise de textos, a classificação de documentos é uma das funcionalidades mais requeridas por usuários que desejam analisar textos ([GRIMES, 2014](#)). Assim, foram realizadas avaliações experimentais

das representações propostas com a aplicação de diferentes algoritmos de classificação.

Nesse contexto, este trabalho foi motivado pela diversidade das aplicações da tarefa de classificação automática de textos, pelo potencial das representações no modelo espaço-vetorial e pela lacuna referente ao tratamento da semântica inerente aos dados em língua natural. Assim como no trabalho de [Turney e Pantel \(2010\)](#), nesta tese o termo semântica é usado em um sentido geral, considerando o sentido ou significado de itens linguísticos, sejam eles palavras, expressões ou documentos completos. Significados mais específicos de semântica, como em Web Semântica ou processamento da semântica por meio de lógica formal, não são considerados. Neste trabalho, o interesse é principalmente identificar aspectos semânticos de textos em língua natural e avaliar seus impactos tanto em relação à interpretabilidade e expressividade das representações quanto em relação à performance obtida em diferentes problemas de classificação de textos.

1.3 Questões de pesquisa e objetivos

Este trabalho de doutorado foi desenvolvido com o propósito geral de avançar as pesquisas da área de Mineração de Textos, em especial na classificação automática de documentos, em relação à incorporação da semântica na representação de coleções de documentos. Assim, o desenvolvimento deste trabalho foi guiado por questões de pesquisa referentes ao estado atual das pesquisas de Mineração de Textos com foco na semântica e também referentes ao impacto da semântica tanto na representação de documentos quanto na tarefa de classificação automática. Essas questões de pesquisa são apresentadas a seguir.

Q1 Como a semântica é tratada no processo de Mineração de Textos?

Essa questão de pesquisa foi dividida em sete questões secundárias:

Q1.1 Quais são os domínios de aplicação que consideram a semântica dos textos de alguma maneira?

Q1.2 Quais idiomas estão sendo tratados?

Q1.3 Quais fontes externas são comumente utilizadas?

Q1.4 Em quais tarefas a semântica aparece com mais frequência?

Q1.5 Quais métodos e algoritmos são comumente utilizados?

Q1.6 Como os textos são representados?

Q1.7 Os usuários participam do processo de Mineração de Textos?

Q2 Que tipo de problema é tradicionalmente tratado pela comunidade de pesquisa?

Q3 Como a semântica interage com os diferentes problemas de classificação existentes?

Q4 A representação *bag-of-words* é adequada a qual tipo de problema de classificação?

Q5 Representações semanticamente enriquecidas podem levar a melhores resultados de classificação?

Q6 A combinação de representações semânticas com a *bag-of-words* pode levar a melhores resultados de classificação?

De acordo com as questões de pesquisa estabelecidas, foram definidos os seguintes objetivos para o desenvolvimento deste trabalho.

1. Mapear os trabalhos realizados na comunidade de Mineração de Textos que se preocupam com a semântica dos textos. Tradicionalmente, os dados textuais são representados com uma *bag-of-words*, tratando os documentos como um conjunto de palavras independentes e desconsiderando os diversos aspectos semânticos inerentes a esse tipo de dados. No entanto, nos últimos anos, diversas pesquisas surgiram apontando a existência de tratamento da semântica dos textos. Esse objetivo está relacionado à questão de pesquisa **Q1**.
2. Analisar o impacto da semântica na tarefa de classificação de textos e diferenciar, do ponto de vista da semântica, os problemas que podem ser tratados com essa tarefa. Cenários reais de aplicação da classificação automática de textos são bastante diversos. Existem diferentes fontes de dados e diferentes objetivos para a aplicação do processo de MT. Esse objetivo está relacionado às questões de pesquisa **Q2** e **Q3**.
3. Propor e desenvolver modelos de representação de coleções de documentos com base em informações semânticas de diferentes naturezas. Considerando-se que cada problema requer que diferentes aspectos ou características estejam presentes na representação para que sejam descobertos por algoritmos de extração de padrões, diferentes abordagens podem ser exploradas para a incorporação da semântica na representação dos dados textuais. Esse objetivo está relacionado às questões de pesquisa **Q4**, **Q5** e **Q6**.

1.4 Principais resultados

Os principais resultados que foram obtidos com o desenvolvimento deste projeto, visando atingir os objetivos propostos, são listados a seguir.

R1 Mapeamento sistemático da literatura sobre a Semântica na Mineração de Textos. Foi conduzida uma revisão sistemática da literatura desse amplo tema, abrangendo a literatura de Mineração de Textos que considera a semântica dos textos. Seguindo um protocolo bem definido, revisou-se 1.693 estudos e como resultado obteve-se um mapeamento do tema. A partir da realização desse mapeamento sistemático respondeu-se a questão de pesquisa de **Q1**, por meio das questões secundárias **Q1.1** a **Q1.7**.

R2 Identificação e definição de níveis de complexidade semântica em problemas de Mineração de Textos. Foi realizada uma análise de alguns dos fenômenos linguísticos que interferem no significado dos textos e das principais coleções de textos de *benchmarking* utilizadas pela comunidade de Mineração de Textos. Com essa análise foram definidos dois níveis de complexidade semântica para problemas de Mineração de Textos, chamados de organização por tópico e organização semântica. Com esse trabalho foram endereçadas as questões de pesquisa **Q2** e **Q3**.

R3 Criação de *datasets* para possibilitar a avaliação em problemas com diferentes complexidades semânticas. Durante o desenvolvimento deste projeto foram coletadas e analisadas diferentes coleções de textos. Para possibilitar a avaliação de representações de textos em problemas de diferentes níveis de complexidade semântica, foram selecionadas duas coleções de textos (uma com documentos em inglês e outra com documentos em português) e, para cada uma dessas coleções, foram gerados três *datasets* diferentes. Assim, para uma mesma coleção de documentos, desenvolveu-se *datasets* que representam objetivos de classificação distintos, correspondendo a problemas de organização por tópico ou de organização semântica. Esse resultado viabilizou a obtenção do resultado **R4**.

R4 Proposta, desenvolvimento e avaliação de modelos de representação de documentos considerando diferentes aspectos semânticos. A semântica foi abordada por quatro frentes diferentes, sendo propostos sete modelos de representação de coleções de documentos:

- *gBoED* - representação baseada em expressões do domínio;
- *Uni-based* - representação baseada em hiperônimos de conceitos;
- *SR-based Terms* e *SR-based Sentences* - representações baseadas em papéis semânticos;
- *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec* - representações baseadas em *embeddings* de palavras e conceitos.

Com o desenvolvimento, análise e avaliação experimental dessas representações, comparando-as com diferentes métodos da literatura, foram tratadas as questões de pesquisa **Q4**, **Q5** e **Q6**.

R5 Disponibilização dos recursos e ferramentas desenvolvidos. Considerando-se a continuidade desta pesquisa, a reprodução dos resultados e a aplicação dos métodos desenvolvidos, os resultados deste trabalho foram disponibilizados na seguinte página *web* no site do Laboratório de Inteligência Computacional (LABIC): <http://sites.labic.icmc.usp.br/rsinoara/tese/>. Nessa página estão disponíveis para a comunidade todos os *datasets* construídos, as representações geradas e avaliadas, as ferramentas e procedimentos desenvolvidos para a geração de tais representações, bem como todos os resultados das avaliações experimentais apresentadas nesta tese.

R6 Ferramentas para geração de representações da literatura. Além dos resultados obtidos diretamente com o desenvolvimento deste trabalho, também foram obtidos resultados a partir da colaboração com outros alunos de mestrado e iniciação científica. A partir dessas colaborações, foram desenvolvidas as seguintes ferramentas para a geração de representações da literatura, possibilitando a avaliação das representações propostas.

- Bag of Vectors² - ferramenta para geração de representações de coleções de documentos com base em modelos de *word embeddings*, como os modelos propostos por Mikolov *et al.* (2013b);
- PRED.ARG³ - ferramenta para geração de representações de documentos propostas por Persson, Johansson e Nugues (2009), que utilizam informações de papéis semânticos.
- Doc2Vec tools⁴ - ferramenta para treinamento de modelos Doc2Vec, proposto por Le e Mikolov (2014).

Esse resultado viabilizou a obtenção do resultado **R4**.

R7 Sistema para apoio à rotulação manual de textos. A partir da colaboração com outros alunos do grupo de pesquisa, foi desenvolvido o sistema RotuLABIC⁵, sistema para apoio à rotulação manual de textos utilizando Aprendizado de Máquina (PARAVIA *et al.*, 2015). Esse resultado viabilizou a obtenção do resultado **R3**.

1.5 Organização do texto

O restante desta tese está organizado como se segue.

Capítulo 2 - Fundamentos e definição do problema. Nesse capítulo são apresentados conceitos essenciais para o entendimento deste trabalho e o posicionamento do mesmo na área de pesquisa na qual ele está inserido. Assim, primeiramente é apresentada uma visão geral sobre o processo de Mineração de Textos, com especial atenção para a representação de documentos e a tarefa de classificação automática de textos. Aspectos sobre a semântica e tarefas de PLN que podem auxiliar na construção de representações de coleções textuais mais ricas também são apresentados brevemente nesse capítulo. Logo após, é apresentado o mapeamento da literatura de Mineração de Textos que foi desenvolvido neste trabalho. No Capítulo 2 também discute-se o impacto da semântica na Mineração de Textos. O problema tratado neste trabalho é definido, sendo apresentado o que foi denominado como

² Ferramenta Bag of Vectors: <<https://github.com/joao4ntunes/text-mining/>>

³ PRED.ARG: <<http://labic.icmc.usp.br/material/16>>

⁴ Doc2Vec tools: <<http://labic.icmc.usp.br/material/18>>

⁵ RotuLABIC: <<http://sites.labic.icmc.usp.br/rotulabic/>>

níveis de complexidade semântica, bem como os aspectos semânticos abordados nesta tese. Nesse contexto, também são apresentadas as coleções de textos e conjuntos de dados (*datasets*) utilizados nas avaliações experimentais realizadas, estabelecendo seus níveis de complexidade semântica.

Capítulo 3 - Explorações iniciais: representações baseadas em conhecimento do domínio e hiperônimos. Nesse capítulo são apresentados dois trabalhos iniciais realizados em busca da incorporação da semântica nas representações de coleções de documentos. São apresentadas e avaliadas duas propostas de modelos de representação semanticamente enriquecida: (i) *gBoED*, representação baseada em expressões do domínio; e (ii) *Uni-based*, representação baseada em desambiguação lexical de sentidos e representação vetorial de conceitos por meio de hiperônimos.

Capítulo 4 - Representações baseadas em papéis semânticos. A investigação da incorporação da semântica na representação de textos por meio de informação sobre papéis semânticos é apresentada nesse capítulo. Primeiramente são apresentados trabalhos relacionados ao uso de papéis semânticos na classificação automática de textos. Em seguida, são apresentados dois modelos propostos e desenvolvidos neste trabalho: *SR-based Terms* e *SR-based Sentences*. Tais modelos são analisados em relação a expressividade e limitações na construção de seus atributos, e avaliados na classificação automática de textos.

Capítulo 5 - Representações baseadas em *embeddings*. Nesse capítulo é apresentada a investigação da incorporação da semântica sob a perspectiva dos modelos preditivos de semântica distribucional (*embeddings*). São apresentados os trabalhos relacionados a representação da semântica de maneira latente, incluindo o uso de métodos de modelagem de tópicos e métodos desenvolvidos em torno de modelos preditivos de semântica distribucional. É apresentada a proposta de três modelos de representação de coleções textuais (*NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*), bem como os recursos e ferramentas utilizados na construção desses modelos propostos. No Capítulo 5 também são apresentadas análises desses modelos em relação ao espaço semântico em que as representações vetoriais dos documentos são projetadas e em relação ao potencial para identificação de similaridade textual. Por fim, é apresentada a avaliação experimental das representações na tarefa de classificação automática de textos.

Capítulo 6 - Conclusões. Nesse capítulo são apresentadas as considerações finais desta tese. As questões de pesquisa apresentadas nesta Introdução são retomadas, sendo sintetizadas as contribuições deste trabalho. Na sequência são apresentadas as publicações originadas deste trabalho, as limitações e direções para trabalhos futuros.

Apêndice A - Processo de rotulação da coleção *BEST sports - Top 4*. Nesse apêndice é apresentado o processo realizado para rotulação manual da coleção de documentos *BEST sports*

- *Top 4*. Essa rotulação possibilitou a construção dos *datasets* dessa coleção, apresentados no [Capítulo 2](#).

Apêndice B - Configuração experimental. Nesse apêndice são apresentados os detalhes da configuração experimental das avaliações apresentadas nos Capítulos [3](#), [4](#) e [5](#). São apresentados os procedimentos e ferramentas utilizados para a geração de todas as representações avaliadas, tanto as representações propostas quanto as representações da literatura. Também são apresentados os algoritmos de aprendizado utilizados, as estratégias para combinação de classificadores e os critérios de avaliação.

Apêndice C - Análises e resultados extras. Análises realizadas durante o desenvolvimento deste trabalho e resultados complementares são apresentados nesse apêndice.

FUNDAMENTOS E DEFINIÇÃO DO PROBLEMA

2.1 Considerações iniciais

Ao longo dos anos, diferentes áreas de pesquisa se desenvolveram em torno do processamento e da análise de textos. Segundo [Grobelnik \(2011\)](#), áreas de pesquisa como Mineração de Textos, Processamento de Língua Natural, Recuperação de Informação e Web Semântica lidam com textos de maneira independente, com técnicas próprias, e juntas podem contribuir para a solução do problema de interpretação de textos como um todo. [Grobelnik \(2011\)](#) também define três dimensões pelas quais os trabalhos de processamento de textos podem ser classificados: (i) representação: como os textos são representados; (ii) técnica: quais algoritmos e técnicas são utilizadas para manipular os textos; e (iii) tarefa: qual problema está sendo tratado.

Pode-se perceber que a análise de textos corresponde a uma vasta gama de áreas de pesquisa. Assim, torna-se necessário limitar o escopo desta pesquisa. Neste trabalho, o interesse está no problema de organização automática de coleções de documentos, por meio da aplicação do processo de MT e algoritmos de AM, utilizando representações no modelo espaço-vetorial. O principal foco deste trabalho é avaliar o impacto de aspectos semânticos no resultado da Mineração de Textos. Conforme já apresentado no Capítulo 1, nesta tese o termo semântica é usado em um sentido geral, considerando o sentido ou significado de itens linguísticos, sejam eles palavras, expressões ou documentos completos.

Nesse contexto, neste capítulo são apresentados alguns fundamentos sobre os quais essa tese foi desenvolvida, bem como uma revisão da literatura. Na Seção 2.2 é descrito o processo de MT, com foco principal na representação dos documentos e na tarefa de classificação automática de textos. Na Seção 2.3 é apresentada uma discussão sobre análise semântica, mostrando o impacto da semântica no entendimento de textos e as tarefas de análise semântica que podem

auxiliar na obtenção de representações semânticas dos textos. Logo após, na Seção 2.4, é apresentado um mapeamento da literatura sobre a semântica na Mineração de Textos, que foi realizado com o objetivo de conhecer e organizar os trabalhos desenvolvidos na área de MT e que possuem uma preocupação especial com aspectos semânticos envolvidos. Por fim, na Seção 2.5, é apresentada uma discussão sobre níveis de complexidade semântica, problema que é o foco deste trabalho. Como os níveis de complexidade semântica estão relacionados diretamente às coleções de textos e à natureza das classes, nessa seção também são detalhadas as coleções de documentos utilizadas nas avaliações experimentais e as abordagens adotadas, neste trabalho, em direção à representação da semântica dos textos.

2.2 Mineração de textos

Apesar de não haver um consenso entre diferentes comunidades de pesquisa (MINER *et al.*, 2012), a Mineração de Textos pode ser vista como a aplicação de um conjunto de técnicas usadas para analisar dados não estruturados e descobrir padrões que não eram conhecidos previamente (AGGARWAL; ZHAI, 2012a). Assim, a MT pode ser tratada como uma especialização da Mineração de Dados. Enquanto a Mineração de Dados trata os dados estruturados, a MT lida com textos escritos em língua natural (dados não estruturados). Com o crescente aumento e variedade de documentos textuais, tanto em redes sociais e *Web* em geral quanto internamente em organizações, as técnicas de MT têm se tornado essenciais no apoio à descoberta de conhecimento. Com isso, as fontes de textos, bem como as aplicações da MT, são variadas.

De forma geral, o processo de MT pode ser visto como um processo formado por cinco etapas, conforme ilustrado na Figura 2. Esse processo se inicia com a especificação de seus objetivos na etapa de Identificação do Problema. Nesta etapa, o analista, especialista em MT, deve delimitar o escopo da mineração, preferivelmente trabalhando com um especialista do domínio de aplicação. Devem ser definidas as coleções de textos que serão mineradas e como os resultados serão utilizados. As especificações definidas na etapa de Identificação do Problema guiarão as próximas etapas do processo de MT, as quais podem ser executadas em ciclos de preparação dos dados (etapa de Pré-processamento), descoberta de conhecimento (etapa de Extração de Padrões) e avaliação do conhecimento (etapa de Pós-processamento).

A etapa de Pré-processamento trata-se da preparação dos dados para a extração de padrões. Definido o escopo e os objetivos do processo, na etapa de Pré-processamento busca-se colocar os dados em um formato adequado para a extração de conhecimento, normalmente sendo realizadas atividades de tratamento, limpeza e redução do volume de dados disponível na base. É nessa etapa que os documentos são representados de maneira a torná-los processáveis pelos algoritmos usados para extração de padrões. As atividades realizadas na etapa de Pré-processamento são cruciais para o sucesso de todo o processo de MT. Os dados pré-processados devem preservar os padrões ocultos nos documentos para que os padrões de interesse possam

Figura 2 – Processo de Mineração de Textos

Fonte: Adaptada de [Sinoara, Antunes e Rezende \(2017\)](#).

ser descobertos na próxima etapa do processo. Desse fato vem a importância do modelo de representação de textos adotado. Na Seção 2.2.1 é apresentada uma discussão sobre como os dados textuais podem ser representados na Mineração de Textos, bem como iniciativas na direção da representação de aspectos semânticos.

Com a coleção de documentos devidamente formatada e tratada, pode-se iniciar a etapa de Extração de Padrões. As tarefas a serem realizadas são definidas de acordo com o objetivo final do processo de extração de conhecimento. Na etapa de Extração de Padrões, o analista aplica um algoritmo de aprendizado adequado para extrair os padrões dos dados pré-processados. A escolha do algoritmo é feita com base nos dados disponíveis e no tipo de conhecimento que se deseja descobrir. Quando o objetivo do processo de Mineração de Textos é organizar uma coleção de documentos, há duas tarefas básicas de MT que podem ser aplicadas: classificação e agrupamento. Neste trabalho, a tarefa de classificação de textos foi aplicada com o objetivo de avaliar os métodos de representação de coleções de documentos propostos. Assim, a tarefa de classificação é apresentada brevemente na Seção 2.2.2.

Uma vez obtidos os padrões acerca dos dados, esses devem ser avaliados e interpretados na etapa de Pós-processamento. Assim como as demais, essa etapa também deve ser guiada pelos objetivos definidos no início do processo. Pode-se avaliar diversos aspectos do conhecimento

extraído, como representatividade, novidade, validade e aplicabilidade. Essa avaliação pode ser realizada junto a um especialista do domínio ou por meio da aplicação de medidas objetivas de avaliação.

Após a etapa de Pós-processamento, caso o conhecimento extraído cumpra os objetivos estabelecidos para o processo de MT, o mesmo pode ser disponibilizado aos usuários, dando início à etapa final do processo (Utilização do Conhecimento). Caso contrário, outro ciclo deve ser executado, realizando mudanças nas atividades de preparação dos dados e/ou em parâmetros da extração de padrões. Se forem necessárias mudanças nos objetivos estabelecidos ou nas coleções de textos utilizadas, o processo de MT deve ser reiniciado na etapa de Identificação do Problema.

2.2.1 Pré-processamento e representação de documentos

A principal atividade realizada na etapa de Pré-processamento é a representação dos textos em um formato aceito pelo algoritmo a ser utilizado na Extração de Padrões. Os algoritmos tradicionais de Aprendizado de Máquina, que normalmente são utilizados na mineração de dados estruturados, assumem que os dados são apresentados em um formato conhecido como matriz atributo-valor. Nessa matriz, cada instância (ou exemplo) corresponde à uma linha e seus atributos (características que descrevem esse exemplo) correspondem às colunas.

A fim de representar coleções de textos como uma matriz atributo-valor, a coleção de documentos pode, da mesma maneira que os dados estruturados, ser representada em um modelo espaço-vetorial. A ideia desse modelo é representar cada documento de uma coleção como um ponto no espaço, ou seja, um vetor em um espaço vetorial, sendo que pontos que estão próximos nesse espaço representam documentos semanticamente similares (TURNERY; PANTEL, 2010). Nesse modelo, cada documento é representado por um vetor, cujas dimensões correspondem a atributos dos documentos. Normalmente, usa-se a palavra “termo” para se referir a dimensões que se baseiam em palavras dos textos (sejam palavras simples, sequências ou conjuntos de palavras) e a palavra “atributo” para dimensões de outras naturezas (ROSSI, 2016). Assim, as representações de textos tradicionais são chamadas de matriz documento-termo e são definidas como se segue. Considerando uma coleção de documentos $D = \{d_1, d_2, \dots, d_N\}$ contendo N documentos e um conjunto de M termos dessa coleção $T = \{t_1, t_2, \dots, t_M\}$, a coleção de documentos pode ser representada como uma matriz documento-termo formada pela união de N vetores compostos por M dimensões. Essa matriz é ilustrada na Figura 3.

Os valores da matriz documento-termo correspondem ao peso de cada termo em cada documento, ou seja, w_{d_i, t_j} corresponde ao peso do termo t_j no documento d_i . Normalmente esses pesos são calculados com base na frequência dos termos nos documentos. As medidas mais comuns são: (i) frequência do termo (TF - do inglês *term frequency*), que corresponde ao de número de vezes que o termo aparece no documento; (ii) frequência do termo ponderada pelo inverso da frequência de documento (TF-IDF - do inglês *term frequency - inverse document*

Figura 3 – Ilustração da representação de uma coleção de N documentos como uma matriz documento-termo

	t_1	t_2	$\cdots$	t_{M-1}	t_M
d_1	w_{d_1,t_1}	w_{d_1,t_2}	$\cdots$	$w_{d_1,t_{M-1}}$	w_{d_1,t_M}
d_2	w_{d_2,t_1}	w_{d_2,t_2}	$\cdots$	$w_{d_2,t_{M-1}}$	w_{d_2,t_M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
d_{N-1}	w_{d_{N-1},t_1}	w_{d_{N-1},t_2}	$\cdots$	$w_{d_{N-1},t_{M-1}}$	w_{d_{N-1},t_M}
d_N	w_{d_N,t_1}	w_{d_N,t_2}	$\cdots$	$w_{d_N,t_{M-1}}$	w_{d_N,t_M}

Fonte: Elaborada pelo autor.

frequency), que corresponde à frequência do termo ponderada pelo inverso da frequência de documento (número de documentos nos quais o termo aparece); e (iii) binária, que corresponde à presença (valor 1) ou ausência (valor 0) do termo no documento. De acordo com Rossi (2016), outros métodos foram propostos na literatura para cálculo desses pesos visando aumentar a performance de classificação, porém com contribuições limitadas à tarefas específicas.

Assim como o cálculo dos pesos, também existem diferentes possibilidades para a definição dos termos. Na representação de textos mais tradicional, conhecida como *bag-of-words*, cada termo corresponde a uma palavra presente na coleção de documentos. Essa representação assume que os termos são independentes e a ordem com que eles ocorrem nos documentos não é considerada. Uma alternativa para manter a relação de ordem entre as palavras é o uso de sequências de palavras (n-gramas). Nessa representação, cada termo é formado por uma sequência de n palavras que ocorrem em ao menos um documento da coleção. No entanto, o uso de n-gramas aumenta o número de termos a serem considerados. Quanto maior o valor de n , maior é a dimensionalidade da matriz. Além disso, termos formados por mais de uma palavra são mais específicos do que termos formados por apenas uma palavra, o que pode levar a uma diminuição da frequência dos termos na coleção de documentos. Por exemplo, o termo “text mining” será menos frequente que os termos independentes “text” e “mining”, visto que estes podem ocorrer em outros contextos. Esses fatos agravam os problemas da alta dimensionalidade e alta esparsidade que já ocorrem na representação BOW. Essa representação possui essas características porque uma coleção de documentos pode conter milhares de palavras e um de seus documentos pode conter apenas um pequeno conjunto dessas palavras. Essas características fazem com que muitos dos algoritmos de extração de padrões sejam ineficientes ao lidar com este tipo de representação.

Visando reduzir o número de termos e amenizar os problemas da alta dimensionalidade e esparsidade, pode-se utilizar algumas técnicas de pré-processamento, como as listadas a seguir.

Remoção de *stopwords*. A remoção de *stopwords* visa a eliminação de palavras que não trazem informação relevante para o processo de MT. Essas palavras, chamadas de *stopwords*, normalmente são palavras que possuem as funções de artigos, preposições, pronomes e

conjunções. No entanto, também podem ser identificadas *stopwords* específicas do domínio de aplicação do processo, ou seja, palavras que sabidamente são frequentes na coleção e que não distinguem classes ou grupos que espera-se identificar com a Mineração de Textos.

Normalização A normalização dos textos visa eliminar as diversas variações que as palavras podem sofrer, como por exemplo variações de gênero e número dos substantivos e conjugações dos verbos. A normalização pode ser realizada por meio de três técnicas: (i) radicalização (*stemming*), que reduz cada palavra ao seu radical (ou palavra raiz); (ii) lematização, que reduz cada palavra a seu lema (ou forma canônica), ou seja, verbos são reduzidos ao infinitivo, e substantivos e adjetivos são reduzidos a forma masculina singular; e (iii) substantivação, que transforma a palavra para que ela tenha o comportamento sintático/semântico semelhante a de um substantivo. Uma comparação entre as três técnicas de normalização na classificação de documentos é apresentada por [Conrado, Gutiérrez e Rezende \(2012\)](#).

Seleção de atributos A seleção de atributos visa selecionar os atributos mais relevantes da coleção de documentos, tornando o conjunto de atributos mais conciso, porém não menos representativo em relação ao conjunto original. Busca-se identificar os atributos que são importantes para o problema, selecionando aqueles que tem maior influência na definição da classe e/ou eliminando os redundantes ou com alto grau de ruído. A seleção de atributos pode ser feita por meio da aplicação de uma medida de avaliação sobre os atributos e posterior seleção daqueles que possuem os melhores valores para essa medida, baseada em um *ranking* ou em um corte por limiar. Esse processo pode ser realizado tanto de maneira supervisionada, quando se possui dados rotulados, quanto de maneira não-supervisionada ([GUYON; ELISSEEFF, 2003](#); [NOGUEIRA; REZENDE, 2009](#)).

Essas técnicas de pré-processamento reduzem a dimensionalidade mantendo termos como atributos da representação. Outras técnicas também podem ser utilizadas para reduzir a dimensionalidade da representação, porém gerando novos atributos, que não mantêm uma relação explícita com os termos da representação inicial. [Rossi \(2016\)](#) classifica essas técnicas em duas categorias: extração de atributos e extração de tópicos. Entre as principais técnicas de extração de atributos encontra-se o *Latent Semantic Indexing* e entre as técnicas de extração de tópicos encontra-se o *Latent Dirichlet Allocation* ([AGGARWAL; ZHAI, 2012a](#); [KUHN; DUCASSE; GÍRBA, 2007](#); [BLEI; NG; JORDAN, 2003](#)). Com o uso dessas técnicas, formas alternativas de expressar o mesmo conteúdo são reduzidas a uma representação comum. Com isso, além da redução da dimensionalidade, também é feito um trabalho na direção de se considerar a semântica dos textos. Porém, apesar de remover ruídos causados por sinônimos e termos polissêmicos, os textos ainda são tratados como um conjunto de palavras independentes e desordenadas. Portanto, relacionamentos e ideias contidos nos textos ainda não são representados.

Considerando a representação semântica de itens linguísticos, *word embeddings* tem ganhado muita atenção em pesquisas em semântica distribucional, com a geração de representações distribuídas a partir de modelos preditivos baseados em redes neurais. Um destaque nas pesquisas de *word embeddings* foi a proposta dos modelos de aprendizado de vetores *Continuous Bag-of-Words* e *Skip-gram* (MIKOLOV *et al.*, 2013a). Nessa linha, para a representação de itens linguísticos maiores, como sentenças e documentos, Le e Mikolov (2014) propuseram uma abordagem inspirada em *word embeddings*. Tal abordagem, inicialmente chamada de *Paragraph Vector* e posteriormente de *Doc2Vec*, consiste em treinar uma rede neural de uma camada intermediária para predizer palavras em documentos. Após o treinamento, os pesos da camada intermediária podem ser usados como representação vetorial do documento. Experimentos realizados por Le e Mikolov (2014) reportam resultados competitivos com métodos estado da arte em tarefas de análise de sentimentos e recuperação da informação. Outros autores também reportam bons resultados com o uso do modelo *Doc2Vec*, em tarefas como similaridade semântica textual, identificação de perfil de autor (*author profiling*), análise de sentimentos e classificação (LEE; YOON, 2017; LAU; BALDWIN, 2016; MARKOV *et al.*, 2016; LIANG; FOTHERGILL; BALDWIN, 2015).

Com o objetivo de incorporar a semântica por meio da identificação dos conceitos presentes nos documentos textuais, alguns métodos utilizam fontes externas de informação, como ontologias, dicionários ou tesouros. O método *Explicit Semantic Analysis*, por exemplo, utiliza a *Wikipedia*¹ para obter um vetor de conceitos para representar os documentos (GABRILOVICH; MARKOVITCH, 2007). Seguindo essa mesma linha, Spanakis, Siolas e Stafylopatis (2012) melhoraram a qualidade de agrupamentos hierárquicos com o uso de uma representação de textos que utiliza tanto os conceitos da *Wikipedia* quanto outros atributos, como links e categorias, disponíveis nos artigos da *Wikipedia*. A identificação dos conceitos presentes em cada documento se dá pela anotação morfossintática (*POS tagging*) e identificação dos sintagmas nominais (*noun phrases*), que formam os conceitos candidatos. Os conceitos candidatos que correspondem a um artigo na *Wikipedia* são promovidos a conceitos. Caso mais de um artigo seja encontrado para um mesmo conceito candidato é feita a desambiguação do mesmo. Os conceitos identificados no documento, juntamente com atributos calculados com base em informações do respectivo artigo da *Wikipedia*, formam o vetor que representa o documento. Abordagens como essa são utilizadas para enriquecer a representação dos documentos no modelo espaço-vetorial.

Para se representar relações entre documentos textuais ou entre as entidades que compõem os documentos, como autoria e citações, a representação por redes (ou grafos) pode ser utilizada como uma alternativa às representações baseadas no modelo espaço-vetorial. As redes são compostas por objetos, que representam as entidades de um problema, e as relações entre esses objetos. As relações entre objetos permitem a extração de padrões que são dificilmente capturados por algoritmos baseados no modelo espaço-vetorial, além de serem úteis para melho-

¹ Wikipedia: <<http://www.wikipedia.com.br>>

rar a qualidade dos padrões extraídos (LU; GETOOR, 2003). Diversos trabalhos relacionados à extração de padrões de dados modelados em rede têm sido desenvolvidos (SUN; HAN, 2012; ROSSI *et al.*, 2012; HAN, 2012). Uma coleção de documentos pode ser representada por meio de uma rede heterogênea bipartida, na qual existem objetos do tipo documento e do tipo termos. Nesse tipo de rede, vértices do tipo documento são ligados a vértices do tipo termo quando o termo existe no documento. O peso dessas conexões pode ser baseado na frequência do termo no documento. Rossi *et al.* (2012) obtiveram bons resultados utilizando esse tipo de rede em problemas de classificação.

Outras configurações também podem ser utilizadas para gerar grafos que representam o conteúdo de documentos. Solé *et al.* (2005) discutem a modelagem de línguas naturais em redes e apresentam três tipos de redes de linguagem que são formadas considerando diferentes aspectos da linguagem: redes de coocorrência (palavras ligadas por ocorrerem juntas em uma sentença), redes sintáticas (palavras ligadas com base em dependências sintáticas) e redes semânticas (palavras são ligadas por relações semânticas). Representações baseadas nesses tipos de redes têm sido utilizadas em tarefas de classificação e agrupamento de textos (MISHRA *et al.*, 2012; WU; XUAN; PAN, 2011; WU *et al.*, 2010; LIU; WANG; WANG, 2008; MONTES-Y-GOMEZ *et al.*, 2001). Também são encontradas representações de textos em grafos nas quais os vértices são sujeitos e objetos das sentenças e as arestas são os verbos (GARRIDO *et al.*, 2013; CARVALHO; FREITAS; SILVA, 2013).

Apesar de representar informações sobre relacionamentos, assim como as tradicionais representações no modelo espaço-vetorial, as redes ainda são limitadas em relação à semântica dos textos. Baseando-se em representações que não são capazes de manter informações semânticas de interesse, os algoritmos de extração de padrões também perdem informações valiosas para a qualidade do conhecimento resultante. Para determinadas aplicações, a falta de informações semânticas pode ter um impacto determinante no resultado final (STAVRIANOU; ANDRITSOS; NICOLYANNIS, 2007). Assim, embora diversas pesquisas tenham sido desenvolvidas para o desenvolvimento de novos modelos de representação, tanto em redes quanto no modelo espaço-vetorial, ainda há espaço para desenvolvimento de métodos e estruturas mais elaborados para a incorporação da semântica na Mineração de Textos (AGGARWAL; ZHAI, 2012a; TURNEY; PANTEL, 2010; STAVRIANOU; ANDRITSOS; NICOLYANNIS, 2007).

2.2.2 Classificação automática de textos

Nessa seção são apresentados conceitos fundamentais da classificação automática de documentos por meio da aplicação de algoritmos de Aprendizado de Máquina. Tais fundamentos são apresentados visto que os métodos para representação de coleções de documentos apresentados nesta tese foram avaliados na tarefa de classificação de textos.

Considerando-se o uso de algoritmos de aprendizado indutivo supervisionados, o problema de classificação automática de textos é definido como se segue. Dados um conjunto de

classes (C) e uma coleção de documentos rotulados (D), documentos cuja classe é conhecida, um algoritmo indutivo supervisionado induz uma função F que mapeia os documentos de D a classes de C ($F : D \mapsto C$). A função F é chamada de modelo de classificação (ou classificador) e é utilizada para prever a classe de novos documentos. Esse processo é ilustrado na Figura 4.

Figura 4 – Esquema ilustrativo da classificação automática de textos por meio de aprendizado indutivo supervisionado

Fonte: Rossi (2016).

Com a coleção de documentos representada no modelo espaço-vetorial, diversos algoritmos de aprendizado indutivo supervisionado podem ser aplicados para induzir um classificador. Os seguintes algoritmos estão entre os mais utilizados e recomendados para a classificação de textos (ROSSI, 2016; AGGARWAL; ZHAI, 2012a; SEBASTIANI, 2002): (i) Naïve Bayes e variantes (método probabilístico), (ii) C4.5 (método baseado em procura, árvore de decisão), (iii) *Support Vector Machine* (método baseado em otimização), e (iv) *k-Nearest Neighbors* (método baseado em distâncias). Ainda considerando as representações no modelo espaço-vetorial, além dos algoritmos tradicionais que utilizam como entrada uma matriz, também é possível aplicar algoritmos de aprendizado baseados em redes bipartidas, visto que essas redes podem ser diretamente derivadas de representações no modelo espaço-vetorial. Esse é o caso dos algoritmos IMBHN (*Inductive Model based on Bipartite Heterogeneous Networks*): $IMBHN^C$ (ROSSI *et al.*, 2014) e $IMBHN^R$ (ROSSI; LOPES; REZENDE, 2016). Por apresentarem uma boa performance de classificação, esses algoritmos foram considerados junto com os tradicionais algoritmos de aprendizado indutivo supervisionado nas avaliações realizadas neste trabalho.

A classificação automática de textos pode ser aplicada em diversos problemas, como organização e recuperação de notícias, artigos e outros tipos de documentos, filtragem de *e-mails* e detecção de *spam*, desambiguação lexical de sentidos, análise de sentimentos e mineração de opinião (ROSSI, 2016; AGGARWAL; ZHAI, 2012a; SEBASTIANI, 2002). Dada a grande

aplicabilidade da tarefa de classificação automática de textos, diversas pesquisas têm sido desenvolvidas. Tais pesquisas buscam melhorar cada vez mais a qualidade dos classificadores e aproveitar as diferentes particularidades de cada problema. Uma das diversas abordagens alternativas para se realizar classificação é a combinação de classificadores.

As decisões de classificadores individuais (chamados de classificadores base) podem ser combinadas ao se classificar novos exemplos, formando o que é chamado de *ensemble* de classificadores. Quando os classificadores individuais são precisos e diversos, o *ensemble* de classificadores será mais preciso do que qualquer um dos membros individuais (DIETTERICH, 2000). De modo geral, os classificadores base de um *ensemble* podem ser construídos de três maneiras: (i) alterando o conjunto de treinamento, na qual cada classificador é gerado por uma amostra distinta dos exemplos de treinamento; (ii) alterando os atributos de entrada, na qual cada classificador é gerado utilizando um conjunto distinto de atributos que formam os exemplos; e (iii) alterando os algoritmos de aprendizado empregado e/ou seus parâmetros, na qual cada classificador é gerado utilizando um algoritmo de aprendizado diferente ou variações dos parâmetros de um mesmo algoritmo de aprendizado. O foco deste trabalho está em representações alternativas para coleções de documentos, portanto trabalha-se com a disponibilidade de diferentes representações dos mesmos dados. Neste contexto, a combinação de classificadores considerando-se a variação dos atributos de entrada (variação dos métodos de representação empregados) foi explorada nas avaliações realizadas neste trabalho.

Para a construção de *ensembles* de classificadores, também deve ser definido como é feita a composição dos votos entre os diferentes classificadores base, ou seja, como os votos dos classificadores serão combinados. Cevikalp e Polikar (2008) classifica as diferentes abordagens que podem ser utilizadas para essa combinação em duas categorias: seleção de classificadores e fusão de classificadores. Abordagens de seleção de classificadores usam apenas um ou poucos classificadores que seriam avaliados como o(s) mais confiável(is). Neste caso, pode-se, por exemplo, atribuir ao novo exemplo a classe que foi predita pelo classificador com maior confiança. Abordagens de fusão de classificadores usam alguma forma de voto ponderado entre a saída de vários classificadores. Um exemplo de fusão de classificadores é realizar a soma das confianças dos diferentes classificadores para cada classe e atribuir ao exemplo não rotulado a classe que tiver a maior confiança final.

Na classificação automática de textos, a avaliação experimental de classificadores é normalmente realizada medindo-se a performance de classificação (SEBASTIANI, 2002). Diversas medidas de performance podem ser calculadas com base nos valores de uma matriz de confusão, que apresenta o número de exemplos classificados corretamente e incorretamente por um determinado classificador. Na Figura 5 é ilustrada uma matriz considerando uma classe $c_i \in C$.

A Acurácia de um classificador é dada pela porcentagem de documentos corretamente

Figura 5 – Matriz de confusão para a classe c_i

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

Fonte: Elaborada pelo autor.

classificados, conforme a [Equação 2.1](#).

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

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

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

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

Para problemas multiclasse, nos quais existem duas ou mais classes e cada documento é rotulado com apenas uma classe, as medidas Precisão e Revocação de cada classe individualmente podem ser sumarizadas considerando-se duas abordagens ([SEBASTIANI, 2002](#); [ROSSI, 2016](#)): *micro-averaging* (Equações 2.4 e 2.5) e *macro-averaging* (Equações 2.6 e 2.7). As duas abordagens dão ênfases diferentes para a distribuição dos exemplos nas diferentes classes. Como pode ser observado nas Equações 2.4 e 2.5, por realizar a soma dos termos individualmente para cada classe, a abordagem *micro-averaging* corresponde à medida Acurácia ([Equação 2.1](#)). Portanto, as medidas Precisão^μ, Revocação^μ e Acurácia resultam em um mesmo valor. Já na abordagem *macro-averaging* (Equações 2.6 e 2.7), são dados pesos iguais às classes, independente do número de exemplos. Com isso, essa abordagem tende a enfatizar mais a classificação correta de exemplos de classes minoritárias, que possuem uma quantidade muito pequena de exemplos em relação a outras classes, do que a abordagem *micro-averaging*.

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

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

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

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

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

$$F1 = 2 * \frac{Precisão * Revocação}{Precisão + Revocação} \quad (2.8)$$

Nessa seção foi apresentada uma breve revisão sobre a classificação automática de textos por meio de algoritmos de AM, visto que as representações de documentos propostas foram avaliadas na classificação de textos. Informações mais detalhadas sobre esse extenso assunto podem ser encontradas em diversos livros e artigos de revisão da literatura, como os trabalhos de [Aggarwal \(2014\)](#), [Faceli et al. \(2011\)](#), [Sebastiani \(2002\)](#), [Mitchell \(1997\)](#).

2.3 Análise semântica

Na Linguística, Semântica é o ramo que compreende o estudo do significado das palavras ([OLIVEIRA, 2012](#)). A Linguística apresenta diversas frentes de estudo da semântica, vertentes que dão um enfoque diferente no estudo do significado. O estudo da semântica vai desde a relação entre as palavras e seres/coisas do mundo real até as mudanças de sentido que as palavras sofrem com o tempo. Assim, quando pensa-se em semântica pensa-se em sentido, em significado de algo.

[Riemer \(2010\)](#) apresenta uma introdução sobre Semântica, mostrando as diferentes abordagens para o estudo do significado. Tal tema é bastante abrangente e não há consenso entre os pesquisadores da Linguística sobre os limites da Semântica. Um exemplo é a relação entre Semântica e Pragmática. A Semântica pode ser vista como responsável pelo estudo do significado de sentença (*sentence meaning*), aquele formado pela composição do significado das palavras que estão presentes em uma frase ou expressão. Já a Pragmática pode ser vista como responsável pelo estudo do significado do enunciado (*utterance meaning*, também chamado de significado do falante, *speaker meaning*), que é o significado que as palavras apresentam em determinado uso e contexto (tempo, lugar, falante), podendo ser afetado por ironia, metáfora e outras figuras de linguagem. No entanto, não há consenso sobre essa relação entre Semântica e Pragmática ([FOSSILE, 2009](#); [ARAÚJO, 2007](#)).

Assim, essa seção não visa cobrir todo o tema, e sim apresentar alguns conceitos básicos relacionados à semântica e contextualizar o trabalho apresentado nesta tese. O interesse deste trabalho está na utilização do significado que documentos escritos em língua natural têm para o usuário, dado o contexto e objetivo no momento da análise. Portanto, nesta seção são apresentadas

algumas relações semânticas que têm impacto no significado dos textos e algumas tarefas de Processamento de Língua Natural que lidam com a semântica.

As palavras apresentam diversas relações entre si. Conforme apresentado por [Pietroforte \(2010\)](#), entre essas relações estão:

Sinonímia - Termos sinônimos podem se substituir em determinado contexto. Por exemplo: as palavras “novo” e “jovem” são sinônimas quando se trata da característica de um ser vivo, como em “homem novo”/“homem jovem”, porém “jovem” não pode substituir “novo” quando se trata da característica de um objeto, como em “livro novo” ([PIETROFORTE, 2010](#)).

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

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

Hiponímia - Um termo é hipônimo de outro termo que é seu hiperônimo. Exemplo: “futebol” é hipônimo de “esporte”.

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

Holonímia - Um termo é holônimo de outro termo quando existe uma relação parte - todo entre eles. Holônimo é o termo que corresponde ao todo na relação parte-todo. Exemplo: “carro” (todo) é holônimo de “freio” (parte).

Meronímia - Um termo é merônimo de outro termo que é o seu holônimo. Por exemplo: “freio” é parte de “carro”, portanto “freio” é merônimo de “carro”.

Polissemia - Um termo é polissêmico quando possui mais de um significado. Exemplo: o termo “manga” pode tanto se referir à uma fruta quanto à uma parte de uma camisa.

Além do significado das palavras, o significado de uma frase (ou sentença) também depende da sua estrutura gramatical. Por exemplo, as frases “João matou o bandido” e “O bandido matou João” possuem as mesmas palavras, porém, dada a estrutura sintática diferente

(alternância entre sujeito e objeto), elas apresentam significados distintos (MÜLLER; VIOTTI, 2010). Além disso, assim como as palavras, as sentenças também possuem relações entre si:

Paráfrase - Corresponde à noção de sinonímia estendida para sentenças.

Acarretamento - Corresponde à noção de hiponímia estendida para sentenças.

Contradição - Duas sentenças são contraditórias quando elas não podem ser simultaneamente verdadeiras.

Ambiguidade - Uma sentença é ambígua quando ela pode ter mais de um sentido. A ambiguidade de uma sentença pode ser causada por uma palavra ambígua, por diferentes estruturas sintáticas possíveis, ou por uma ambiguidade semântica (causada por relações anafóricas, dêiticas ou de escopo, relações descritas na sequência).

Relação Anafórica - Ocorre quando um pronome presente na sentença se refere a um nome citado anteriormente na mesma sentença.

Relação Dêitica - Ocorre quando um pronome presente na sentença se refere a um ente que existe no contexto.

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

Essas relações semânticas entre palavras e sentenças influenciam como as pessoas interpretam os textos e podem ser importantes para a Mineração de Textos. O entendimento de textos escritos em língua natural é um processo complexo, que se dá por meio do conhecimento das palavras e de seus significados, das relações existentes entre as palavras, bem como do conhecimento de mundo e do contexto no qual o texto foi escrito. Por exemplo, as frases “João matou o bandido” e “O bandido matou João” são representadas igualmente no modelo *bag-of-words* (apresentado na Seção 2.2.1, página 40), pois possuem as mesmas palavras na mesma frequência. Assim, a representação *bag-of-words* não é suficiente para distinguir os sentidos dessas duas sentenças.

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

Reconhecimento de Entidades Nomeadas - Tarefa de extração de informação que envolve processar um texto e identificar as ocorrências de palavras ou expressões pertencentes a categorias de entidades nomeadas (AMARAL; VIEIRA, 2014; GRISHMAN; SUNDHEIM, 1996). São exemplos de categorias de entidades nomeadas: Pessoa, Organização e Local.

Além das entidades identificadas por nome próprio, também é comum o reconhecimento de expressões temporais e numéricas.

Anotação de Papéis Semânticos - Tarefa que busca identificar o predicado de uma oração e atribuir papéis semânticos a seus argumentos (FONSECA; ROSA, 2012; PALMER; GILDEA; XUE, 2010). Com os papéis semânticos obtém-se informações do tipo “quem fez o que para quem”, além de “como fez” e “quando fez”. Como exemplos de papéis semânticos tem-se Agente (aquele que inicia a ação), Paciente (aquele afetado pela ação), Instrumento (algo ou meio utilizado para efetuar a ação) e Local (lugar de um objeto ou ação).

Desambiguação Lexical de Sentidos - Tarefa que busca determinar qual sentido uma palavra apresenta quando é utilizada em determinado contexto (NÓBREGA; PARDO, 2014; MORO; RAGANATO; NAVIGLI, 2014; AGIRRE; EDMONDS, 2007). Normalmente essa tarefa é realizada com o apoio de recursos léxicos, como a WordNet², que agrupa as palavras em conjuntos de sinônimos e apresenta relacionamentos entre esses conjuntos e seus membros.

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

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

Similaridade Semântica - Tarefa que busca determinar o grau de equivalência semântica entre um par de itens linguísticos, que podem ser palavras, conceitos ou sentenças, por exemplo. Diversas medidas e abordagens têm sido propostas para medir similaridade semântica e corpus têm sido construídos para possibilitar a avaliação das propostas (HARISPE *et al.*, 2015; PILEHVAR; NAVIGLI, 2015; FONSECA *et al.*, 2016; AGIRRE *et al.*, 2015; LEE; PINCOMBE; WELSH, 2005).

As relações semânticas apresentadas e o tratamento delas por meio das tarefas da área de PLN podem contribuir para a obtenção de melhores resultados na Mineração de Textos. Na próxima seção é apresentado um mapeamento sistemático sobre a semântica na Mineração de Textos, realizado com o objetivo de mapear os diversos trabalhos desenvolvidos em MT que consideram a semântica dos textos de alguma maneira.

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

2.4 Mapeamento da semântica na Mineração de Textos

Dados textuais não estão naturalmente em um formato adequado para a extração de padrões, o que traz desafios adicionais ao processo automático de descoberta de conhecimento. O significado dos textos em língua natural dependem, basicamente, dos três níveis de conhecimento linguístico: léxico, sintático e semântico. Cada nível é mais complexo e requer um processamento mais rebuscado do que o nível anterior. Expressividade versus custo de processamento é um *trade-off* comum no processamento de língua natural. Assim, os componentes léxico e sintático têm sido mais amplamente explorados na Mineração de Textos (AGGARWAL; ZHAI, 2012a). No entanto, houve um crescimento no interesse do tratamento semântico dos textos entre os pesquisadores de MT, com o objetivo de melhorar os resultados da mineração. Esse crescimento pode ser atribuído ao progresso da capacidade de processamento computacional, com consequente redução do tempo de processamento, e à evolução da área de PLN, que permite um processamento mais profundo dos textos puros.

A semântica dos textos pode ser considerada nas três principais etapas do processo de MT: Pré-processamento, Extração de Padrões e Pós-processamento. Na etapa de Pré-processamento, os dados podem ser representados com base em algum aspecto semântico, além de apenas o léxico. Na Extração de Padrões, informações semânticas podem guiar a geração do modelo, permitindo uma seleção mais apurada das hipóteses consideradas. Já na etapa de Pós-processamento, os padrões extraídos podem ser avaliados com base em aspectos semânticos. Em todos os casos, a incorporação da semântica na MT possibilita um avanço nos resultados obtidos com a mineração tradicional. Um tratamento adequado da semântica dos textos pode levar a resultados mais apropriados em certas aplicações (AGGARWAL; ZHAI, 2012a). A semântica tem um papel importante no significado dos textos e pode, por exemplo, ser crucial para diferenciar documentos que, apesar de usarem o mesmo vocabulário, apresentam ideias diferentes sobre um mesmo assunto, como é o caso do exemplo apresentado na Introdução desta tese (Figura 1, Página 27).

O termo “*semantics*” tem sido usado em vários trabalhos de Mineração de Textos. Assim, foi realizado um mapeamento sistemático da literatura neste tema³. Tal revisão da literatura foi realizada com o objetivo de mapear os estudos de Mineração de Textos que consideram a semântica, identificando e organizando as diferentes frentes de pesquisa realizadas no sentido da incorporação da semântica no processo de MT. Os estudos considerados neste mapeamento foram analisados principalmente em relação à domínio de aplicação, tarefas de Mineração de Textos, métodos e recursos aplicados e participação do usuário no processo.

Nessa seção é apresentado brevemente o protocolo utilizado para realizar esse estudo e um resumo dos principais resultados obtidos. Detalhes do processo de revisão adotado e dos

³ “Semântica” pode estar relacionada a diferentes temas de pesquisa, sendo a maioria deles estudados na área de Processamento de Língua Natural, como as tarefas apresentadas na Seção 2.3. No entanto, o foco do mapeamento realizado não está na semântica em si, mas sim em trabalhos de Mineração de Textos em que a semântica é considerada de alguma maneira.

resultados obtidos podem ser vistos no artigo de [Sinoara, Antunes e Rezende \(2017\)](#).

2.4.1 Protocolo do mapeamento

Mapeamento sistemático corresponde ao levantamento sistemático da literatura de determinada área de pesquisa, realizado com o objetivo de prover uma visão geral da área e uma indicação da quantidade de estudos sendo realizados ([KITCHENHAM; CHARTERS, 2007](#)). Este segue um protocolo semelhante ao de uma revisão sistemática convencional, porém as principais diferenças entre os dois métodos são a quantidade de estudos considerados e a profundidade da análise desses estudos ([PETERSEN et al., 2008](#)). Um mapeamento sistemático, em geral, possui termos de pesquisa mais abrangentes e, portanto, um número maior de estudos são identificados. Com isso, a análise feita sobre cada estudo não é tão profunda quanto à análise feita em uma revisão sistemática tradicional e normalmente seus resultados são apresentados de forma gráfica.

O mapeamento apresentado nesta seção pode ser visto como um mapeamento *quasi-sistemático* por ter duas diferenças em relação à definição clássica de revisões sistemáticas da literatura ([KITCHENHAM; CHARTERS, 2007](#)). A primeira refere-se à participação dos pesquisadores no processo de revisão. Apesar do planejamento do mapeamento ter sido realizado por dois pesquisadores, a seleção dos estudos e a extração de informações foram conduzidas por apenas um pesquisador. A segunda diferença está nos tipos de estudos que foram considerados. Como esse mapeamento foi realizado com o objetivo de se conhecer toda a produção científica da área, além dos estudos primários também foram considerados os estudos secundários. Os principais pontos do protocolo utilizado nesse mapeamento são apresentados a seguir.

Questões de Pesquisa: o trabalho foi guiado por uma questão principal e outras secundárias, que especificam a principal.

Questão principal: Como a semântica é tratada no processo de Mineração de Textos?

Questões secundárias:

- Quais são os domínios de aplicação que consideram a semântica dos textos de alguma maneira?
- Quais idiomas estão sendo tratados?
- Quais fontes externas são comumente utilizadas?
- Em quais tarefas a semântica aparece com mais frequência?
- Quais métodos e algoritmos são comumente utilizados?
- Como os textos podem ser representados?
- Os usuários participam do processo de Mineração de Textos?

Fontes e Termos de Pesquisa: as pesquisas foram realizadas em cinco importantes bibliotecas digitais e fontes de pesquisa da área de computação: ACM Digital Library⁴, IEEE Xplore⁵, Science Direct⁶, Web of Science⁷ e Scopus⁸. Sempre que possível os termos de busca foram procurados tanto no título quanto nas palavras-chaves dos artigos. Foi utilizada a seguinte expressão de busca geral: *semantic* AND text* AND (mining OR representation OR clustering OR classification OR association rules)*. A *string* de busca foi adaptada para cada fonte de pesquisa.

Estratégia de Seleção: todos os estudos retornados pelas buscas passaram pela etapa de seleção. Os estudos foram selecionados com base no título, resumo e outras informações do artigo (como número de páginas, por exemplo). Foram marcados como duplicados os estudos retornados por mais de uma fonte de pesquisa, deixando apenas uma entrada por artigo. Nessa etapa foram rejeitados os artigos que se encaixaram em pelo menos um dos critérios de exclusão listados a seguir.

- Publicação com uma página, pôsteres, apresentações, resumos e índices de anais de eventos e editoriais.
- Publicação hospedada em serviços inacessíveis.
- Publicação em idioma diferente de inglês ou de português.
- Publicação que não se refere a tratamento/manipulação de textos ou que não se referem à semântica dos textos.

Estratégia de Extração de Informações: a extração de informações foi realizada com a leitura do resumo dos artigos aceitos na etapa de seleção. As informações extraídas referem-se principalmente a sete importantes dimensões da Mineração de Textos: 1. domínio de aplicação, 2. idioma, 3. fonte de conhecimento externo, 4. tarefa de Mineração de Textos, 5. métodos e algoritmos, 6. modelo de representação, e 7. interação com usuário. As informações que buscou-se extrair de cada estudo são apresentadas na [Figura 6](#).

2.4.2 Resultados do mapeamento

Esse mapeamento foi realizado com base em 1.693 estudos, selecionados entre 3.984 estudos identificados nas cinco fontes de pesquisa. Foram executados dois ciclos das fases de mapeamento sistemático. O primeiro ciclo foi executado com buscas realizadas de janeiro de 2014 e o segundo, correspondente a uma atualização do mapeamento, foi executado com buscas realizadas em fevereiro de 2016. Os resultados da análise quantitativa da produção científica,

⁴ ACM Digital Library: <<http://portal.acm.org/>>

⁵ IEEE Xplore: <<http://ieeexplore.ieee.org/>>

⁶ Science Direct: <<http://www.sciencedirect.com/>>

⁷ Web of Science: <<http://isiknowledge.com/>>

⁸ Scopus: <<http://www.scopus.com/>>

Figura 6 – Formulário para extração de informações durante a condução do mapeamento sistemático da literatura.

Questions - Information Extraction

- Paper and information extraction data:
 1. Date
 2. Title, authors, publication year, journal or conference
- Study data:
 3. Study type (primary or secondary)
 4. Is any application domain presented? Which one?
 5. Does the study deal with an specific language? Which one?
 6. Is any external knowledge source used? Which one?
 7. Which text mining task is addressed?
 8. Which method or algorithm is applied or proposed?
 9. How do texts and their semantics are normally represented?
 10. Does a user or domain expert take part in the process? How?

Fonte: Sinoara, Antunes e Rezende (2017).

apresentada na Figura 7, confirmou alguns indicativos prévios dos pesquisadores sobre o tema de pesquisa e apontou outras características interessantes da área.

Em relação ao domínio de aplicação (Figura 7a), correspondendo a área de pesquisa ou gênero textual, o mapeamento mostrou que os domínios mais preocupados com a semântica ao se realizar alguma tarefa de MT são os domínios de saúde e ciências biológicas (33,9%) e documentos da *Web*, incluindo páginas *web*, *blogs*, fóruns, redes sociais e filtragem de *e-mails* (22,7%). As áreas de saúde e ciências biológicas são tradicionalmente preocupadas com padronização e classificação de seus conceitos e relacionamentos. Isso normalmente é feito por meio da criação de taxonomias ou ontologias, que são inerentemente ligadas à semântica dos textos. Já a presença de documentos da *Web* pode ser explicada pelo constante crescimento, tanto em quantidade como em variedade, desse tipo de conteúdo.

Em relação aos idiomas tratados (Figura 7b), encontrou-se grande presença do idioma chinês. Depois do idioma inglês, cuja presença é dominante em pesquisas da área (45%), o chinês foi o mais citado (26,4%). Considerando os idiomas, percebe-se uma lacuna em relação à estudos específicos para o tratamento de textos em idiomas diferentes de inglês ou chinês. O idioma português, por exemplo, foi citado em apenas 4 dos estudos considerados neste mapeamento.

A presença do idioma chinês também aparece ao se analisar o uso de fontes externas de

Figura 7 – Resultados do mapeamento dos estudos de mineração de textos que consideram a semântica.

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

conhecimento (Figura 7c). A base de conhecimento HowNet⁹, que cobre os idiomas chinês e inglês, foi a terceira fonte externa de conhecimento mais utilizada entre os estudos considerados. A fonte externa mais utilizada é a WordNet¹⁰ (29,9%), seguida por Wikipedia¹¹ (16,6%) e HowNet (5,2%).

Considerando-se os métodos e algoritmos utilizados nos estudos (Figura 7d), o tratamento

⁹ HowNet: <<http://www.keenage.com/>>

¹⁰ WordNet: <<http://wordnet.princeton.edu/>>

¹¹ Wikipedia: <<http://www.wikipedia.org/>>

da semântica latente, em especial com a aplicação de *Latent Semantic Index*, destaca-se entre os estudos considerados. Além disso, a aplicação de algoritmos de AM tradicionais, tais como SVM, k-NN e K-means são frequentemente aplicados nesses estudos. Os resultados indicam que muitos trabalhos também fazem uso de métodos de PLN e/ou de fontes externas de conhecimento.

A participação de especialistas do domínio pode ser crucial em aplicações reais do processo de MT. No entanto, como pode ser visto na [Figura 7e](#), a participação de usuários é pouco explorada em artigos científicos. A dificuldade inerente na avaliação de métodos que consideram a interação com usuários é uma razão provável para a escassez de estudos com esse tipo de abordagem.

Em relação às tarefas de Mineração de Textos, a classificação de textos é a tarefa mais aplicada em estudos de MT que consideram a semântica, aparecendo em 27,4% dos estudos. Em seguida aparece a tarefa de agrupamento de textos, em 17% dos estudos considerados ([Figura 7f](#)). Entre as tarefas de mineração mais específicas, que podem ser resolvidas em parte pela classificação ou agrupamento, aparecem a extração de informação (8,9%), recuperação de informação (7,9%) e a análise de sentimentos (7,2%). A análise de sentimentos é vista como um destaque nesse mapeamento, visto que é uma área mais recente e aparece com frequência semelhante a áreas mais consolidadas como extração e recuperação de informação. Além disso, a SentiWordNet¹², que é um recurso léxico para análise de sentimentos e mineração de opinião, aparece em entre as fontes externas mais citadas nos estudos (2,4%).

Considerando-se a representação dos textos, observou-se a busca por representações no modelo espaço-vetorial enriquecidas em relação à tradicional *bag-of-words*, além de representações baseadas em rede. Grande parte dos estudos são desenvolvidos para aplicações ou domínios específicos. Além da representação dos textos, a semântica também é incorporada ao processo por meio de uso de fontes externas de conhecimento, como redes semânticas e ontologias.

Durante a realização desse mapeamento da literatura, além das questões apresentadas na [Figura 6](#), também buscou-se identificar as coleções de textos utilizadas nos estudos. Foi identificado o uso de 211 coleções ou fontes de textos, sendo que 10 aparecem em mais de 46% dos estudos cujas coleções/fontes de documentos foram identificadas. Essas 10 coleções/fontes são: 20-newsgroups, Twitter, MEDLINE, Reuters-21578, TREC, PubMed, *Web*, Reuters (versão não especificada), Wikipedia e OHSUMED.

O mapeamento indica a predominância de pesquisas que tratam a semântica de maneira latente e a aplicação de métodos tradicionais de Aprendizado de Máquina. Também reafirma a lacuna no tratamento de idiomas diferentes de inglês e chinês, e a importância das tarefas de classificação e agrupamento em aplicações do processo de Mineração Textos. Detalhes sobre o mapeamento e os resultados completos são apresentados no artigo de [Sinoara, Antunes e Rezende \(2017\)](#). No contexto apresentado por este mapeamento, na próxima seção é apresentada uma

¹² SentiWordNet: <http://sentiwordnet.isti.cnr.it/>

nova perspectiva sobre a tarefa de classificação de textos considerando-se aspectos semânticos dos dados, definindo-se assim o problema tratado neste trabalho.

2.5 Definição do problema: níveis de complexidade semântica na mineração de textos

Nesta seção é discutido o problema estudado neste trabalho, sendo apresentada uma nova perspectiva sobre o problema de classificação de textos. Sob essa nova perspectiva, é definido o que se considera níveis de complexidade semântica envolvida nos problemas de classificação. As coleções de textos e os *datasets* utilizados em avaliações experimentais, que ilustram esses níveis de complexidade semântica, também são descritos e analisados nesta seção, bem como os aspectos semânticos que são abordados nos próximos capítulos desta tese.

2.5.1 Níveis de complexidade semântica

Como já apresentado na Seção 2.3, o entendimento da língua natural é um processo complexo. A fim de se entender o significado de textos escritos em língua natural é necessário ter conhecimento sobre: (i) vocabulário utilizado, ou seja, conhecer o significado das palavras; (ii) gramática do idioma, ou seja, conhecer as regras que definem como as palavras são utilizadas e combinadas; (iii) relações semânticas entre os itens linguísticos, tais como sinonímia e hiperonímia; e (iv) conhecimento de mundo e do contexto no qual os textos foram escritos. Textos são uma fonte rica de conhecimento, porém seu formato não estruturado e passível de ambiguidade, sarcasmo, ironia e outros fenômenos que podem alterar o significado composicional do que é dito (RIEMER, 2010), trazem grandes desafios ao processo de descoberta automática de conhecimento.

Pode-se perceber algumas características importantes e que podem ter impacto no processo de MT com alguns exemplos. Na introdução desta tese foram apresentados os exemplos que são aqui repetidos, por comodidade, na Figura 8.

Figura 8 – Exemplo 1: Documentos ilustrativos

Ex1.D1 A Empresa Alfa adquiriu a Empresa Beta.

Ex1.D2 A Empresa Beta adquiriu a Empresa Alfa.

Ex1.D3 A Empresa Beta foi adquirida pela Empresa Alfa.

Ex1.D4 A Empresa Alfa comprou a Empresa Beta.

Fonte: Elaborada pelo autor.

Para o entendimento dessas sentenças e a relação entre elas, além do vocabulário, estão envolvidos:

Relação sujeito x objeto. As sentenças *Ex1.D1* e *Ex1.D2* são diferenciadas pela sintaxe, pela ordem com que as mesmas palavras aparecem. Ambas possuem as mesmas palavras, porém o sujeito de *Ex1.D1* é o objeto de *Ex1.D2* e vice-versa.

Voz ativa x voz passiva. As sentenças *Ex1.D1* e *Ex1.D3* reportam o mesmo fato, apesar de terem sujeitos e objeto/agente da passiva opostos.

Sinonímia. As sentenças *Ex1.D1* e *Ex1.D3* reportam o mesmo fato, apesar de usarem verbos diferentes.

A fim de ir um pouco mais adiante nessa análise e ilustrar como esses e outros fatores podem afetar o problema de classificação automática de documentos, apresenta-se um outro exemplo. Considere a existência de uma coleção de notícias sobre diferentes esportes e que precisa ser classificada por esporte. Neste cenário, a questão que envolve a organização dessa coleção seria “Qual é o assunto do documento?” ou “Sobre o que é cada documento?”. Suponha que nessa coleção exista os documentos apresentados na [Figura 9](#).

Figura 9 – Exemplo 2: Documentos ilustrativos

<i>Ex2.D1</i>	<i>Guga é o campeão do Tennis Masters Cup. Ele venceu Agassi por três sets a zero no jogo final.</i>
<i>Ex2.D2</i>	<i>Hamilton larga na pole position e vence o Grande Prêmio do Canadá. Após colisão, Massa abandona a prova.</i>

Fonte: Elaborada pelo autor.

O documento *Ex2.D1* possui os termos “Guga”, “Tennis Masters Cup”, “sets”, “Agassi” e “jogo”. E o documento *Ex2.D2* possui os termos “Hamilton”, “pole position”, “Grande Prêmio”, “Massa” e “prova”. Esses termos são bem característicos de seus esportes. Assim, considerando esses termos, pode-se dizer que os documentos são de dois grupos (ou classes) distintos: *Ex2.D1* é sobre Tênis e *Ex2.D2* é sobre Fórmula 1. Nesse exemplo, cada esporte (cada grupo esperado ou classe conhecida) possui seus termos (ou palavras-chave) específicos. Os documentos de um mesmo esporte terão palavras similares. Portanto, a classe (ou grupo esperado) pode ser determinada em grande parte pelo vocabulário utilizado.

No entanto, usuários diferentes ou situações diferentes podem requerer outras naturezas de classificação ou organização dos mesmos documentos. Considere agora que é desejado organizar a mesma coleção de notícias de esportes de outra perspectiva. Pode ser interessante

organizar os documentos em relação ao desempenho de atletas brasileiros em competições. Portanto, nesse novo cenário, a questão que envolve a organização da coleção de documentos seria “Esse documento refere-se a vitória de um atleta brasileiro?”. Considerando novamente os documentos *Ex2.D1* e *Ex2.D2* da Figura 9, para este caso, as informações importantes são “Guga é o campeão” e “Massa abandona a prova”. E para organizar corretamente esses documentos é necessário saber que Guga e Massa são atletas brasileiros. Como isso, pode-se dizer que *Ex2.D1* refere-se a uma vitória de brasileiro e *Ex2.D2* refere-se a uma derrota.

Nesse contexto, nesta tese os problemas de classificação de documentos são divididos em dois níveis de complexidade semântica. O primeiro nível, que é chamado de organização por tópico, consiste em problemas de classificação que dependem basicamente do vocabulário. Nesse problema, cada classe possui termos bastante característicos, e, portanto, o léxico (vocabulário) possui grande relevância para representar o conteúdo dos documentos. Pode-se dizer que os documentos podem ser diferenciados em grande parte pelas palavras utilizadas¹³.

O segundo nível de complexidade semântica engloba os demais problemas de classificação de documentos. Esse segundo nível é chamado de organização semântica, no sentido de que se necessita mais do que apenas o léxico para resolvê-lo. Tais problemas requerem uma análise mais profunda, além apenas das palavras, visto que os documentos de classes distintas podem usar o mesmo vocabulário.

Os problemas de organização de documentos tratados na área de MT são tradicionalmente problemas do primeiro nível de complexidade semântica, a organização por tópico. Seja por meio da classificação ou do agrupamento, normalmente espera-se organizar os documentos com base no assunto dos mesmos. Esse fato fica claro logo na introdução de Sebastiani (2002). Tal artigo apresenta uma revisão completa sobre classificação automática de textos por meio de Aprendizado de Máquina, sendo bastante conceituado na área, com 2,462 citações na base de dados da Web of Science¹⁴, (585 citações nos últimos 3 anos). Em sua revisão, o autor apresenta a tarefa de classificação automática de textos como sendo uma tarefa de detecção de tópicos, com a rotulação de textos escritos em língua natural por meio da atribuição de uma categoria temática presente em uma lista de categorias pré-definidas. Nas palavras do autor: “*Text categorization (TC — a.k.a. text classification, or topic spotting), the activity of labeling natural language texts with thematic categories from a predefined set*” (SEBASTIANI, 2002, p. 1). Pode-se perceber que essa definição corresponde aos problemas do primeiro nível de complexidade semântica definido nesta tese como organização por tópico.

A predominância dos problemas de organização por tópico nas pesquisas de MT também

¹³ Vale notar que as palavras de conteúdo (como substantivos, adjetivos e verbos) carregam mais informação do que as palavras funcionais (como artigos, preposições e conjunções) e, portanto, são mais importantes para a organização dos documentos.

¹⁴ Número de citações em 20 de dezembro de 2017: <http://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=2&SID=6FUvF5Z7wsz4F97SzaH&page=1&doc=1>

pode ser verificada ao se analisar coleções de textos de *benchmarking*. Conforme apresentado na Seção 2.4.2, as principais coleções ou fontes de textos utilizadas nos estudos mapeados são: 20-newsgroups, Twitter, MEDLINE, Reuters-21578, TREC, PubMed, *Web*, Reuters (versão não especificada), Wikipedia e OHSUMED. Em relação a coleções de textos de *benchmarking* específicas para classificação ou agrupamento de textos, Rossi, Marcacini e Rezende (2013) apresentam 45 coleções, entre as quais aparecem 7 das 10 mais citadas no mapeamento sistemático (considerando-se a coleção completa ou um subconjunto da coleção). Neste trabalho, essas 45 coleções de *benchmarking* foram analisadas e, no Quadro 1, é apresentada uma descrição do problema tratado nessas coleções e a classificação dos mesmos em relação ao nível de complexidade semântica (coluna “Classificação semântica”), conforme definido nesta tese. Vale notar que muitas coleções são disponibilizadas apenas no formato estruturado, normalmente na representação *bag-of-words*, e, portanto, não foi possível confirmar a natureza das classes de algumas coleções. Nesses casos, marcados no Quadro 1 com um “*”, a informação apresentada é o que se pode deduzir a partir das informações disponíveis sobre a coleção de textos em questão.

No Quadro 1, pode-se verificar a predominância de problemas de organização por tópico entre as coleções de *benchmarking*. Apenas três das 45 coleções são classificadas como organização semântica. Essas três coleções são coleções de análise de sentimentos e visam a classificação do sentimento (polaridade) no nível de documento. A classificação do sentimento é um caso particular de problemas do segundo nível de complexidade semântica. Nesse tipo de classificação, palavras e expressões de sentimento são indicadores importantes do sentimento manifestado no documento. Palavras de sentimento são palavras que normalmente são utilizadas para expressar sentimentos positivos ou negativos, tais como “bom”, “péssimo”, “incrível”. Dada a importância de tais palavras na análise de sentimentos, várias pesquisas têm o foco na criação de listas de palavras de sentimento (*sentiment lexicon*) (LIU, 2012). Um exemplo de recurso léxico bastante utilizado em aplicações de análise de sentimentos para textos em inglês é a SentiWordNet (BACCIANELLA; ESULI; SEBASTIANI, 2010). Para textos em português, Balage Filho, Pardo e Aluísio (2013) comparam três recursos léxicos de sentimento: LIWC, OpinionLexicon e SentiLex. Assim, a classificação de sentimento pode ser resolvida por meio de métodos supervisionados ou não supervisionados (LIU, 2012). Os métodos não supervisionados normalmente se baseiam nas listas de palavras e expressões de sentimento. Já os métodos supervisionados normalmente utilizam algoritmos de AM tradicionais.

Enquanto em problemas de classificação por tópico os termos de domínio são importantes para distinguir as classes dos documentos, na classificação de sentimento as palavras de sentimento, que normalmente são adjetivos ou advérbios, são de grande importância. Esse fato pode justificar os bons resultados reportados com o uso da BOW. No entanto, apesar de importantes, as palavras de sentimento não são suficientes para resolver o problema de análise de sentimentos (LIU, 2012), encaixando esse problema no segundo nível de complexidade semântica. Entre os desafios desse problema estão: (i) palavras de sentimento podem ter orientações (positiva ou negativa) diferentes em diferentes contextos ou domínios de aplicação; (ii) uma

Quadro 1 – Problemas tratados em coleções de textos de *benchmarking* apresentadas por Rossi, Marcacini e Rezende (2013)

Coleção	Fonte	Natureza dos documentos	Natureza das classes	Classificação semântica
20-Newsgroups	-	mensagens de <i>e-mail</i>	assuntos dos newsgroups (diversos domínios)	organização por tópico
ACM	ACM digital library	artigos científicos de ciências da computação	áreas de pesquisa	organização por tópico
Classic4	CACM (<i>Communications of the ACM</i>), CISI (recuperação de informação), CRANFIELD (sistemas aeronáuticos) e MEDLINE	artigos científicos (diferentes formatos)	coleções de origem dos documentos (domínios específicos)	organização por tópico
CSTR	Departamento de Ciências da Computação da <i>University of Rochester</i>	relatórios técnicos	áreas de pesquisa	organização por tópico
Dmoz-Business-500	DMOZ - Open Directory Project	páginas <i>web</i>	categorias das páginas	organização por tópico
Dmoz-Computers-500	DMOZ - Open Directory Project	páginas <i>web</i>	categorias das páginas	organização por tópico
Dmoz-Health-500	DMOZ - Open Directory Project	páginas <i>web</i>	categorias das páginas	organização por tópico
Dmoz-Science-500	DMOZ - Open Directory Project	páginas <i>web</i>	categorias das páginas	organização por tópico
Dmoz-Sports-500	DMOZ - Open Directory Project	páginas <i>web</i>	categorias das páginas	organização por tópico
Enron-Top-20	-	mensagens de <i>e-mail</i>	assuntos tratados nas mensagens*	organização por tópico
FBIS	TREC	notícias	assuntos das notícias*	organização por tópico
Hitech	San Jose Mercury newspaper	notícias	assuntos das notícias (categorias)	organização por tópico
Industry-Sector	-	páginas <i>web</i> de empresas	setores de atuação das empresas	organização por tópico
Irish-Sentiment	RTE News, The Irish Times e The Irish Independent	artigos de jornais online	polaridades dos artigos	organização semântica
La1s	TREC	notícias	assuntos das notícias (categorias)	organização por tópico
La2s	TREC	notícias	assuntos das notícias (categorias)	organização por tópico
LATimes	união de La1s e La2s	notícias	assuntos das notícias (categorias)	organização por tópico
Multi-Domain-Sentiment	Amazon	opiniões sobre produtos	polaridades das opiniões	organização semântica
New3	San Jose Mercury newspaper	notícias	tópicos das notícias	organização por tópico
NSF Collection	<i>National Science Foundation</i>	textos científicos	áreas de pesquisas	organização por tópico

Continua na página seguinte.

Quadro 1 – Problemas tratados em coleções de textos de *benchmarking* apresentadas por Rossi, Marcacini e Rezende (2013) (continuação)

Continuação da página anterior.				
Coleção	Fonte	Natureza dos documentos	Natureza das classes	Classificação semântica
Oh0	OHSUMED	textos da área médica	tópicos dos textos (diversos)	organização por tópico
Oh5	OHSUMED	textos da área médica	tópicos dos textos (diversos)	organização por tópico
Oh10	OHSUMED	textos da área médica	tópicos dos textos (diversos)	organização por tópico
Oh15	OHSUMED	textos da área médica	tópicos dos textos (diversos)	organização por tópico
Ohscal	OHSUMED	textos da área médica	tópicos dos textos (diversos)	organização por tópico
Ohsumed-400	OHSUMED	textos da área médica	doenças relacionadas	organização por tópico
Opinosis	Tripadvisor (hotéis), Edmunds (carros) e Amazon (eletrônicos)	opiniões sobre produtos	aspectos dos produtos	organização por tópico
Pubmed-Cancer	Pubmed	textos científicos sobre câncer	órgãos afetados ou tipos de câncer	organização por tópico
Re0	Reuters-21578	notícias	assuntos das notícias (categorias)	organização por tópico
Re1	Reuters-21578	notícias	assuntos das notícias (categorias)	organização por tópico
Re8	Reuters-21578	notícias	assuntos das notícias (categorias)	organização por tópico
Review-Polarity	-	opiniões sobre filmes	polaridades das opiniões	organização semântica
Reviews	San Jose Mercury newspaper	notícias	assuntos das notícias (categorias)	organização por tópico
SpamAssassin	-	mensagens de <i>e-mail</i>	classificações em <i>spam</i> ou não <i>spam</i>	organização por tópico
SpamTrec-3000	-	mensagens de <i>e-mail</i>	classificações em <i>spam</i> ou não <i>spam</i>	organização por tópico
SyskillWebert	-	páginas <i>web</i>	categorias das páginas	organização por tópico
Tr11	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr12	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr21	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr23	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr31	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr41	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
Tr45	TREC	textos relevantes a certas consultas	tópicos*	organização por tópico
WAP	WebACE Project	páginas <i>web</i>	categorias das páginas	organização por tópico
WebKB	WebKB Project	páginas <i>web</i>	categorias das páginas	organização por tópico

sentença contendo palavras de sentimento pode não expressar um sentimento; (iii) o autor do texto pode estar sendo sarcástico; e (iv) sentenças sem palavras de sentimento podem conter opiniões implícitas.

Motivado pela riqueza e complexidade das línguas naturais, pela existência de diferentes objetivos de classificação no mundo real e pelo potencial da classificação automática de textos com o uso de algoritmos de AM, pretende-se nesta tese chamar a atenção aos problemas de maior complexidade semântica. Espera-se explorar a classificação de documentos sob a perspectiva de tipos mais diversos de classificação, cujos resultados ainda são deficientes e apresentam-se como uma grande lacuna na área de pesquisa de Mineração de Textos por meio de algoritmos de Aprendizado de Máquina.

2.5.2 Coleções de documentos e datasets utilizados

Com o objetivo de se analisar os dois níveis de complexidade semântica em problemas de classificação, as avaliações experimentais apresentadas nesta tese foram realizadas, principalmente, com duas coleções de textos: *BEST sports* e **SE-ABSA15**, sendo que cada coleção possui versões diferentes. As versões foram criadas para representar cenários reais de aplicação, nos quais diferentes usuários e situações necessitam de diferentes classificações para uma mesma coleção de textos. Cada versão forma um conjunto de dados (*dataset*) diferente, ou seja, cada versão pode ser vista como um *gold standard* distinto, referente a um objetivo específico de classificação para a coleção de documentos. Além dessas duas coleções principais, uma terceira coleção de textos (*BBC*) também foi utilizada para complementar as avaliações experimentais neste trabalho. As três coleções de textos e os respectivos *datasets* são apresentados a seguir.

BEST sports. É uma coleção de notícias sobre esportes, escritas em português (SINOARA; REZENDE, 2018). As notícias foram extraídas do *website BEST sports*¹⁵, cujo foco principal é a divulgação de notícias e resultados de Jogos Olímpicos e principais campeonatos mundiais de vários esportes. A coleção contém 881 notícias curtas (entre 383 e 2779 caracteres), publicadas entre agosto de 1999 e agosto de 2008. A maioria das notícias são do ano de 2004, principalmente devido aos Jogos Olímpicos de Verão de Atenas. De acordo com a classificação do *website*, cada notícia tem um rótulo de classe que corresponde a um esporte específico ou a um evento esportivo, de um total de 66 classes. Notícias com rótulo correspondente a um esporte específico reportam resultados de jogos ou competições de tal esporte. Por exemplo, um documento com o rótulo “Fórmula 1” pode relatar o pódio do primeiro Grande Prêmio da temporada de 2004. Já uma notícia com rótulo correspondente a um evento esportivo relata um acontecimento relacionado ao evento em geral. Por exemplo, um documento com o rótulo “Jogos Olímpicos de Verão” pode relatar a cerimônia de abertura de uma edição específica dos Jogos Olímpicos.

¹⁵ *BEST sports* - Arquivo de notícias: <<http://bestsports.com.br/db/notarqhome.php>>

Considerando essa classificação, a coleção *BEST sports* é desbalanceada. As 10 classes com mais documentos totalizam 502 documentos, ou seja, mais da metade do número total de documentos da coleção. Esse primeiro *dataset* da coleção *BEST sports* é chamado de *BS-full*.

A partir da coleção *BEST sports*, criou-se um subconjunto contendo as quatro maiores classes da coleção: “Fórmula 1”, “Motovelocidade”, “Futebol” e “Tênis”. Esse subconjunto, denominado *Best sports - Top 4 (BS-Top4)*, é composto por 283 notícias. Essas notícias foram manualmente rotuladas considerando-se a performance de atletas brasileiros. Cada documento recebeu uma das seguintes classes “Brasileiro venceu”, “Brasileiro não venceu”, “Brasileiro não citado” e “Não definido”. Os detalhes do processo de rotulação manual são apresentados no Apêndice A.

Com os resultados da rotulação manual, obteve-se três *datasets* para a coleção *Best sports - Top 4*: (i) *BS-topic*: classificação por esporte; (ii) *BS-semantic*: classificação por desempenho de atleta brasileiro; e (iii) *BS-topic-semantic*: classificação por ambos, esporte e desempenho de atleta brasileiro.

SE-ABSA15. É uma coleção de documentos composta por opiniões de usuários sobre restaurantes, *laptops* e hotéis. Os documentos foram criados e disponibilizados para a avaliação conjunta *SemEval-2015 Aspect Based Sentiment Analysis* (PONTIKI *et al.*, 2015). Essa coleção de análise de sentimentos é de alta qualidade, visto que foi criada seguindo um processo controlado e bem definido. Por ter sido realizada por anotadores treinados, a anotação de polaridade foi realizada de maneira consistente. Isso reduz os efeitos da subjetividade do anotador, como pode ocorrer quando a polaridade ou sentimento é derivado de avaliação fornecida pelo próprio usuário autor do texto opinião, como, por exemplo, por meio de um número estrelas. Além disso, os textos apresentam poucos problemas ortográficos que são comuns em textos de opinião.

A coleção original¹⁶ apresenta 830 documentos. Para os experimentos realizados nesta tese, a coleção *SE-ABSA15* é formada por 815 destes documentos. A remoção de 15 documentos da coleção original foi realizada pelos seguintes motivos: (i) 1 documento foi removido por não ter polaridade anotada; e 14 documentos foram removidos por não ser possível identificar papéis semânticos no texto (tal informação era necessária para a realização dos experimentos apresentados no Capítulo 4).

As anotações da base original atribuem polaridade (positivo, negativo ou neutro) a cada aspecto das entidades avaliadas nas revisões. Para a geração dos *datasets* da coleção *SE-ABSA15*, a polaridade do documento foi dada pela polaridade predominante entre as polaridades dos aspectos avaliados no documento. Aspectos avaliados como neutro foram contabilizados em ambas as polaridades, positiva e negativa. Documentos que possuem o mesmo número de aspectos positivos e negativos foram rotulados como neutro. Portanto,

¹⁶ Coleção original disponível em: <<http://alt.qcri.org/semeval2015/task12/>>

vale ressaltar que, neste trabalho, a classe neutro refere-se a documentos para os quais não foi possível identificar a polaridade positiva ou negativa em nível de documento, não referindo-se, assim, a documentos que não apresentam opinião. Com isso, foram gerados três *datasets* para essa coleção: (i) *SE-product*: classificação por tipo de produto (restaurante, *laptop* ou hotel); (ii) *SE-polarity*: classificação por polaridade (positivo, negativo ou neutro); e (iii) *SE-product-polarity*: classificação por ambos, tipo de produto e polaridade da opinião.

BBC. É uma coleção de 2225 notícias rotuladas do *website* BBC news¹⁷ (GREENE; CUNNINGHAM, 2006). A classe dos documentos corresponde a uma categoria, podendo ser “Negócios” (*Business*), “Entretenimento” (*Entertainment*), “Política” (*Politics*), “Esporte” (*Sport*) ou “Tecnologia” (*Tech*).

O Quadro 2 apresenta a descrição das coleções de textos e respectivos *datasets*. Os *datasets* são variados em termos de idioma, número de documentos e número de classes. Além disso, eles também apresentam os dois níveis de complexidade semântica estudados neste trabalho. Considerando os objetivos de classificação (classes pré-definidas) e o conteúdo dos documentos, pode-se dizer que *BBC*, *SE-product* e *BS-topic* são *datasets* do primeiro nível de complexidade semântica, ou seja, organização por tópico. Em *BBC* e *BS-topic* as classes são categorias das notícias e em *SE-product* são as categorias de produtos. Já *SE-polarity*, *SE-product-polarity*, *BS-semantic* e *BS-topic-semantic* estão no segundo nível de complexidade semântica, ou seja, organização semântica. Em *SE-polarity* e *SE-product-polarity*, as classes envolvem a polaridade do texto de opinião e, portanto, envolvem o conceito do que é positivo ou negativo no domínio em questão. Em *BS-semantic* e *BS-topic-semantic*, as classes envolvem o desempenho de atletas brasileiros nos jogos ou competições esportivas.

Para realizar uma análise objetiva dos conjuntos de dados e os níveis de complexidade semântica, foi realizada uma análise intrínseca dos dados, utilizando o critério da largura de silhueta (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010; ROUSSEEuw, 1987). Considerando um documento $d_i^{c_j}$ um dos n documentos pertencente à classe c_j , a largura de silhueta de $d_i^{c_j}$ ($s(d_i^{c_j})$) é calculada segundo a Equação 2.9.

$$s(d_i^{c_j}) = \frac{b(d_i^{c_j}) - a(d_i^{c_j})}{\max(a(d_i^{c_j}), b(d_i^{c_j}))} \quad (2.9)$$

$$a(d_i^{c_j}) = \frac{1}{n-1} \sum_{r=1}^n (dist(d_i^{c_j}, d_r^{c_j})) \quad (2.10)$$

$$b(d_i^{c_j}) = \min_{c_k \in C, k \neq j} \left(\frac{1}{|c_k|} \sum_{r=1}^{|c_k|} (dist(d_i^{c_j}, d_r^{c_k})) \right) \quad (2.11)$$

A largura de silhueta de um documento é uma relação entre a distância média deste documento a todos os outros documentos da mesma classe (Equação 2.10) e a distância média a

¹⁷ Coleção disponível em: <http://mlg.ucd.ie/datasets/bbc.html>

Quadro 2 – Descrição das coleções de textos e *datasets*

Coleção de textos	Idioma	Natureza dos documentos	#Docs	#Palavras por doc.		
				min.	max.	média
<i>BBC</i>	Inglês	Notícias	2225	89	4432	384,04
<i>SE-ABSA15</i>	Inglês	<i>Reviews</i>	815	4	572	75,61
<i>BS-Top4</i>	Português	Notícias	283	64	457	192,20

Coleção de textos	<i>Dataset</i>	#classes	Desvio padrão das classes	Classe majoritária	Largura de silhueta média	Classif. semântica
<i>BBC</i>	<i>BBC</i>	5	2,73%	22,97%	0,0465	tópico
<i>SE-ABSA15</i>	<i>SE-product</i>	3	26,28%	53,74%	0,0507	tópico
	<i>SE-polarity</i>	3	31,80%	66,75%	0,0102	semântica
	<i>SE-product-polarity</i>	9	13,27%	32,88%	0,0024	semântica
	<i>BS-topic</i>	4	4,91%	32,16%	0,1978	tópico
<i>BS-Top4</i>	<i>BS-semantic</i>	4	9,04%	32,86%	0,0083	semântica
	<i>BS-topic-semantic</i>	15	3,85%	16,96%	0,0173	semântica

todos os documentos da classe mais próxima (Equação 2.11). Ela permite analisar a compactação e a separação das classes conhecidas nos dados. Documentos que apresentam um valor alto (próximo a 1) para largura de silhueta são bem separáveis de documentos de outras classes. Valores baixos (próximos a 0) para largura de silhueta indicam que os documentos estão nas bordas das classes, ou seja, estão próximos a outras classes. E documentos com valores negativos (próximos a -1) para largura de silhueta são documentos que estão distantes dos documentos de sua classe e, portanto, provavelmente teriam sido colocados na classe errada. Nessa análise, as classes de referência para o cálculo da medida são as classes conhecidas (*gold standard*). Assim, documentos com valores negativos são interpretados como documentos difíceis de serem classificados corretamente, considerando a representação dos dados e a medida de distância utilizadas. No caso desta análise, a representação utilizada é a *bag-of-words* e a medida de distância é a Cosseno. Essa medida de distância foi selecionada por ser adequada a dados esparsos, sendo a medida mais comum na Mineração de Textos.

As larguras de silhueta dos conjuntos de dados, apresentadas no Quadro 2 e nas Figuras 10 e 11, são dadas pela média da largura de silhueta de todos os exemplos do respectivo *dataset*. Os maiores valores de largura de silhueta são dos *datasets* do primeiro nível de complexidade semântica (*BBC*, *SE-product* e *BS-topic*). Já os demais *datasets* são mais difíceis de classificar e apresentam valores mais baixos de largura de silhueta média, visto que suas classes são semanticamente mais complexas do que uma classificação por tópico.

2.5.3 Abordagens adotadas em direção à incorporação de aspectos semânticos na representação de textos

A tradicional representação *bag-of-words*, com a representação dos documentos por vetores cujas dimensões são termos presentes na coleção de documentos, baseia-se na *bag of*

Figura 10 – Largura de silhueta dos cenários das coleções *BBC* e *SE-ABSA 15*

Fonte: Elaborada pelo autor.

Nota – Em cada gráfico, a linha pontilhada indica a largura de silhueta média do respectivo *dataset*. A largura de silhueta foi calculada com base na representação *bag-of-words* utilizando a medida de distância Cosseno.

words hypothesis da área de recuperação de informação (TURNERY; PANTEL, 2010). Essa hipótese afirma que a relevância de um documento para uma consulta (*query*) pode ser estimada ao se representar o documento e a consulta como um saco de palavras (*bag of words*), ou seja, a frequência das palavras nos documentos indicam a relevância do documento para a consulta. A ideia é que o vetor que representa o documento captura um aspecto do significado do documento, que é o tema ou assunto do documento. Isso funciona bem para os problemas do primeiro nível de complexidade semântica discutidos anteriormente. Uma justificativa intuitiva para esse fato é que o tópico de um documento será influenciado pela escolha de palavras do autor ao escrever o documento (TURNERY; PANTEL, 2010). Assim, para se resolver problemas de maior complexidade semântica, que vão além da classificação por tópico, a representação *bag-of-words*, em geral, não é suficiente.

Figura 11 – Largura de silhueta dos cenários da coleção *BEST sports* (a) *BS-topic* - largura de silhueta média: 0,1978(b) *BS-semantic* - largura de silhueta média: 0,0083 (c) *BS-topic-semantic* - largura de silhueta média: 0,0173

Fonte: Adaptada de [Sinoara, Scheicher e Rezende \(2017\)](#).

Nota – Em cada gráfico, a linha pontilhada indica a largura de silhueta média do respectivo *dataset*. A largura de silhueta foi calculada com base na representação *bag-of-words* utilizando a medida de distância Cosseno.

Conforme apresentado neste capítulo, diversos fenômenos linguísticos influenciam o significado dos textos. Neste trabalho, são propostas e avaliadas diferentes abordagens para representar documentos no modelo espaço-vetorial, segundo diferentes perspectivas do tratamento semântico dos textos. Segundo [Widdows \(2004\)](#) modelos matemáticos são ferramentas poderosas para representar o significado de língua natural, porém não é viável a modelagem de todos os detalhes da língua. Toda abstração matemática usada como modelo linguístico contém simplificações e o desafio está em encontrar simplificações inteligentes que possibilitem que as características principais do evento modelado sejam aparentes. Neste trabalho, foram exploradas diferentes abordagens para representações visando a representação de diferentes características semânticas dos textos. Tais abordagens, apresentadas a seguir, visam manter parte das informações semânticas da língua natural na representação estruturada das textos.

Conhecimento do usuário. O conhecimento de especialista sobre o domínio e a coleção de documentos foi explorado na representação *gBoED*. Essa abordagem é apresentada no Capítulo 3.

Desambiguação lexical de sentidos e hiperônimos. A representação de documentos com base em termos desambiguados e vetores formados por hiperônimos foram exploradas na representação baseada em vetores *NASARI unified*. Essa abordagem é apresentada no Capítulo 3.

Papéis semânticos. Rótulos de papéis semânticos foram explorados na representações *SR-based Terms* e *SR-based Sentences*. Essa abordagem é apresentada no Capítulo 4.

Desambiguação lexical de sentidos e representação distribuídas de itens linguísticos. Termos desambiguados e representações vetoriais de palavras e de sentidos (*synsets* ou conjuntos de sinônimos) foram exploradas nas representações *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*. Essa abordagem é apresentada no Capítulo 5.

2.6 Considerações finais

Neste capítulo foram discutidos temas relacionados à análise de textos. Foram apresentados os conceitos básicos sobre Mineração de Textos, que são essenciais para a contextualização deste trabalho. Foi dado enfoque especial para a tarefa de representação de dados textuais, tema de estudo deste trabalho, e para a tarefa de classificação automática de textos, tarefa utilizada para avaliar os modelos de representação propostos. Um dos desafios da Mineração de Textos é o tratamento da semântica dos documentos. Isso foi destacado com a breve apresentação de algumas relações semânticas que influenciam o entendimento dos textos escritos em língua natural e de tarefas de Processamento de Língua Natural que podem apoiar o enriquecimento da representação dos documentos.

Um mapeamento sistemático da literatura realizado com o objetivo de obter uma visão geral sobre como a semântica tem sido tratada na Mineração de Textos, cujos resultados foram apresentados neste capítulo, indica a diversidade de trabalhos de MT que consideram a semântica dos documentos de alguma maneira. No entanto, esse estudo da literatura também mostrou que os trabalhos se concentram em domínios específicos, como biomedicina, e há uma carência na análise de textos em português.

Com o objetivo de apresentar as tarefas de organização (classificação ou agrupamento) de coleções textuais sob uma perspectiva semântica, neste capítulo foi definido o problema estudado neste trabalho. As principais coleções de textos de *benchmarking* foram analisadas e mostrou-se que os problemas tradicionalmente tratados nas pesquisas de Mineração de Textos são relacionados a uma organização por tópico. Por meio de exemplos, foi mostrado que esse tipo de organização pode ser resolvida por meio de termos isolados dos domínios envolvidos.

Porém, no mundo real, há problemas semanticamente mais complexos, sendo este um grande desafio para área de Mineração de Textos.

A fim de resolver tais problemas, é necessário que os modelos de representação de textos mantenha os padrões de interesse nos dados para que possam ser descobertos pelos algoritmos de Aprendizado de Máquina. Neste trabalho, foram propostos e avaliados diferentes modelos de representação de textos. Escolheu-se trabalhar com representações no modelo espaço-vetorial por sua simplicidade, por permitirem a utilização de tradicionais e bem estudados algoritmos de AM e pelo potencial que eles apresentam. Os modelos propostos e a avaliação dos mesmos são apresentados nos próximos capítulos desta tese.

EXPLORAÇÕES INICIAIS: REPRESENTAÇÕES BASEADAS EM CONHECIMENTO DO DOMÍNIO E HIPERÔNIMOS

3.1 Considerações iniciais

Neste capítulo são apresentadas explorações iniciais em busca da incorporação da semântica na representação de coleções de documentos. São apresentadas, como explorações iniciais no tema, duas representações de coleções de documentos propostas visando o enriquecimento semântico da representação em duas abordagens diferentes. A primeira representação, chamada de *generalized-bag-of-expressions-of-domain* (*gBoED*), visa o uso de conhecimentos de especialistas fornecidos por meio de listas de termos do domínio. Essa representação, apresentada na Seção 3.2, é uma generalização de uma representação proposta para um problema específico da área de desenvolvimento de produtos, mais especificamente Sistemas Produto-Serviço (MARQUES *et al.*, 2015).

A segunda abordagem, chamada de *Unified vector-based* (*Uni-based*), visa a incorporação de relações semânticas de hiperonímia/hiponímia (relações taxonômicas) por meio do uso de vetores de representação de conceitos NASARI *unified* (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016). Para isso, essa representação é gerada a partir dos documentos desambiguados. A abordagem NASARI, bem como a representação de coleções de documentos proposta, são apresentadas na Seção 3.3.

3.2 Representação baseada em expressões do domínio

Conforme apresentado no Capítulo 2, dada a importância da etapa de Pré-processamento para o sucesso do processo de Mineração de Textos, diversos trabalhos têm sido desenvolvidos com o objetivo de encontrar modelos de representação que sejam mais adequados às diferentes aplicações do processo. Considerando aspectos semânticos e conhecimento prévio que o usuário tem sobre a coleção de documentos, Marques *et al.* (2015) desenvolveram uma abordagem específica para o domínio de artigos da área de desenvolvimento de produtos, mais especificamente Sistemas Produto-Serviço. Tal trabalho foi realizado com o objetivo de facilitar a análise de artigos da área como parte de um projeto maior, que visa o desenvolvimento e disponibilização de um portal de conhecimentos das áreas de inovação, desenvolvimento de produtos, gestão do ciclo de vida de produtos e sustentabilidade¹.

Assim, em uma análise exploratória do uso de técnicas de Mineração de Textos na identificação da aplicação prática de métodos e ferramentas em artigos sobre Sistemas Produto-Serviço, Marques *et al.* (2015) propuseram um novo modelo, chamado de *bag-of-expressions-of-domain* (BoED). Nesse modelo, a coleção de artigos é representada com base em expressões do domínio, com o objetivo de identificar duas classes de documentos da área de Sistemas Produto-Serviço: (i) documentos que apresentam desenvolvimento teórico de métodos ou ferramentas; e (ii) documentos que apresentam aplicações práticas dos mesmos métodos ou ferramentas.

Assim como a BOW, a BoED também é uma representação no modelo espaço-vetorial. A principal diferença entre as duas representações é que, enquanto na BOW os termos são palavras independentes, na BoED os termos são formados por expressões do domínio. As expressões do domínio são formadas a partir de três listas de termos do domínio. Com isso, os atributos da representação carregam informações semânticas obtidas por meio dessas listas, que normalmente são fornecidas pelo usuário. Tais listas são descritas a seguir, conforme a definição de Marques *et al.* (2015).

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

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

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

$$A = \{a_1(s_{11}, \dots, s_{1i}), a_2(s_{21}, \dots, s_{2i}), \dots, a_p(s_{p1}, \dots, s_{pi})\}$$

¹ Portal de Conhecimentos: <<http://www.portaldeconhecimentos.org.br/>>

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

$$T = \{t_1(s_{11}, \dots, s_{1i}), t_2(s_{21}, \dots, s_{2i}), \dots, t_q(s_{q1}, \dots, s_{qi})\}$$

Segundo [Marques et al. \(2015\)](#), a tarefa de geração das três listas pode ser feita de forma manual ou utilizando técnicas de Mineração de Textos. Quando realizada de forma manual, deve-se selecionar um conjunto de artigos de referência e gerar, com apoio de especialistas, as três listas a partir da leitura desses artigos. Para a geração por meio de Mineração de Textos, foi explorado o uso de técnicas de reconhecimento de entidades nomeadas e regras de associação. No entanto, apesar dos resultados serem promissores, algumas deficiências foram identificadas na geração automática dessas listas.

Na representação BoED, cada expressão do domínio é composta por um termo da lista M associado a um termo da lista A ou da lista T . As expressões do domínio são buscadas em cada uma das sentenças do documento. A quantidade de ocorrências de cada expressão em cada documento é verificada e a BoED é construída para aquele conjunto de textos. Para ilustrar esse processo e comparar a representação BoED com a tradicional BOW, considere os seguintes documentos hipotéticos apresentados na [Figura 12](#).

Figura 12 – Exemplo 3: Documentos ilustrativos

D1 This paper proposes a Quality Function Deployment method.

D2 This paper proposes a QFD method.

D3 This paper presents a case study on Quality Function Deployment method.

Fonte: Adaptada de [Scheicher et al. \(2016\)](#).

Nota – Esses documentos ilustrativos foram apresentados por [Scheicher et al. \(2016\)](#) e foram criados com base nas listas de termos definidas por [Marques et al. \(2015\)](#).

[Marques et al. \(2015\)](#) apresenta as seguintes listas de termos geradas manualmente por um especialista do domínio.

1. $M = \{\text{Analytic Hierarchy Process (AHP), Brainstorming, Computer Aided Design (CAD), Conjoint Analysis, Delphi, Design for Assembly (DFA), Design for Disassembly (DFD), Design Structure Matrix (DSM), Eco-costs/Value Ratio Model (EVR Model), Failure Mode}$

and Effects Analysis (FMEA), Focus Group, Kansei Engineering, Life Cycle Assessment (LCA), Product-Service Blueprint, Quality Function Deployment (QFD), Technology Roadmap (TRM), Theory of Inventive Problem Solving (TRIZ)}

2. $A = \{\text{use (uses, using, used, usage), apply (applies, applying, applied, application), validate (validates, validating, validated), case study (case research, action research, cases, real case, practical case)}\}$
3. $T = \{\text{develop (develops, developing, developed), propose (proposes, proposing, proposed), introduce (introduces, introducing, introduced), suggest (suggests, suggesting, suggested), provide (provides, providing, provided)}\}$

Considerando essas três listas, as representações BOW (com remoção de *stopwords*) e BoED para os três documentos ilustrativos da Figura 12 são apresentadas na Figura 13. Nessa figura, é possível verificar a diferença entre as representações BOW e BoED. Os documentos $D1$ e $D2$ possuem o mesmo significado, pois QFD é a sigla de *Quality Function Deployment* (conforme a lista M). No entanto, na representação BOW os dois documentos são representados por vetores bastante distintos. Já na representação BoED, $D1$ e $D2$ são representados pelo mesmo vetor.

Figura 13 – Representações dos documentos $D1$, $D2$ e $D3$ da Figura 12

	paper	proposes	quality	function	deployment	method	QFD	presents	case	study
$D1$	1	1	1	1	1	1	0	0	0	0
$D2$	1	1	0	0	0	1	1	0	0	0
$D3$	1	0	1	1	1	1	0	1	1	1

(a) *bag-of-words*

	Quality-Function-Deployment_propose	Quality-Function-Deployment_case-study
$D1$	1	0
$D2$	1	0
$D3$	0	1

(b) *bag-of-expressions-of-domain*

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

Para possibilitar a aplicação da BoED para representação de coleções de documentos de qualquer domínio, foi proposta uma generalização da representação BoED, chamada de *gBoED*. Essa proposta torna a representação independente do domínio e independente do número de classes apresentado pela coleção de documentos.

3.2.1 Modelo proposto para representação de coleções de documentos: gBoED

Conforme apresentado na seção anterior, a representação BoED foi desenvolvida para representar conjuntos de documentos de uma área específica e com o objetivo específico de distinguir documentos que apresentam desenvolvimento teórico de documentos que apresentam aplicações práticas. A fim de possibilitar a aplicação dessa ideia de usar expressões do domínio em outros problemas e avaliá-la de maneira mais ampla, foi proposta uma generalização da BoED, denominada *generalized-bag-of-expressions-of-domain* (gBoED).

O ponto central da BoED é a geração de expressões do domínio a partir das três listas de termos do domínio definidas por Marques *et al.* (2015). Cada expressão é formada por um método ou ferramenta pertencente à lista M , seguido de um elemento de uma das duas outras listas (listas A e T). Os elementos da lista A são palavras ou expressões comumente utilizadas quando o autor do documento quer apresentar um uso prático de um método ou ferramenta. Da mesma forma, os elementos da lista T são utilizados quando se tem a apresentação do desenvolvimento teórico de um método ou ferramenta. Portanto, as listas A e T podem ser vistas como listas que possuem termos que são importantes ou relevantes para uma classe de documentos, ou seja, as listas A e T referem-se, respectivamente, à documentos de aplicação e documentos de desenvolvimento teórico.

Assim, para possibilitar a aplicação da representação baseada em expressões do domínio em problemas de classificação de qualquer domínio, a lista de métodos e ferramentas foi redefinida como um lista de termos do domínio. Já as listas de expressões que indicam desenvolvimento teórico ou aplicação prática (classes ou grupos esperados no problema de Marques *et al.* (2015)) foram redefinidas para listas de identificadores de classe. A definição dessas listas é apresentada a seguir.

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

$$\text{Domain_Keywords} = \{k_1, k_2, \dots, k_i\}$$

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

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

$$Class_Keywords_Set = \{\{ck_{11}, ck_{12}, \dots, ck_{1l}\}, \dots, \{ck_{m1}, ck_{m2}, \dots, ck_{mz}\}\}$$

Cada elemento da m -ésima lista do conjunto $Class_Keywords_Set$, ck_{mn} , é formado por um termo identificador de classe t e seus sinônimos s , isto é, $ck_{mz} = \{t_z\} \cup \{s_1, \dots, s_p\}$.

Assim como na BoED, na $gBoED$ os atributos são formados por expressões do domínio criadas a partir da combinação de elementos da lista $Domain_Keywords$ com elementos de uma das listas do conjunto $Class_Keyword_Set$. Um esquema da representação $gBoED$ para uma coleção de N documentos é apresentado na Figura 14.

Figura 14 – Esquema da representação de coleção de documentos $gBoED$

	$k_1_ck_{11}$	...	$k_1_ck_{1j}$	...	$k_i_ck_{11}$	...	$k_i_ck_{1j}$	...	$k_1_ck_{m1}$	...	$k_1_ck_{mz}$	...	$k_i_ck_{m1}$	...	$k_i_ck_{mz}$
d_1															
d_2															
$\vdots$															
d_{N-1}															
d_N															

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

3.2.2 Avaliação experimental - representação $gBoED$

Nessa seção é apresentada a avaliação experimental realizada com o objetivo de analisar o uso da representação $gBoED$ na classificação de documentos, considerando os diferentes níveis de complexidade semântica apresentados na Seção 2.5 (Página 58).

3.2.2.1 Datasets, representações de textos e configuração experimental

Os experimentos foram realizados utilizando-se os *datasets* das coleções *BS-Top4* e *SE-ABSA15*, apresentados na Seção 2.5.2 (Página 64). Para cada *dataset* foram geradas as representações BOW e $gBoED$. Para a construção das representações $gBoED$ da coleção *BS-Top4* foram utilizadas listas que contêm nomes de atletas brasileiros (lista de termos do domínio) e verbos usados para expressar vitórias e derrotas (listas de identificadores de classe). Para a coleção *SE-ABSA15* foram usadas listas que contêm aspectos dos produtos (lista de termos do domínio) e palavras positivas e negativas (listas de identificadores de classe).

As representações geradas foram avaliadas de acordo com a acurácia dos classificadores construídos utilizando-se sete algoritmos de AM. Também foi avaliada a combinação de classificadores gerados por ambas as representações, utilizando três estratégias para compor os votos entre os diferentes classificadores base: *Sum of Confidences* (SC), *Weighted Sum of Confidences* (WSC) e *Most Confident* (MC). Detalhes sobre a geração das representações, bem como sobre a configuração experimental adotada, são apresentados no Apêndice B.

3.2.2.2 Resultados - gBoED

Na [Tabela 1](#) são apresentados os melhores valores de acurácia obtidos na execução de cada algoritmo para cada *dataset* das coleções de documentos *BEST sports Top 4* e *SE-ABSA15*. São apresentados os resultados obtidos com classificadores gerados a partir das representações BOW e *gBoED*, bem como com a combinação de classificadores gerados com as duas representações. Pode-se notar que a representação *gBoED* não leva a boas acurácias quando é utilizada de modo independente, sendo que os melhores valores de acurácia obtidos são sempre menores do que as melhores acurácias obtidas com o uso da BOW. Apesar de ser formada por atributos mais expressivos do que simples palavras, a *gBoED* não apresenta informação suficiente para possibilitar a classificação correta dos documentos.

No entanto, ao se observar a combinação de classificadores gerados com as duas representações, verifica-se que a combinação obteve resultados melhores do que os obtidos com o uso da *gBoED* nos *datasets* de organização semântica. Vale notar que, por serem problemas de maior complexidade semântica, esses foram os problemas considerados para criar as listas de termos usadas na geração da *gBoED*.

Assim, considerando-se os diferentes níveis de complexidade semântica, verifica-se que para os *datasets* do primeiro nível (*BS-topic* e *SE-product*) a representação BOW é suficiente para atingir acurácia próxima a 100%. Já para os *datasets* do segundo nível (*BS-semantic*, *BS-topic-semantic*, *SE-polarity* e *SE-product-polarity*), a BOW não é suficiente, obtendo acurácias mais baixas. Nesses cenários, a combinação de classificadores gerados com a BOW e com a *gBoED* levou a melhores valores de acurácia na maioria das configurações testadas.

A representação *gBoED* apresentada nessa seção aborda a semântica dos textos por meio de conhecimento do domínio obtido a partir das listas de termos e identificadores de classe. Os resultados obtidos indicam que, a partir de listas adequadas ao objetivo da classificação, pode-se obter melhora nos resultados ao se considerar o uso combinado da BOW com a *gBoED*. A partir dessa combinação de classificadores gerados com as duas representações, a grande quantidade de dados, porém sem informação semântica, da BOW pode ser unida aos dados semanticamente mais ricos da *gBoED*, que possui atributos expressivos porém menos frequentes. Na próxima seção é apresentada uma outra abordagem para incorporação da semântica na representação de coleções de documentos, que tem como base a representação semântica de conceitos.

3.3 Representação baseada em hiperônimos

Visando a incorporação de relações semânticas de hiperonímia/hiponímia na representação de coleções de documentos, foi explorado o uso de vetores de representação de conceitos *NASARI unified*. *NASARI* (*Novel Approach to a Semantically-Aware Representation of Items*) é uma abordagem para representação semântica de sentidos de palavras ou conceitos. Proposta

Tabela 1 – Melhores acurácias para as coleções *BS-Top4* e *SE-ABSA15*

	BOW	<i>gBoED</i>	BOW+ <i>gBoED</i>		
			SC	WSC	MC
SE-product	99,5077	89,9353	99,5077	99,5077	99,5077
IMBHN ^C	98,1587	86,0193	98,5276	98,5276	98,1587
IMBHN ^R	99,1388	89,6914	99,0169	99,0169	99,1388
J48	92,2704	80,6158	92,2704	92,2704	92,2704
k-NN	98,5245	89,9353	98,7699	98,7699	98,6480
MNB	99,5077	86,2572	99,5077	99,5077	99,5077
NB	92,7612	84,5348	93,9898	94,4791	94,2307
SVM	96,4408	89,4550	96,3189	96,3189	96,4408
SE-polarity	84,5438	69,0786	84,5453	84,5453	84,5438
IMBHN ^C	80,4908	65,0316	80,6143	80,6128	80,4908
IMBHN ^R	82,8214	64,2939	81,9587	82,2057	82,8214
J48	71,5266	66,5056	72,6257	72,3893	72,7492
k-NN	77,2900	67,7281	76,9151	76,6787	77,1740
MNB	84,5438	66,7525	84,5453	84,5453	84,5438
NB	70,3071	63,6886	70,1867	70,3086	70,3071
SVM	81,6110	69,0786	81,2436	81,3640	81,6110
SE-product-polarity	83,6811	62,2162	83,8046	83,8046	83,6811
IMBHN ^C	77,4345	60,1235	77,7989	78,0367	77,5565
IMBHN ^R	73,9762	60,7347	72,5083	72,6302	73,9762
J48	71,0479	56,1999	71,6652	71,4213	71,6667
k-NN	75,8220	61,2195	75,9455	75,9455	75,9455
MNB	83,6811	62,2162	83,8046	83,8046	83,6811
NB	68,9521	57,0671	69,5604	69,5649	69,0696
SVM	77,8034	60,8627	78,0473	78,0473	77,8034
BS-topic	100,0000	78,4360	100,0000	100,0000	100,0000
IMBHN ^C	98,9286	67,2167	98,9286	98,9286	98,9286
IMBHN ^R	99,6429	68,9532	99,6429	99,6429	99,6429
J48	96,8227	61,8719	96,8227	96,8227	96,8227
k-NN	99,6552	73,1158	99,6552	99,6552	99,6552
MNB	100,0000	78,4360	100,0000	100,0000	100,0000
NB	99,6429	71,4163	99,6429	99,6429	99,6429
SVM	100,0000	71,7118	100,0000	100,0000	100,0000
BS-semantic	68,9532	53,7315	69,2857	69,2980	68,9532
IMBHN ^C	64,6552	44,2118	65,3818	65,3571	64,6552
IMBHN ^R	68,9532	43,1034	69,2857	69,2980	68,9532
J48	59,0517	46,3054	61,4655	63,2266	61,5025
k-NN	65,3818	53,7315	68,5345	68,1650	66,4778
MNB	59,7414	52,9926	63,9778	63,6330	60,4557
NB	57,6108	46,6379	59,0271	57,9557	59,0025
SVM	63,6576	48,8300	64,7167	64,7167	63,6576
BS-topic-semantic	66,8596	44,5074	68,9901	68,9901	66,8596
IMBHN ^C	62,5739	42,0443	67,0813	68,1281	62,9310
IMBHN ^R	57,9926	38,5099	56,5764	56,5764	57,9926
J48	55,1478	31,8596	56,2069	56,5640	57,2537
k-NN	65,7512	44,5074	66,7734	67,5000	65,7512
MNB	62,5985	44,1749	65,0369	63,9655	65,0369
NB	57,2537	43,8054	57,6108	57,6108	57,2537
SVM	66,8596	38,5345	68,9901	68,9901	66,8596

Nota – Acurácias maiores que as obtidas com o *baseline* BOW são apresentadas em negrito. A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia de cada representação para o respectivo *dataset*.

por Camacho-Collados, Pilehvar e Navigli (2016), a abordagem NASARI é composta por três variações de representação vetorial de conceitos: *lexical*, *embedded* e *unified*. Tais representações são obtidas por meio da exploração de diferentes fontes de conhecimento, tendo como base as relações semânticas da WordNet (FELLBAUM, 1998), o mapeamento entre os *synsets* da

WordNet e os artigos da Wikipedia existente na base de conhecimento BabelNet (NAVIGLI; PONZETTO, 2012) e os textos e relacionamentos dos artigos da Wikipedia.

De modo geral, os vetores NASARI representam conceitos relacionados a *synsets* da BabelNet a partir de informações contextuais obtidas de um corpus de referência. No caso do trabalho apresentado por Camacho-Collados, Pilehvar e Navigli (2016), a Wikipedia é utilizada como corpus de referência. Assim, para cada *synset* s da BabelNet é gerado um sub-corpus que representa as informações contextuais de s . Cada sub-corpus contém a página (artigo) da Wikipedia correspondente ao *synset*, se houver, além de páginas que possuam um link para a página do *synset* e de páginas correspondentes a hiperônimo e hipônimos de s na taxonomia da BabelNet. Em seguida, a informação contextual é transformada em representações vetoriais usando toda a Wikipedia como corpus de referência.

No contexto da representação de conceitos, NASARI apresenta as seguintes vantagens: (i) não requer uma grande quantidade de textos desambiguados e dados anotados; (ii) fornece representações já ligadas a uma base de sentidos (BabelNet); (iii) possui uma grande cobertura, incluindo conceitos e entidades nomeadas; (iv) não é restrita à representação de conceitos definidos na WordNet; (v) não é limitada ao idioma inglês; (vi) não é restrita a uma tarefa específica; e (vii) possibilita a comparação entre diferentes idiomas e itens linguísticos, como palavras e sentidos. Uma visão geral das três representações vetoriais da abordagem NASARI é apresentada a seguir.

NASARI *lexical*. Nessa representação vetorial cada conceito é representado por um vetor cujas dimensões são palavras individuais. Essa representação é similar à tradicional *bag-of-words*. No entanto, os pesos das palavras não são calculados por meio das medidas de frequência tradicionais. Os autores propõem o uso da medida *lexical specificity* e, assim, o peso de cada palavra é calculado com base nas frequências da palavra tanto no corpus de referência quanto no sub-corpus que corresponde às informações contextuais do conceito.

NASARI *embedded*. Essa representação combina a representação *lexical* com representações de palavras chamadas de *word embeddings*. A representação vetorial de palavras é abordada em mais detalhes nesta tese no Capítulo 5. As principais vantagens da representação NASARI *embedded* são: (i) ela representa conceitos em um espaço de baixa dimensionalidade; e (ii) os conceitos são representações no mesmo espaço semântico das *word embeddings* e, portanto, se um conceito de NASARI *embedded* e uma palavra presente no conjunto de *word embeddings* são relacionados, os respectivos vetores estarão próximos no espaço. Os vetores NASARI *embedded* foram utilizados na proposta apresentada no Capítulo 5.

NASARI *unified*. Nessa representação, cada dimensão do vetor corresponde a um *synset* da BabelNet. O primeiro passo para a construção de NASARI *unified* é o agrupamento de palavras do sub-corpus que possuam pelo menos um sentido compartilhando o mesmo

hiperônimo. As dimensões da representação são formadas pelos hiperônimos que superam um determinado valor de *lexical specificity*. Os pesos das dimensões são calculados utilizando *lexical specificity* com base em todos os hipônimos do hiperônimo. Uma vantagem da representação NASARI *unified* é que suas dimensões são *synsets* da BabelNet e, portanto, são interpretáveis e multilíngues.

A abordagem NASARI foi avaliada pelos seus autores considerando diferentes *datasets* de *benchmarking* e idiomas variados e apresentou resultado superior a métodos da literatura em diferentes tarefas de PLN, como similaridade semântica, agrupamento de sentidos, *domain labeling* e desambiguação lexical de sentidos (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016). Os autores disponibilizaram para a comunidade os vetores NASARI *lexical*, *embedded* e *unified* que foram gerados utilizando a Wikipedia como corpus de referência².

3.3.1 Modelo proposto para representação de coleções de documentos: Uni-based

Visando aproveitar informações provenientes de relações semânticas de hiperonímia/hiponímia incorporadas na representação de conceitos NASARI *unified*, foi proposto neste trabalho um novo modelo de representação de coleções de documentos. Esse modelo, chamado de *Unified vector-based (Uni-based)*, se baseia na desambiguação lexical de sentidos e no uso de vetores NASARI *unified* para representar coleções de documentos.

A desambiguação é utilizada para identificar os conceitos presentes no textos. O processo de desambiguação, realizado com a ferramenta Babelfy, é descrito na Seção 5.4 do Capítulo 5 (Página 116), visto que é o mesmo processo realizado para a geração dos outros modelos propostos. Após a execução do sistema Babelfy, cada fragmento dos documentos é desambiguado, retornando um conjunto de *synsets* da BabelNet.

A construção da representação *Uni-based* é ilustrada na Figura 15. O processo recebe como entrada o conjunto de documentos desambiguados (D') e o conjunto de vetores NASARI *unified* (U). Para cada documento, os vetores correspondentes aos *synsets* identificados no documento são buscados no conjunto NASARI *unified*. Assim, um subconjunto de vetores é retornado para cada documento. Vale notar que algumas das dimensões (*synsets* da BabelNet) desses vetores se sobrepõem, ou seja, um mesmo *synset* pode estar presente em mais de um vetor NASARI *unified*. Para construir o vetor que representa o documento, os vetores retornados são somados e é aplicada a função logarítmica para normalizar os pesos das dimensões. Como os vetores de todos os documentos calculados, a coleção de documentos é representada pela matriz cujas linhas são documentos e as colunas correspondem a *synsets* da BabelNet.

² NASARI: <<http://lcl.uniroma1.it/nasari/>>

Figura 15 – Construção da representação *Uni-based*

Fonte: Elaborada pelo autor.

3.3.2 Avaliação experimental - representação *Uni-based*

Nessa seção é apresentada a avaliação experimental realizada com o objetivo de analisar o uso da representação *Uni-based* na classificação de documentos, considerando-se os diferentes níveis de complexidade semântica apresentados na Seção 2.5 (Página 58).

3.3.2.1 Datasets, representações de textos e configuração experimental

Os experimentos foram realizados utilizando-se os *datasets* das coleções *BS-Top4* e *SE-ABSA15*, apresentados na Seção 2.5.2 (Página 64). Para cada *dataset* foram geradas as representações BOW e *Uni-based*. Para a construção das representações *Uni-based* foram utilizados os vetores disponibilizados por Camacho-Collados, Pilehvar e Navigli (2016).

Assim como na avaliação experimental da representação *gBoED*, apresentada na Seção 3.2.2, as representações geradas foram avaliadas de acordo com a acurácia dos classificadores construídos utilizando-se sete algoritmos de AM, além da combinação de classificadores gerados por ambas as representações, utilizando três estratégias para compor os votos entre os diferentes classificadores base: SC, WSC e MC. Detalhes sobre a geração das representações, bem como sobre a configuração experimental adotada, são apresentados no Apêndice B.

3.3.2.2 Resultados - *Uni-based*

Na Tabela 2 são apresentados os melhores valores de acurácia obtidos na execução de cada algoritmo para cada *dataset* das coleções de documentos *BS-Top4* e *SE-ABSA15*. São apresentados os resultados obtidos com classificadores gerados a partir das representações BOW e *Uni-based*, bem como com a combinação de classificadores gerados com as duas representações.

Comparando as melhores acurácias obtidas por cada algoritmo, pode-se verificar que o uso da representação *Uni-based* obteve resultado superior ao resultado obtido com o uso da

Tabela 2 – Melhores acurácias para as coleções *BS-Top4* e *SE-ABSA15*

	BOW	Uni-based	BOW+Uni-based		
			SC	WSC	MC
SE-product	99,1388	97,1771	99,1388	99,1388	99,1388
IMBHN ^C	98,1587	97,1725	98,7715	98,7715	98,5200
IMBHN ^R	99,1388	95,8266	99,1388	99,1388	99,1388
J48	92,2704	95,7031	95,8251	95,7031	95,7031
k-NN	98,5245	96,6862	99,0154	99,0154	98,5245
NB	92,7612	93,7383	93,8678	93,8678	93,7383
SVM	96,4408	97,1771	97,2990	97,2990	97,2990
SE-polarity	82,8214	73,7459	83,6781	83,6781	82,8214
IMBHN ^C	80,4908	68,7218	80,7362	80,7362	80,4908
IMBHN ^R	82,8214	73,7459	83,6781	83,6781	82,8214
J48	71,5266	58,9190	72,7492	73,3665	71,8910
k-NN	77,2900	72,0265	77,0446	77,0446	77,1740
NB	70,3071	50,4456	70,4336	70,4321	70,3071
SVM	81,6110	73,6269	81,2436	81,2436	81,6110
SE-product-polarity	77,8034	71,6561	78,9009	78,9009	77,8034
IMBHN ^C	77,4345	67,6152	78,6540	78,6540	77,5565
IMBHN ^R	73,9762	64,2909	73,7248	73,9702	74,0982
J48	71,0479	62,8214	70,8055	70,8055	71,0479
k-NN	75,8220	69,8178	75,8160	75,9410	75,8235
NB	68,9521	52,7778	69,5694	69,4460	68,9521
SVM	77,8034	71,6561	78,9009	78,9009	77,8034
BS-topic	100,0000	100,0000	100,0000	100,0000	100,0000
IMBHN ^C	98,9286	99,6429	100,0000	100,0000	99,6429
IMBHN ^R	99,6429	99,6429	100,0000	100,0000	100,0000
J48	96,8227	97,8695	97,8695	97,8695	97,8695
k-NN	99,6552	98,5837	100,0000	100,0000	100,0000
NB	99,6429	95,0616	99,6429	99,6429	99,6429
SVM	100,0000	100,0000	100,0000	100,0000	100,0000
BS-semantic	68,9532	55,9236	69,2980	68,9409	68,9532
IMBHN ^C	64,6552	51,2931	64,6552	64,6552	64,6552
IMBHN ^R	68,9532	52,7217	69,2980	68,9409	68,9532
J48	59,0517	45,6158	59,3966	59,3966	59,0517
k-NN	65,3818	53,8054	65,7389	65,7389	65,7389
NB	57,6108	48,0788	57,6108	57,6108	57,6108
SVM	63,6576	55,9236	63,6576	63,6576	63,6576
BS-topic-semantic	66,8596	56,3054	66,5271	66,5271	66,8596
IMBHN ^C	62,5739	54,8768	66,4409	66,4409	62,5985
IMBHN ^R	57,9926	47,0567	57,9803	57,9803	57,9926
J48	55,1478	42,0813	55,8744	55,8744	55,1478
k-NN	65,7512	53,4483	66,1084	66,1084	66,4655
NB	57,2537	46,3300	57,2537	57,2537	57,2537
SVM	66,8596	56,3054	66,5271	66,5271	66,8596

Nota – Acurácias maiores que as obtidas com o *baseline* BOW são apresentadas em negrito. A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia de cada representação para o respectivo *dataset*.

BOW em apenas cinco casos, sendo três para o *dataset SE-product* e dois para *BS-topic*. Esses dois *datasets* correspondem ao primeiro nível de complexidade semântica, a organização por tópico, e apresentam acurácia bastante elevadas mesmo com o uso da tradicional BOW.

Nos cenários de maior complexidade semântica (*SE-polarity*, *SE-product-polarity*, *BS-semantic* e *BS-topic-semantic*), os classificadores gerados com a representação *Uni-based* não obtiveram resultados melhores do que a BOW. Já a combinação de classificadores gerados com as duas representações melhorou os resultados da BOW em todos os *datasets*. No entanto, o

aumento da acurácia em relação ao uso da BOW foi verificado com mais frequência nos *datasets* de organização por tópico.

3.4 Considerações finais

Conforme apresentado no capítulo anterior, o modelo de representação adotado para coleções de documentos durante o processo de Mineração de Textos pode ter grande impacto no resultado final do processo. Nesse contexto, neste capítulo foram apresentadas duas novas representações de textos que foram exploradas inicialmente neste trabalho.

A primeira representação, *gBoED*, é construída com base em listas de termos do domínio. Com isso, incorpora-se aspectos semânticos e conhecimento do domínio em uma representação no modelo espaço vetorial. Por serem formados por termos do domínio e termos que são relevantes para as classes dos documentos, os atributos da *gBoED* são mais expressivos do que os atributos da BOW. A segunda representação de documentos proposta e explorada, *Uni-based*, é formada a partir de vetores da abordagem NASARI *unified*. Para gerar a representação *Uni-based*, é realizada uma combinação dos vetores NASARI *unified* para os termos desambiguados presentes nos documentos. Assim, nessa representação são incorporadas as informações de hiperônimos dos termos presentes nos documentos.

Para analisar o uso dessas representações na MT, foram realizadas avaliações experimentais das representações propostas considerando-se a classificação de coleções de documentos com diferentes níveis de complexidade semântica. Os resultados indicam que as representações semânticas, apesar de serem mais expressivas do que a BOW, não apresentam informações suficientes para a identificação de padrões. No entanto, a combinação das representações semânticas com a BOW se mostrou promissora.

O trabalho apresentado neste capítulo corresponde a explorações iniciais da incorporação da semântica na representação de documentos. Nos próximos capítulos são apresentadas propostas de novas representações que incorporam a semântica dos documentos por meio de outras abordagens. No Capítulo 4 a semântica é abordada por meio do uso de informações de papéis semânticos e no Capítulo 5 a semântica é abordada de maneira latente, por meio de *embeddings* de palavras e de sentidos.

REPRESENTAÇÕES BASEADAS EM PAPÉIS SEMÂNTICOS

4.1 Considerações iniciais

A importância da semântica na Mineração de Textos, bem como um mapeamento do tema, foi apresentada no Capítulo 2. Uma das frentes de pesquisa que tem potencial para contribuir para uma representação de textos mais completa é a incorporação de informações enriquecidas por meio de técnicas de Processamento de Língua Natural. Nesse contexto, são encontrados trabalhos utilizando reconhecimento de entidades nomeadas (SINOARA *et al.*, 2014; DOAN *et al.*, 2009), anotação de classes morfossintáticas (SPANAKIS; SIOLAS; STAFYLOPATIS, 2012; BEKKERMAN *et al.*, 2007) e anotação de papéis semânticos (OCHOA *et al.*, 2013; SHEHATA; KARRAY; KAMEL, 2010; PERSSON; JOHANSSON; NUGUES, 2009).

Parte do problema discutido na Seção 2.5.1, em especial no exemplo da Figura 8 (Página 58), poderia ser resolvido com informações sobre os papéis semânticos dos argumentos associados aos verbos das sentenças. Os papéis semânticos podem ser úteis para diferenciar documentos que, apesar de utilizarem o mesmo vocabulário, apresentam visões diferentes sobre um mesmo assunto. No entanto, ainda não há muitos trabalhos explorando informações sobre papéis semânticos para melhorar o resultado da classificação automática de textos.

A anotação de papéis semânticos é uma tarefa de PLN que tem o objetivo de identificar e atribuir rótulos de papéis semânticos a argumentos de predadores (geralmente verbos). Os rótulos de papéis semânticos permitem a identificação do evento descrito em uma sentença e seus participantes. Com isso, questões como “quem fez o que a quem?” podem ser respondidas (PALMER; GILDEA; XUE, 2010). Por exemplo, a versão no idioma inglês das três primeiras sentenças da Figura 8 (Página 58) podem ser anotadas como se segue.

1. [Company Alpha]_{AGENT} [acquired]_{VERB} [Company Beta]_{THEME}.

2. [Company Beta]**AGENT** [acquired]**VERB** [Company Alpha]**THEME**.
3. [Company Beta]**THEME** was [acquired]**VERB** [by Company Alpha]**AGENT**.

Nesse exemplo, as entidades que adquiriram algo são rotuladas como AGENT e as entidades que foram adquiridas são rotuladas como THEME. AGENT e THEME são papéis semânticos, ou mais especificamente papéis temáticos (JURAFSKY; MARTIN, 2009), utilizados no recurso léxico VerbNet¹. Apesar de não haver uma definição consensual de uma lista de papéis semânticos na comunidade de pesquisa, a lista de papéis semânticos adotadas no PropBank² (PALMER; GILDEA; KINGSBURY, 2005) tem sido frequentemente utilizada na comunidade de PLN.

O PropBank é um recurso linguístico-computacional, para o idioma inglês, que apresenta anotações de papéis semânticos para o corpus de árvores sintáticas Penn Treebank. Nas definições do PropBank, cada verbo tem um conjunto definido de papéis semânticos, que são argumentos identificados por números (de Arg0 a Arg5) e modificadores de verbos (ArgM), como por exemplo informação sobre local (ArgM-LOC) e tempo (ArgM-TMP). Esse esquema de rótulos baseados em argumentos numerados do PropBank simplifica a lista de rótulos, perdendo expressividade. No entanto, essa simplicidade facilita a tarefa de predição de rótulos por meio de Aprendizado de Máquina, sendo adequado para a automação da tarefa de anotação de papéis semânticos. Por esse fato, essa lista tem sido adotada em sistemas de anotação de papéis semânticos (COLLOBERT *et al.*, 2011; FONSECA; ROSA, 2013b; ALVA-MANCHEGO, 2013).

Como exemplo, o predicador “*acquire*”, usado no exemplo anterior, pode ter os seguintes papéis semânticos, segundo o PropBank³:

- Arg0: *agent, entity acquiring something* (correspondente ao papel AGENT na VerbNet)
- Arg1: *thing acquired* (correspondente ao papel THEME na VerbNet)
- Arg2: *seller* (correspondente ao papel SOURCE na VerbNet)
- Arg3: *price paid*
- Arg4: *benefactive*

Para o português, existe o PropBank.Br⁴, que foi criado seguindo as diretrizes do projeto PropBank, porém com algumas adequações necessárias devido a peculiaridades do idioma português (DURAN; ALUÍSIO, 2012). O PropBank.Br apresenta anotações em textos escritos em português do Brasil sobre diversos assuntos, extraídos do jornal Folha de São Paulo. Em

¹ VerbNet: <<http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>>

² PropBank: <<http://verbs.colorado.edu/~mpalmer/projects/ace.html>>

³ Predicador *acquire*: <<http://verbs.colorado.edu/propbank/framesets-english-aliases/acquire.html>>

⁴ PropBank.Br: <<http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/propbankbr>>

comparação com o PropBank, o recurso do português tem um volume bem reduzido. Segundo [Fonseca e Rosa \(2013b\)](#), o PropBank.Br possui um sétimo do tamanho do PropBank do idioma inglês. Apesar de reduzido, o PropBank.Br é um recurso linguístico computacional importante, possibilitando o treinamento de modelos para anotação automática de papéis semânticos para idioma o português. Vale também notar que esse recurso vem sendo evoluído. Atualmente, estão disponíveis versões atualizadas do PropBank.Br. A versão 1.1⁵ apresenta 5931 instâncias anotadas para 3348 sentenças, e a versão 2⁶ apresenta 7661 instâncias anotadas para 7442 sentenças ([HARTMANN; DURAN; ALUÍSIO, 2016](#)).

Motivado pela riqueza de informações fornecidas por papéis semânticos e visando a melhoria dos resultados da Mineração de Textos, neste trabalho, a utilização de anotações de papéis semânticos foi explorada e avaliada na classificação automática de documentos. Assim, na próxima seção, é apresentada uma revisão dos trabalhos relacionados ao uso de papéis semânticos na classificação de textos. Em seguida, os modelos de representação propostos neste trabalho são descritos e analisados, e a avaliação experimental é apresentada.

4.2 Trabalhos relacionados

Anotações de papéis semânticos ainda são pouco exploradas na representação de documentos na Mineração de Textos. Existem alguns trabalhos em tarefas como construção de ontologias ([OCHOA *et al.*, 2013](#)), sumarização ([KHAN; SALIM; KUMAR, 2015](#)), identificação de eventos ([LLORENS; SAQUETE; NAVARRO-COLORADO, 2013](#)) e classificação de textos ([SHEHATA; KARRAY; KAMEL, 2010](#); [PERSSON; JOHANSSON; NUGUES, 2009](#)).

Para classificação de textos, [Shehata, Karray e Kamel \(2010\)](#) propõem a representação de textos no modelo espaço-vetorial baseada em estruturas de predicador (verbo) e argumentos, extraídas de cada sentença por meio de um anotador de papéis semânticos. Dependendo da construção das sentenças, uma ou mais estruturas de predicador e argumentos podem ser identificadas. Assim, em uma única sentença, um argumento pode ser anotado com diferentes rótulos de papéis semânticos, sendo um para cada estrutura. Os argumentos anotados, que podem ser palavras ou expressões, são chamados pelos autores de conceitos e são considerados candidatos a atributos. Os autores propõem o uso de quatro esquemas para calcular os pesos de cada conceito nos níveis de sentença, documento e corpus: (i) $weight_{ctf}$, é calculado com base na frequência do conceito nas estruturas de predicador e argumentos dos documentos e representa o peso do conceito no nível de sentença; (ii) $weight_{tf}$, é a frequência do conceito no documento (como a tradicional *term frequency*) e representa o peso do conceito no nível do documento; (iii) $weight_{stat}$, é a soma dos dois pesos anteriores; e (iv) $weight_{combined}$, representa a importância do conceito nos níveis de sentença, documento e corpus por meio da combinação de $weight_{ctf}$, $weight_{tf}$ e a frequência de documento do conceito (df , que corresponde ao número de documentos em que o conceito

⁵ PropBank.Br - Download: <http://nilc.icmc.usp.br/portlex/index.php/en/downloadsingl>

⁶ ProBank.Br versão 2: <http://nilc.icmc.usp.br/semanticlp/propbankbr/pbbr-v2.html>.

aparece). Os conceitos com os maiores pesos, em cada caso, são utilizados para representar os documentos. Shehata, Karray e Kamel (2010) avaliaram a performance de classificação dos esquemas de pesos propostos para construir quatro representações independentes no modelo espaço-vetorial. Os resultados reportados indicam que o esquema *weight_{combined}* é mais adequado para as configurações testadas.

A proposta de Shehata, Karray e Kamel (2010) utiliza apenas as estruturas de predador e argumentos e desconsidera os rótulos dos papéis semânticos. Tais rótulos são utilizados na proposta de Persson, Johansson e Nugues (2009), cujas representações no modelo espaço-vetorial também foram avaliadas na classificação de textos. Os autores propõem cinco conjuntos de atributos diferentes, sendo quatro deles baseados em predadores e seus papéis semânticos mais frequentes (Arg0 e Arg1 do PropBank), conforme apresentado a seguir.

1. *VSO triples*: atributos são triplas verbo-sujeito-objeto identificadas nos documentos.
2. *Predicates*: atributos são predadores desambiguados.
3. *Argument 0*: atributos são concatenação do predador desambiguado e seu Arg0.
4. *Argument 1*: atributos são concatenação do predador desambiguado e seu Arg1.
5. *Arguments 0 e 1*: atributos são concatenação do predador desambiguado e seus dois argumentos, Arg0 e Arg1.

Na avaliação experimental, Persson, Johansson e Nugues (2009) fazem combinações desses conjuntos de atributos e da tradicional *bag-of-words*, avaliando assim 64 representações diferentes para os documentos. A combinação entre dois ou mais conjuntos de atributos é feita por meio da união dos conjuntos. Por exemplo, uma das representações testadas corresponde à união dos atributos dos conjuntos BOW, *Predicates* e *Argument 1*. Nessa representação combinada, todos os atributos dos três conjuntos são utilizados como dimensões dos vetores que representam os documentos. De acordo com a avaliação apresentada, os melhores resultados foram obtidos por representações que incluem os atributos dos conjuntos BOW, *Predicates* e um ou mais dos demais conjuntos. Entre as representações que consideram apenas os atributos baseados em anotação de papéis semânticos (desconsiderando, portanto, o conjunto *VSO triples*), a representação que apresentou a melhor performance de classificação foi a formada por BOW, *Predicates* e *Argument 1*.

Os resultados desses trabalhos sugerem que o uso de rótulos de papéis semânticos na representação de documentos pode melhorar a performance de classificação automática de textos. As representações propostas neste trabalho, e descritas na próxima seção, levam em consideração os rótulos de papéis semânticos. No entanto, apresentam um processo de construção mais simples que as representações propostas por Persson, Johansson e Nugues (2009), visto que não é necessária a desambiguação dos termos.

4.3 Modelos propostos para representação de coleções de documentos: *SR-based Terms* e *SR-based Sentences*

Visando aproveitar a informação semântica fornecida pelos rótulos de papéis semânticos, foram propostos dois modelos de representação: *SR-based Terms* e *SR-based Sentences*. Os atributos dessas representações combinam as palavras com os respectivos papéis semânticos. Tais representações mantêm a simplicidade do modelo espaço-vetorial e podem ser utilizadas como entrada para os algoritmos tradicionais de Aprendizado de Máquina. De maneira geral, a construção das representações é realizada em quatro etapas: (i) Identificação das classes morfofossintáticas e dos papéis semânticos nos documentos; (ii) Seleção dos *tokens* de acordo com as classes morfofossintáticas e normalização; (iii) Construção do conjunto de atributos; e (iv) Construção da representação da coleção de documentos. Essas etapas são descritas a seguir.

1. Identificação das classes morfofossintáticas e dos papéis semânticos nos documentos.

A primeira etapa consiste no processamento dos documentos para identificação e anotação de classes morfofossintáticas e papéis semânticos. Com os documentos anotados, tem-se as estruturas de predador e argumentos de cada documento. Os atributos das representações propostas são construídos com base nas estruturas de predador e argumentos identificadas por um rotulador de papéis semânticos. Vale ressaltar que, uma mesma sentença pode apresentar mais de um estrutura de predador e argumentos. Por exemplo, a sentença “*It is rare that hotels in this class serve hot meals, yet they do!*”⁷, apresenta três estruturas de predador e argumentos⁸, listadas na [Figura 16](#). Nessa figura, os rótulos de papéis semânticos estão subscritos e as classes morfofossintáticas estão sobrescritas. A descrição das etiquetas utilizadas na anotação morfofossintática é apresentada no [Quadro 3](#). Essas etiquetas pertencem ao conjunto de etiquetas do Penn Treebank, para textos em inglês.

2. Seleção dos *tokens* de acordo com as classes morfofossintáticas e normalização.

Para cada estrutura de predador e argumentos identificada, os *tokens*, normalmente palavras, são selecionados de acordo com suas classes morfofossintáticas. O objetivo desse passo é remover *tokens* irrelevantes, como preposições e pontuações. Essa seleção pode ser vista como uma remoção de *stopwords*, tarefa comumente realizada na etapa de Pré-processamento do processo de Mineração de Textos. A ideia é que verbos e substantivos são geralmente as classes de palavras mais relevantes para a Mineração de Textos. No entanto, as classes morfofossintáticas de interesse podem variar de acordo com a coleção de textos e o domínio de aplicação. Para análise de sentimentos, por exemplo, os adjetivos também podem ser interessantes, visto que são uma forma direta de expressar opiniões.

⁷ Sentença extraída de um dos documentos da coleção *SE-ABSA15*, descrita na Seção 2.5.2 (Página 64).

⁸ A anotação apresentada foi realizada com a ferramenta SENNA (COLLOBERT *et al.*, 2011).

Figura 16 – Estruturas de predador e argumentos identificadas na sentença “*It is rare that hotels in this class serve hot meals, yet they do!*”.

1. *It* [*is*^{VBZ}]_V [*rare*^{JJ}]_{A2} *that* [*hotels*^{NNS} *in*^{IN} *this*^{DT} *class*^{NN} *serve*^{VB} *hot*^{JJ} *meals*^{NNS}]_{A1}, *yet they do!*
2. *It is rare that* [*hotels*^{NNS} *in*^{IN} *this*^{DT} *class*^{NN}]_{A0} [*serve*^{VB}]_V [*hot*^{JJ} *meals*^{NNS}]_{A1}, *yet they do!*
3. *It is rare that hotels in this class serve hot meals, yet* [*they*^{PRP}]_{A0} [*do*^{VB}]_{V!}

Fonte: Elaborada pelo autor.

Quadro 3 – Conjunto de etiquetas morfossintáticas do Penn Treebank.

Etiqueta	Descrição
DT	determinante (<i>determiner</i>)
IN	preposição ou conjunção (<i>preposition / subordinating conjunction</i>)
JJ	adjetivo (<i>adjective</i>)
NN	nome, no singular (<i>noun, singular or mass</i>)
NNP	nome próprio, no singular (<i>proper noun, singular</i>)
NNS	nome, no plural (<i>noun, plural</i>)
PRP	pronome pessoal (<i>personal pronoun</i>)
VB	verbo, no infinitivo (<i>verb, base form</i>)
VBD	verbo, no passado (<i>verb, past tense</i>)
VBN	verbo, no particípio passado (<i>verb, past participle</i>)
VBP	verbo, no presente, exceto terceira pessoa do singular (<i>verb, non-3rd ps. sing. present</i>)
VBZ	verbo, no presente, terceira pessoa do singular (<i>verb, non-3rd ps. sing. present</i>)

Fonte: Adaptada de [Marcus, Marcinkiewicz e Santorini \(1993\)](#).

Com o objetivo de reduzir a dimensionalidade das representações por meio do agrupamento de variações das palavras, os *tokens* são normalizados, sendo substituídos pelo radical das palavras. O radical dos *tokens* selecionados são chamados de termos e irão compor os atributos das representações propostas.

3. Construção do conjunto de atributos.

É nessa etapa de construção de atributos que as duas representações propostas, *SR-based Terms* e *SR-based Sentences*, se diferenciam. Cada representação tem uma composição diferente para os atributos. Os atributos de *SR-based Terms* são formados por um termo

e seu rótulo de papel semântico. Por exemplo, para a sentença “*It is rare that hotels in this class serve hot meals, yet they do!*”, cujas estruturas de predicador e argumentos são apresentadas na [Figura 16](#), são gerados os seguintes 14 atributos⁹ para a representação *SR-based Terms*:

- [V]_is
- [A2]_rare
- [A1]_hotels
- [A1]_class
- [A1]_serve
- [A1]_hot
- [A1]_meals
- [A0]_hotels
- [A0]_class
- [V]_serve
- [A1]_hot
- [A1]_meals
- [A0]_they
- [V]_do

Já na representação *SR-based Sentences*, mantém-se a estrutura de predicador e argumentos. Assim, cada atributo dessa representação é formado por todos os termos de uma estrutura e seus respectivos papéis semânticos. Para a mesma sentença da [Figura 16](#), são gerados os seguintes três atributos para a representação *SR-based Terms*:

- [V]_is_[A2]_rare_[A1]_hotels_class_serve_hot_meals
- [V]_serve_[A0]_hotels_class_[A1]_hot_meals
- [V]_do_[A0]_they

4. Construção da representação da coleção de documentos.

Para a construção das representações da coleção de documentos, o peso de cada atributo é calculado para cada documento da coleção. Assim como na construção da representação *bag-of-words*, esse cálculo pode ser realizado por qualquer medida. Neste trabalho, optou-se por utilizar a medida TF, visto os bons resultados que apresentou com a representação

⁹ Atributos gerados sem normalização dos termos e considerando-se apenas as classes morfosintáticas referentes a verbos, nomes, pronomes e adjetivos.

bag-of-words no trabalho de Rossi *et al.* (2014). Os autores compararam as medidas TF e TF-IDF utilizando diferentes coleções de documentos e algoritmos de aprendizado supervisionado, e concluíram, com base nos experimentos realizados, que a medida TF é a mais adequada para a classificação de textos quando se realiza a remoção de *stopwords*.

Para ilustrar as representações propostas, considere os documentos da Figura 17. As estruturas de predicador e argumentos dos documentos *D1*, *D2* e *D3*, com anotações de classes morfofossintáticas e de papéis semânticos, são apresentadas na Figura 18¹⁰. Nessa figura, os rótulos de papéis semânticos estão subscritos e as classes morfofossintáticas, cujas descrições das etiquetas são apresentadas no Quadro 3, estão sobrescritas.

Figura 17 – Exemplo 4: Documentos ilustrativos

D1 Company Alpha acquired Company Beta.
D2 Company Beta acquired Company Alpha.
D3 Company Beta was acquired by Company Alpha.

Fonte: Elaborada pelo autor.

Figura 18 – Exemplo 4: Documentos da Figura 17 com anotações de classes morfofossintáticas e papéis semânticos.

D1 [Company^{NNP} Alpha^{NNP}]_{A0} [acquired^{VBD}]_V [Company^{NNP} Beta^{NNP}]_{A1}.
D2 [Company^{NNP} Beta^{NNP}]_{A0} [acquired^{VBD}]_V [Company^{NNP} Alpha^{NNP}]_{A1}.
D3 [Company^{NNP} Beta^{NNP}]_{A1} was^{VBD} [acquired^{VBN}]_V [by^{IN} Company^{NNP} Alpha^{NNP}]_{A0}.

Fonte: Elaborada pelo autor.

Considerando a seleção das classes morfofossintáticas referentes a verbos e nomes, as representações *bag-of-words*, *SR-based Terms* e *SR-based Sentences* para os documentos apresentados na Figura 17 são apresentadas na Figura 19¹¹. Pode-se verificar que, na representação BOW (Figura 19a), os documentos *D1* e *D2*, que apresentam ideias opostas, são representados pelo mesmo vetor. Já os documentos *D1* e *D3*, que possuem o mesmo significado, são representados por vetores diferentes. Tal problema não ocorre ao se utilizar os modelos propostos (Figuras 19b

¹⁰ A anotação apresentada foi realizada com a ferramenta SENNA (COLLOBERT *et al.*, 2011).

¹¹ Os termos foram radicalizados pelo algoritmo de Porter: <<http://tartarus.org/~martin/PorterStemmer/>>

e 19c). Nesses modelos, o vetor que representa $D1$ é igual ao vetor de $D3$ e diferente do vetor de $D2$.

Figura 19 – Representações dos documentos $D1$, $D2$ e $D3$ da Figura 17

	compani	alpha	acquir	beta	wa
$D1$	2	1	1	1	0
$D2$	2	1	1	1	0
$D3$	2	1	1	1	1

(a) *bag-of-words*

	[A0]_ compani	[A1]_ compani	[A0]_ alpha	[A1]_ alpha	[V]_ acquir	[A0]_ beta	[A1]_ beta
$D1$	1	1	1	0	1	0	1
$D2$	1	1	0	1	1	1	0
$D3$	1	1	1	0	1	0	1

(b) *SR-based Terms*

	[V]_acquir -[A0]_compani_alpha -[A1]_compani_beta	[V]_acquir -[A0]_compani_beta -[A1]_compani_alpha
$D1$	1	0
$D2$	0	1
$D3$	1	0

(c) *SR-based Sentences*

Fonte: Elaborada pelo autor.

4.4 Análise das representações *SR-based Terms* e *SR-based Sentences*

Conforme pode ser verificado no exemplo da Figura 19, os atributos das representações *SR-based Terms* e *SR-based Sentences* são mais ricos em informação do que os atributos da *bag-of-words*. Enquanto os atributos da BOW são apenas termos, sendo palavras normalizadas ou não, os atributos das representações propostas apresentam informações sobre o papel semântico desses termos nas sentenças. Por exemplo, o termo “compani”, que corresponde a um único atributo na BOW (Figura 19a), corresponde a dois termos, “[A0]_compani” e “[A1]_compani”, na representação *SR-based Terms* (Figura 19b). Nesse caso, o termo “compani” aparece na coleção de documentos desempenhando dois papéis semânticos distintos, Arg0 e Arg1, para os diferentes predicadores. No caso da representação *SR-based Sentences*, cada atributo corresponde a uma estrutura de predicador e argumentos. Assim, os atributos dessa representação são ainda mais ricos, contendo um predicador ligado a seus argumentos.

Pode-se perceber que os atributos das representações propostas, por serem mais informativos, são também mais específicos do que os atributos da BOW. Como consequência, esses

atributos são menos frequentes na coleção de documentos, gerando representações mais esparsas. Um único termo da BOW pode dar origem a mais de um termo na representação *SR-based Terms*, visto que o termo pode aparecer com diferentes papéis semânticos. Assim, a dimensionalidade de *SR-based Terms* tende a ser maior do que a dimensionalidade da BOW. Já no caso de *SR-based Sentences*, os atributos são formados por todos os termos da estrutura de predador e argumentos, unindo, portanto, vários termos em um único atributo. O resultado é uma dimensionalidade maior do que a da BOW e menor do que a da representação *SR-based Terms*. Para ilustrar esses fatos, na Tabela 3 são apresentadas as características dos conjuntos de atributos das representações propostas e da representação BOW, em duas coleções de documentos descritas na Seção 2.5.2 (Página 64).

Tabela 3 – Características dos conjuntos de atributos das representações propostas

Coleção de documentos	Representação	Número de atributos	Esparsidade
<i>BS-Top4</i>	BOW	1.312	94,68%
	<i>SR-based Terms</i>	5.615	95,58%
	<i>SR-based Sentences</i>	4.548	99,93%
<i>SE-ABSA15</i>	BOW	1.880	98,80%
	<i>SR-based Terms</i>	12.064	99,54%
	<i>SR-based Sentences</i>	10.820	99,92%

Conforme apresentado, as representações propostas são representações enriquecidas com informações de papéis semânticos, apresentando atributos mais expressivos do que a BOW. No entanto, isso acarreta algumas limitações. Como essas informações são obtidas por meio de sistemas de anotação de papéis semânticos, deve-se notar que a cobertura e qualidade desses sistemas podem afetar a qualidade das representações propostas. Erros desses sistemas são propagados para a representação da coleção de documentos. Além disso, devido à própria natureza dos papéis semânticos, parte do conteúdo dos documentos pode não ser representado. Caso o sistema de anotação de papéis semânticos não atribua rótulos aos termos de uma sentença, ela será omitida nas representações propostas. Os seguintes fatores podem levar a essa situação: (i) textos informais e/ou de baixa qualidade; (ii) limitação dos sistemas a predadores verbais; e (iii) frases sem verbos.

Textos informais ou de baixa qualidade podem limitar a criação de atributos nas representações propostas, visto que problemas como gírias e erros gramaticais podem prejudicar o desempenho do sistema de anotação de papéis semânticos. Em relação ao segundo fator, a limitação dos sistemas a predadores verbais, tem-se que o PropBank originalmente só apresentava anotação de papéis semânticos de verbos e o mesmo ocorre com o PropBank.Br (BONIAL *et al.*, 2014; ALVA-MANCHEGO, 2013; FONSECA, 2013). Assim, os sistemas treinados em corpus que apresentam apenas anotação de verbos, limitam-se a esse tipo de predador, não reconhecendo os predadores nominais. O terceiro fator corresponde à frases curtas, formada por uma ou poucas palavras e sem um verbo. Frases podem ser compostas por apenas uma palavra e ainda assim podem ter um significado para o ouvinte ou leitor, como, por exemplo, a frase “Cuidado!”.

Esses fatores ocorrem com mais frequência em textos informais e/ou textos de opinião. Nos textos de opinião, como é o caso da coleção *SE-ABSA15*, o autor pode expressar sua opinião em frases curtas, formada por uma ou poucas palavras e sem um verbo, podendo até mesmo usar apenas adjetivos. Por exemplo, a opinião pode ser expressa usando “bom”, “limpo e confortável” e “péssimo atendimento”. Como nessas frases não há um verbo, elas não terão rótulos de papéis semânticos atribuídos e, consequentemente, não serão incluídas nas representações *SR-based Terms* e *SR-based Sentences*. Assim, essas representações propostas, apesar de serem mais expressivas, podem representar um conteúdo mais limitado do que o representado com a BOW.

4.5 Avaliação experimental - representações baseadas em papéis semânticos

Nessa seção é apresentada a avaliação experimental realizada com os objetivos de avaliar a performance de classificação em relação a diferentes níveis de complexidade semântica e o impacto do uso de rótulos de papéis semânticos na representação de documentos. Com essa avaliação, discute-se as seguintes perguntas: (i) A representação BOW é adequada a qual tipo de problema de classificação?; (ii) Representações enriquecidas com informações de papéis semânticos podem levar a melhores resultados de classificação?; e (iii) A combinação de representações semânticas com a BOW pode levar a melhores resultados de classificação? Se sim, qual é a melhor maneira de combinar as representações?.

4.5.1 Datasets, representações de textos e configuração experimental

Os experimentos foram realizados utilizando-se os *datasets* das coleções *BEST sports Top 4* e *SE-ABSA15*, apresentados na Seção 2.5.2 (Página 64). Os modelos de representação baseados em papéis semânticos, propostos e apresentados na Seção 4.3, foram avaliados em relação a três outros modelos da literatura. Assim, para cada *dataset*, foram geradas cinco representações no modelo espaço-vetorial, apresentadas a seguir.

- *SR-based Terms*;
- *SR-based Sentences*;
- BOW (*baseline*);
- pred.Arg1, representação proposta por Persson, Johansson e Nugues (2009) (trabalho descrito na Seção 4.2) com atributos formados pelos conjuntos *Predicates* e *Argument 1*;

- BOW+pred.Arg1, representação proposta por Persson, Johansson e Nugues (2009), que é a união do conjunto de atributos anterior (pred.Arg1) com o conjunto de atributos da *bag-of-words*.

Os experimentos foram realizados utilizando variações de parâmetros de sete algoritmos de AM. Também foi avaliada a combinação de classificadores, utilizando três estratégias para compor os votos entre os diferentes classificadores base: *Sum of Confidences* (SC), *Weighted Sum of Confidences* (WSC) e *Most Confident* (MC). Detalhes sobre a geração das representações, bem como sobre a configuração experimental adotada, são apresentados no Apêndice B.

4.5.2 Resultados - papéis semânticos

A execução da configuração experimental adotada resultou em 105 classificadores para cada representação avaliada. A distribuição dos valores de acurácia em cada conjunto de dados é apresentada na Figura 20. De modo geral, a representação BOW apresentou as melhores acurácias e a representação *SR-based Terms* apresentou valores próximos à BOW. Já o uso das representações pred.Arg1 e *SR-based Sentences* levaram a valores de acurácia muito abaixo do *baseline*. As medianas dos resultados da representação pred.Arg1 foram maiores do que as medianas da representação *SR-based Sentences* e menores do que as medianas da representação *SR-based Terms*. A única exceção foi o caso do *dataset SE-polarity*, para o qual as três representações (pred.Arg1, *SR-based Terms* e *SR-based Sentences*) apresentaram o mesmo valor de mediana (66,75%). Outro ponto a ser notado é que, no geral, a representação *SR-based Sentences* apresentou menor variação nos resultados ao se variar as configurações de aprendizado (variações de algoritmos e seus parâmetros), porém com acurácias baixas quando comparada às outras representações.

A primeira linha de *box plots* (Figuras 20a e 20b) corresponde aos *datasets* do primeiro nível de complexidade semântica, a organização por tópico. Esses *datasets* apresentam as maiores acurácias de classificação, com valores máximos próximos a 100%. Para o *dataset BS-topic* a representação BOW apresenta ótimos resultados, com pequena variação de acurácia ao se variar as configurações de aprendizado. Para esse caso, a BOW apresenta baixo intervalo interquartil (diferença entre o terceiro quartil e o primeiro quartil) e valor médio de acurácia de 89,34%. Já para o conjunto *SE-product*, apesar de apresentar valor médio de acurácia de 76,86%, a BOW apresenta grande variação entre as diferentes configurações de aprendizado, sendo que o primeiro quartil é 53,73%.

A segunda linha de *box plots* (Figuras 20c e 20d) corresponde aos *datasets* do segundo nível de complexidade semântica (organização semântica) da coleção de documentos em inglês. Já os *datasets* de organização semântica da coleção de documentos em português são apresentados na terceira linha (Figuras 20e e 20f). Pode-se perceber que os problemas de organização semântica apresentam acurácias mais baixas do que os problemas de organização por tópico. Na

organização semântica, os valores máximos de acurácia da coleção em inglês ficam em torno de 80% e os valores máximos da coleção em português ficam em torno de 60%.

Figura 20 – *Box plots* das acurácias de classificação, por representação utilizada, para cada *dataset*.

(a) *SE-product*

(b) *BS-topic*

(c) *SE-polarity*

(d) *SE-product-polarity*

(e) *BS-semantic*

(f) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Conforme apresentado na Seção 2.2.2 (Página 44), a medida acurácia corresponde à abordagem *micro-averaging* para sumarização das medidas precisão, revocação e F1 em

problemas multiclasse. Nesses experimentos, os resultados de *micro-averaging* apresentam correlação positiva com os resultados obtidos com a abordagem *macro-averaging*. A correlação de Pearson entre essas abordagens, para as representações e combinações de representações avaliadas, é apresentada na Tabela 4. Pode-se verificar que, com exceção das representações *pred.Arg1* e *SR-based Sentences* que apresentam as piores performances de classificação, as medidas $Precisão^M$, $Revocação^M$ e $F1^M$ apresentam uma boa correlação com a acurácia.

Tabela 4 – Correlação das medidas de performance de classificação utilizando as abordagens *micro-averaging* (acurácia) e *macro-averaging*.

Representação	Correlação de Pearson		
	Acurácia x $Precisão^M$	Acurácia x $Revocação^M$	Acurácia x $F1^M$
BOW	0,8866	0,8817	0,8985
<i>pred.Arg1</i>	0,7685	0,7642	0,7850
<i>SR-based Terms</i>	0,8805	0,8768	0,8916
<i>SR-based Sentences</i>	0,6112	0,7488	0,7171
BOW+ <i>pred.Arg1</i>	0,8763	0,8729	0,8890
BOW + <i>SR-based Terms</i> - SC	0,8859	0,8763	0,8948
BOW + <i>SR-based Terms</i> - WSC	0,8862	0,8761	0,8949
BOW + <i>SR-based Terms</i> - MC	0,8868	0,8813	0,8973

A Tabela 5 apresenta a melhor acurácia obtida por cada algoritmo de aprendizado entre todos os parâmetros testados. A melhor acurácia foi avaliada com o objetivo de considerar o melhor caso de cada estratégia de aprendizado. No restante desta seção, os resultados obtidos com a medida acurácia são utilizados para discutir as perguntas estabelecidas no início da seção.

A representação BOW é adequada a qual tipo de problema de classificação?

Os resultados da avaliação experimental indicam que quando o objetivo de classificação é a organização por tópico (primeiro nível de complexidade semântica), a BOW é uma representação bastante adequada. Esse é o caso dos *datasets BS-topic* e *SE-product*, para os quais a acurácia de classificação obtida com o uso da BOW é próxima a 100%. Conforme discutido na Seção 2.5 (Página 58), esses são os cenários que normalmente são tratados na comunidade de pesquisa. Nesses cenários, as palavras têm grande importância na diferenciação das classes. Nos casos dessa avaliação experimental, há palavras específicas para descrever os resultados dos diferentes esportes do conjunto *BS-topic* ou as características dos diferentes produtos do conjunto *SE-product*.

No entanto, quando o objetivo de classificação é diferente da organização por tópico, a acurácia obtida com o uso da BOW é mais baixa. Esse é o caso dos *datasets BS-semantic*, *BS-topic-semantic*, *SE-polarity* e *SE-product-polarity*. Para esses *datasets*, apenas as palavras não são suficientes para se obter uma boa performance de classificação.

Tabela 5 – Melhores acurácias para as coleções *BS-Top4* e *SE-ABSA15*

	BOW	pred.Arg1	SR-based Terms	SR-based Sentences	BOW+SR-based Terms			BOW+ pred.Arg1
					SC	WSC	MC	
SE-product	99,5077	76,4363	96,6877	63,9265	99,6311	99,6311	99,5077	99,0139
IMBHN ^C	98,1587	70,6700	94,3541	44,0515	98,5276	98,5276	98,5291	97,2960
IMBHN ^R	99,1388	72,3893	93,6149	48,8392	99,0169	99,0169	99,1388	98,7715
J48	92,2704	59,0319	87,0009	54,3406	92,2704	92,2704	92,2704	91,9091
k-NN	98,5245	73,0081	80,8582	63,9265	98,7684	98,7684	98,5245	98,3996
MNB	99,5077	76,4363	96,6877	39,4023	99,6311	99,6311	99,5077	99,0139
NB	92,7612	63,0653	84,4143	62,8245	93,3755	93,0066	92,7612	92,0235
SVM	96,4408	73,9822	91,0449	59,8796	96,3189	96,3189	96,4408	95,4577
SE-polarity	84,5438	71,8970	79,0289	69,3180	84,7877	84,6658	84,5438	83,9235
IMBHN ^C	80,4908	61,5854	74,7320	47,2343	81,1066	81,1126	80,4908	78,7669
IMBHN ^R	82,8214	61,9572	75,5872	50,5571	82,9449	82,9449	82,8214	81,5944
J48	71,5266	66,3791	65,6399	68,7067	71,6456	71,6456	71,6456	73,4929
k-NN	77,2900	71,8970	72,1575	69,3180	78,0337	78,0337	77,2930	77,2930
MNB	84,5438	69,4444	79,0289	36,4438	84,7877	84,6658	84,5438	83,9235
NB	70,3071	70,3086	62,3336	66,9904	70,3086	70,8025	70,3071	68,9672
SVM	81,6110	70,3086	75,0888	68,0955	81,2436	81,2436	81,6110	79,8811
SE-product-polarity	83,6811	57,7853	76,8112	43,0608	84,0515	84,0515	83,6811	82,1966
IMBHN ^C	77,4345	45,7603	70,8010	27,4947	78,4026	78,5260	78,7775	77,0461
IMBHN ^R	73,9762	49,9368	64,9036	32,2779	73,4854	73,4809	73,9762	74,0952
J48	71,0479	47,3607	60,1250	37,4195	70,9259	70,9259	71,0479	69,9458
k-NN	75,8220	51,2888	57,8064	42,8365	76,5522	76,4303	75,8220	75,3222
MNB	83,6811	57,7853	76,8112	38,1542	84,0515	84,0515	83,6811	82,1966
NB	68,9521	47,6122	55,9545	41,4755	69,4460	69,5679	68,9521	69,0756
SVM	77,8034	53,2475	69,5664	43,0608	77,6814	77,6814	77,8034	75,9515
BS-topic	100,0000	76,6133	100,0000	48,4236	100,0000	100,0000	100,0000	100,0000
IMBHN ^C	98,9286	76,2562	98,9286	47,7340	99,2857	99,2857	99,2857	99,6429
IMBHN ^R	99,6429	73,7931	100,0000	48,4236	100,0000	100,0000	100,0000	100,0000
J48	96,8227	67,8695	94,0025	32,1798	97,8818	97,5369	97,5369	95,7389
k-NN	99,6552	70,2463	99,2980	48,4236	99,6552	99,6552	100,0000	99,6552
MNB	100,0000	76,6133	99,6429	38,5222	100,0000	100,0000	100,0000	100,0000
NB	99,6429	71,9951	98,2143	41,6872	99,6429	99,6429	99,6429	99,6552
SVM	100,0000	73,1281	99,2857	40,6404	100,0000	100,0000	100,0000	100,0000
BS-semantic	68,9532	53,0788	65,4064	39,5567	68,9409	68,5961	68,9532	68,9532
IMBHN ^C	64,6552	46,6502	62,2906	36,3793	67,1798	67,1798	64,7044	65,6773
IMBHN ^R	68,9532	43,4360	65,4064	37,8079	68,9409	68,5961	68,9532	68,9532
J48	59,0517	45,1601	49,9015	28,2882	59,3966	59,3966	59,0517	51,9704
k-NN	65,3818	53,0788	61,9335	39,5567	67,5246	67,1552	66,0961	64,6921
MNB	59,7414	49,1133	57,9926	24,0025	60,8251	60,0985	60,8251	58,6823
NB	57,6108	46,9951	54,8153	32,4877	57,6108	57,6108	57,6108	58,3005
SVM	63,6576	49,4581	59,0887	31,4655	64,7167	64,7167	63,6576	67,1675
BS-topic-semantic	66,8596	41,3424	63,3128	26,1453	68,2389	68,2389	66,8596	64,6921
IMBHN ^C	62,5739	40,2463	63,3128	23,3005	68,2389	68,2389	66,4409	63,2759
IMBHN ^R	57,9926	40,9729	56,9212	20,8498	58,7069	59,0517	58,0049	58,3374
J48	55,1478	31,4778	53,7315	18,3867	58,3128	57,2414	56,1823	57,9557
k-NN	65,7512	41,3424	61,5517	26,1453	67,1675	67,5246	66,4655	64,6921
MNB	62,5985	40,6527	54,8768	15,5542	64,0148	63,6576	63,6700	58,6946
NB	57,2537	38,8670	55,4803	16,2808	58,9901	57,2537	57,2537	56,1946
SVM	66,8596	35,6773	59,0764	17,3399	66,8842	67,2167	66,8596	64,3842

Nota – Acurácias maiores que as obtidas com o *baseline* BOW são apresentadas em negrito. A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia de cada representação para o respectivo *dataset*.

Representações enriquecidas com informações de papéis semânticos podem levar a melhores resultados de classificação?

As representações avaliadas neste trabalho são mais ricas do que a BOW em relação à expressividade, porém, de modo geral, os classificadores gerados utilizando apenas essas

representações possuem performance de classificação inferior ao dos classificadores gerados com a BOW. Considerando todas as configurações testadas, a representação BOW se apresenta como a primeira no *ranking* das representações. O teste de Friedman $N \times N$ com pós-teste de Nemenyi, rejeita a hipótese nula (hipótese de que não há diferença estatisticamente significativa entre as representações, com nível de confiança de 95% e $p\text{-value} < 2.2 \times 10^{-16}$). O diagrama de diferença crítica dessa análise é apresentado na Figura 21.

Figura 21 – Diagrama de diferença crítica entre as representações avaliadas

Fonte: Elaborada pelo autor.

Nota – *Ranking* médio das representações: BOW = 1,3127, *SR-based Terms* = 2,1722, pred.Arg1 = 2,9389 e *SR-based Sentences* = 3,5762. Diferença crítica, segundo pós-teste de Nemenyi: CD = 0,1870

Conforme apresentado na Seção 4.4, a incorporação das informações sobre papéis semânticos leva a um aumento de número de atributos em comparação com a BOW. As representações propostas na Seção 4.3 combinam termos com seus papéis semânticos e a representação pred.Arg1 utiliza os preditores desambiguados e os mesmos preditores concatenados com os respectivos argumentos Arg1. Assim, as representações semânticas possuem dimensionalidade maior que a BOW e são mais esparsas. A dimensionalidade e esparsidade da BOW e das representações propostas para as coleções de documentos utilizadas nessa avaliação experimental são apresentadas na Tabela 3 (Página 96). Esses fatores podem explicar os resultados inferiores obtidos com as representações semânticas. Apesar dos atributos serem de melhor qualidade, o volume de dados não é suficiente para que os padrões de interesse sejam representados.

A representação *SR-based Sentences* se apresenta como um bom exemplo para essa condição de qualidade versus quantidade. Essa representação possui os atributos mais informativos, conforme pode ser visto na Figura 19 (Página 95). No entanto, essa é a representação que apresenta os piores resultados de classificação. Esses resultados podem ser explicados pelas características do conjunto de atributos da *SR-based Sentences*. Cada atributo dessa representação é uma combinação de atributos da representação *SR-based Terms*. Assim, apesar de *SR-based Sentences* ter um número menor de atributos do que *SR-based Terms*, a representação é mais esparsa, visto que seus atributos são mais específicos e, conseqüentemente, menos frequentes.

Já a representação *SR-based Terms* apresenta valores de acurácia mais próximos aos valores obtidos com a BOW. Para a coleção *BS-Top4*, essa representação obteve melhor performance de classificação do que a BOW em dois casos, com o uso de algoritmos de aprendizado baseados em redes bipartidas. Por apresentar performance semelhante à BOW, porém com maior

expressividade dos atributos, essa representação é indicada em casos que a expressividade dos modelos tem grande valor para o problema tratado.

A combinação de representações semânticas com a BOW pode levar a melhores resultados de classificação? Se sim, qual é a melhor maneira de combinar as representações?

Persson, Johansson e Nugues (2009) realizam a combinação de diferentes conjuntos de atributos por meio da simples união dos conjuntos, ou seja, concatenando as matrizes de duas ou mais representações em uma única matriz. Os resultados obtidos com essa representação combinada são apresentados na última coluna (coluna BOW+pred.Arg1) da Tabela 5. Neste trabalho, é proposta a combinação utilizando *ensemble* de classificadores. Assim, classificadores gerados com a representação BOW foram combinados com classificadores gerados com a representação *SR-based Terms*. Para cada estratégia de combinação (SC, WSC e MC), foram avaliadas cinco variações de pesos entre as representações¹².

As distribuições dos valores de acurácia obtidos para cada abordagem de combinação são apresentadas nas Figuras 22, 23 e 24. Nessas figuras, os pesos apresentados (de 0,1 a 0,9) nos gráficos de cada *dataset* correspondem ao peso do voto do classificador gerado com a BOW na combinação com o classificador gerado com *SR-based Terms*. De modo geral, para as três estratégias de combinação avaliadas, os melhores resultados são obtidos quando se considera os pesos de 0,7 ou 0,9 para o classificador obtido com representação BOW e 0,3 ou 0,1, respectivamente, para o classificador obtido com a representação *SR-based Terms*. Também é observado que o intervalo interquartil para os *datasets* da coleção em inglês (*SE-product*, *SE-polarity* e *SE-product-polarity*) tendem a aumentar conforme se aumenta o peso da representação BOW. Esse fato pode ser influenciado pelo comportamento da representação BOW nessa coleção, visto que os classificadores gerados com a BOW possuem intervalo interquartil maior do que os classificadores gerados com a *SR-based Terms* para essa coleção (Figuras 20a, 20c e 20d). Para a coleção de documentos em português ocorre o inverso. Os classificadores gerados com a BOW para essa coleção possuem intervalo interquartil menor do que os classificadores gerados com a *SR-based Terms*. Assim, ao aumentar o peso da coleção BOW, em geral, o intervalo interquartil da combinação diminui.

As diferentes estratégias de combinação apresentam comportamentos semelhantes. No entanto, pode-se perceber que, para o *dataset BS-topic*, como a classificação utilizando apenas a BOW já atinge ótima performance de classificação, a combinação de classificadores não se faz necessária. Já o *dataset SE-polarity*, que corresponde a um problema de maior complexidade semântica, apresenta menor acurácia quando se constrói classificadores com apenas uma das representações. Para esse *dataset*, a combinação de BOW e *SR-based Terms* apresenta melhores valores para acurácia ao se utilizar as estratégias SC e WSC.

¹² Detalhes sobre as estratégias de combinação adotadas (SC, WSC e MC) são apresentadas no Apêndice B.

Figura 22 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *SR-based Terms*, utilizando a estratégia SC.

(a) *SE-product*

(b) *BS-topic*

(c) *SE-polarity*

(d) *SE-product-polarity*

(e) *BS-semantic*

(f) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 23 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *SR-based Terms*, utilizando a estratégia WSC.

(a) *SE-product*

(b) *BS-topic*

(c) *SE-polarity*

(d) *SE-product-polarity*

(e) *BS-semantic*

(f) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 24 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *SR-based Terms*, utilizando a estratégia MC.

(a) *SE-product*

(b) *BS-topic*

(c) *SE-polarity*

(d) *SE-product-polarity*

(e) *BS-semantic*

(f) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Na [Tabela 5](#) são apresentados os maiores valores de acurácia obtidos pelas três estratégias de combinação, colunas SC, WSC e MC. Essa avaliação experimental sugere que a combinação de BOW com *SR-based Terms* leva a melhora na acurácia da classificação, em especial nos cenários de classificação com maior complexidade semântica. Nesses casos, ou seja, nos *datasets SE-polarity, SE-product-polarity, BS-semantic* e *BS-topic-semantic*, a combinação de classificadores obteve acurácia maior do que o classificador gerado com a BOW em mais da metade dos algoritmos de aprendizado considerados. Além disso, nesses casos encontram-se as maiores diferenças entre as acurácias obtidas com a BOW e com a combinação de classificadores. Para o *dataset BS-topic-semantic*, a acurácia obtida com o algoritmo IMBHN^C subiu de 62,5739% (classificador gerado com a BOW) para 68,2389% (combinações SC e WSC). Já o uso do algoritmo J48, a acurácia passou de 55,1478% (classificador gerado com a BOW) para 58,7069% (combinação SC).

Na [Figura 25](#) é apresentado o diagrama de diferença crítica considerando todas as representações e combinação de representações avaliadas (combinações separadas por peso do classificador gerado com a BOW na combinação de classificadores). Para essa análise, realizou-se o teste de Friedman $N \times N$ com pós-teste de Nemenyi, que rejeitou a hipótese nula (hipótese de que não há diferença estatisticamente significativa entre as representações), com nível de confiança de 95% e $p\text{-value} < 2.2 \times 10^{-16}$. Nos diagramas da [Figura 25](#), as representações são apresentadas de acordo com o *ranking* médio obtido considerando-se todas as configurações testadas. As representações conectadas por uma linha não apresentam diferenças estatisticamente significantes entre si. Pode-se perceber que, ao considerar os pesos mais altos para a representação BOW ([Figuras 25d](#) e [25e](#)) a combinação de classificadores fica à frente no *ranking*, porém sem diferenças estatisticamente significantes em relação aos classificadores gerados com a BOW.

4.6 Considerações finais

Nesse capítulo foi apresentado o trabalho realizado visando a geração de representações de documentos mais ricas, por meio da incorporação de rótulos de papéis semânticos. Dois modelos de representação foram propostos e avaliados na classificação automática de documentos. Tais modelos diferem da proposta de [Persson, Johansson e Nugues \(2009\)](#) em dois aspectos. Primeiro, a construção dos atributos propostos neste trabalho é mais simples, já que não é necessário realizar a desambiguação dos textos. A segunda diferença está na combinação realizada com a representação BOW. Enquanto o método da literatura realiza a combinação das representações por meio da união dos conjuntos, neste trabalho foi avaliada a combinação por meio de *ensemble* de classificadores.

Os resultados obtidos na avaliação experimental confirmam a hipótese de que os problemas do primeiro nível de complexidade semântica (organização por tópico) podem ser satisfatoriamente resolvidos com a representação *bag-of-words*, porém essa representação não é

Figura 25 – Diagrama de diferença crítica entre as representações avaliadas

(a) Combinações com peso 0,1 para BOW

(b) Combinações com peso 0,3 para BOW

(c) Combinações com peso 0,5 para BOW

(d) Combinações com peso 0,7 para BOW

(e) Combinações com peso 0,9 para BOW

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,4185$.

suficiente para a resolução de problemas mais complexos, chamados de organização semântica. Tal hipótese foi levantada na definição do problema apresentada na Seção 2.5 do Capítulo 2. Com o uso de *datasets* diferentes para as mesmas coleções de documentos, bem como diferentes algoritmos de classificação e variações de parâmetros desses algoritmos, notou-se ótimas performances na organização por tópico e performances baixas em problemas de organização semântica.

Em relação à performance das representações baseadas em papéis semânticos, verificou-se que individualmente, tanto as representações propostas quanto a representação da literatura, não apresentam informações suficientes para a classificação de documentos. Classificadores gerados com essas representações apresentam performance inferior aos classificadores gerados com a BOW. Por apresentar resultados inferiores, porém próximos aos obtidos com a BOW, os classificadores gerados com a representação *SR-based Terms* só seriam indicados para casos em que a expressividade dos atributos tenha uma grande importância na utilização do modelo gerado.

Quando a representação *SR-based Terms* é combinada com a *bag-of-words*, os resultados obtidos são melhores do que os resultados obtidos apenas com o uso da BOW. No entanto, nos experimentos realizados, os *ensembles* de classificadores gerados com ambas as representações não apresentaram diferenças estatisticamente significantes em relação ao *baseline* BOW. Porém, as diferenças obtidas considerando-se apenas os *datasets* de maior complexidade semântica, ou seja, os problemas de organização semântica sugerem que a representação enriquecida com informações semânticas pode melhorar o resultado da classificação de documentos quando combinada com a representação BOW. Buscando-se melhorar esse resultado por meio do uso de outros aspectos semânticos, realizou-se um trabalho com representações de documentos geradas a partir de representações vetoriais de palavras e sentidos de palavras. Tal trabalho é apresentado no próximo capítulo.

REPRESENTAÇÕES BASEADAS EM *EMBEDDINGS*

5.1 Considerações iniciais

No trabalho apresentado no capítulo anterior, buscou-se incorporar a semântica na representação de coleções de documentos por meio de rótulos de papéis semânticos. Neste capítulo, a semântica dos textos é abordada de maneira latente, sob a perspectiva dos modelos preditivos de semântica distribucional. A semântica distribucional é uma área de pesquisa que estuda e desenvolve teorias e métodos para o cálculo de similaridade semântica entre itens linguísticos, como palavras e expressões. Os modelos de semântica distribucional (*distributional semantic models*) baseiam-se na hipótese distribucional (*distributional hypothesis*), que afirma que palavras que ocorrem em contextos similares tendem a ter significados similares (TURNERY; PANTEL, 2010). Aplicando essa hipótese a representações de palavras no modelo espaço-vetorial, tem-se que palavras podem ser representadas por vetores cujas dimensões são contextos. Assim, palavras cujos vetores são similares tendem a ter sentidos similares.

Analogamente à representação de documentos por meio de uma *bag-of-words*, os modelos de semântica distribucional tradicionais são baseados na contagem de contextos. Um contexto pode ser definido de diversas maneiras, como janelas de palavras ou dependências sintáticas (TURNERY; PANTEL, 2010; LEVY; GOLDBERG, 2014). Modelos construídos dessa maneira tradicional, também podem ser chamados de modelos de contagem (*count models*) (BARONI; DINU; KRUSZEWSKI, 2014). Nos últimos anos surgiu uma nova abordagem para construção de modelos de semântica distribucional. Tal abordagem, inicialmente desenvolvida para modelagem de língua, faz uso de redes neurais artificiais para prever a próxima palavra dado um contexto. Como palavras similares aparecem em contextos similares, a rede aprende a atribuir vetores similares a palavras similares. Assim, como resultado desse treinamento, obtém-se um conjunto de vetores que modelam os contextos em que as palavras do corpus são observadas. Esses

modelos de semântica distribucional são chamados de modelos preditivos (*predictive models*), *neural language models* ou, simplesmente, *embeddings*.

Avanços como esses na representação de itens linguísticos (como palavras, expressões e sentidos de palavras) têm sido importantes em tarefas de semântica lexical, tais como desambiguação lexical de sentidos, anotação de papéis semânticos, identificação de similaridade semântica e analogia (TURNERY; PANTEL, 2010; BARONI; DINU; KRUSZEWSKI, 2014; CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016). A representação semântica de itens linguísticos é fundamental para o entendimento mais profundo de textos escritos em línguas naturais. Em especial, tem-se a representação dos sentidos, que é mais precisa do que a representação de palavras, visto que essa última é afetada pela polissemia e homonímia. Como uma mesma palavra pode assumir diferentes significados, a representação de sentidos resulta em representações mais precisas (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016).

Nesse contexto e com o objetivo de gerar representações semanticamente mais ricas para coleções de documentos e, conseqüentemente, melhorar os resultados da classificação de textos, neste trabalho foram propostos três modelos de representação de coleções de documentos baseados na representação vetorial de palavras e sentidos. Os modelos propostos, chamados de *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, unem a riqueza de informações obtida com a desambiguação lexical de sentidos e a disponibilidade de bons vetores pré-treinados de palavras e sentidos para construir representações de documentos em espaços de dimensionalidade reduzida.

Na próxima seção é apresentada uma revisão dos trabalhos relacionados a representações de baixa dimensionalidade e que incorporam a semântica de maneira latente. Os recursos linguísticos e ferramentas utilizados neste trabalho são apresentados na Seção 5.3. Os modelos propostos neste trabalho são descritos e analisados nas Seções 5.4 e 5.5, respectivamente. Por fim, na Seção 5.6 é apresentada a avaliação experimental dos modelos propostos na classificação de textos.

5.2 Trabalhos relacionados

Conforme apresentado anteriormente, a representação de documentos textuais é uma questão importante na Mineração de Textos, visto que o modelo de representação adotado deve manter os padrões a serem descobertos. Apesar da semântica ter um papel importante no entendimento de dados em língua natural, ela não é considerada nos modelos tradicionais de representação de documentos, que são baseados em frequência de palavras ou expressões. Com o objetivo de superar essa limitação dos modelos tradicionais, aparecem como alternativas promissoras as abordagens que tratam a semântica de maneira latente, baseadas principalmente em métodos de modelagem de tópicos ou *word embeddings*.

Latent Dirichlet allocation (BLEI; NG; JORDAN, 2003), método estado da arte em modelagem de tópicos, tem sido usado para representar coleções de documentos em um espaço

semântico de baixa dimensionalidade (CRAIN *et al.*, 2012; LU; MEI; ZHAI, 2011; LIU *et al.*, 2011; ZRIGUI *et al.*, 2012; ROSSI, 2016). Os tópicos identificados pelo LDA podem ser vistos como atributos e, assim, a distribuição dos tópicos nos documentos pode ser vista como uma representação da coleção de documentos. LDA aplica um modelo probabilístico para encontrar padrões de co-ocorrência que correspondam a tópicos semânticos na coleção de documentos. O resultado é um espaço de baixa dimensionalidade, também chamado de espaço semântico, no qual formas alternativas usadas para expressar um mesmo conceito são projetadas em uma representação comum. Com isso, espera-se que o uso do espaço semântico obtido com a aplicação do método LDA reduza ruídos causados por termos sinônimos e polissêmicos na representação de documentos.

Lu, Mei e Zhai (2011) avaliaram o LDA como método para gerar representação de textos em diferentes tarefas de MT, comparando-o com o modelo *bag-of-words* e o espaço gerado pelo método *Probabilistic Latent Semantic Analysis*. Para a tarefa de classificação de textos, os autores indicam que o espaço semântico gerado com o LDA pode ser mais efetivo que o espaço original de alta dimensionalidade, em especial quando o conjunto de dados de treinamento é pequeno. Em outros trabalhos, a representação obtida com o LDA também apresentou resultados melhores do que outros modelos de representação e em diferentes conjuntos de dados (ZRIGUI *et al.*, 2012; LIU *et al.*, 2011).

Em uma outra linha de pesquisa, vários trabalhos relacionados à geração de representações vetoriais para palavras utilizando-se modelos de redes neurais têm sido desenvolvidos na área de semântica lexical nos últimos anos. Um destaque nessa área foi a proposta dos modelos de aprendizado de vetores chamados de *Continuous Bag-of-Words* e *Skip-gram* (MIKOLOV *et al.*, 2013a). Após essa proposta, diversos trabalhos foram desenvolvidos com extensões ou aplicações desses modelos. Para verificar a eficiência das representações vetoriais geradas por meio de modelos de redes neurais, Baroni, Dinu e Kruszewski (2014) realizaram um estudo comparando os vetores gerados por meio da predição (*context-predicting vectors* ou *word embeddings*) com os modelos tradicionais baseados em contagem de co-ocorrências (*context-counting vectors*) em diferentes tarefas de semântica lexical. Os autores concluíram que as *word embeddings* são modelos de semântica distribucional superiores, obtendo resultados excelentes mesmo com apenas poucas variações dos parâmetros de treinamento.

Nesse contexto, Levy e Goldberg (2014) propõem o uso de um novo método para definir o contexto na geração de *word embeddings*. Enquanto o método proposto por Mikolov *et al.* (2013a) usa janelas de palavras para definir o contexto de palavras, Levy e Goldberg (2014) propõem o uso de informações de dependência sintática entre as palavras. Levy e Goldberg (2014) mostram que os contextos obtidos pelo método original dão ênfase a similaridades topicais, enquanto que os contextos gerados com base em dependências sintáticas são mais indicados para identificar similaridades funcionais entre as palavras. Outros trabalhos fazem uso de outros tipos de anotações no texto de entrada para gerar as *word embeddings*, como *supersenses* (FLEKOVA;

GUREVYCH, 2016) e sentidos das palavras (IACOBACCI; PILEHVAR; NAVIGLI, 2015).

Nessa linha, Camacho-Collados, Pilehvar e Navigli (2016) apresentam a abordagem NASARI *embedded* para construção de *embeddings* de sentidos de palavras no mesmo espaço vetorial de *word embeddings* pré-treinadas. Além de NASARI *embedded*, a abordagem NASARI inclui outras versões (apresentadas no Capítulo 3). Em avaliações experimentais realizadas pelos autores, a representação NASARI *embedded* apresentou bons resultados, tanto para identificação de similaridade entre palavras quanto para o agrupamento de sentidos. Camacho-Collados, Pilehvar e Navigli (2016) disponibilizaram a representação vetorial para milhões de conceitos e entidades nomeadas pertencentes à base BabelNet, descrita na próxima seção. Por serem representações ligadas aos *synsets* da BabelNet, que estão disponíveis em vários idiomas, apesar de terem sido construídos a partir de textos no idioma inglês, os vetores NASARI *embedded* são independentes de idioma.

Considerando a representação de parágrafos ou documentos completos, Le e Mikolov (2014) propuseram uma abordagem chamada *Paragraph Vector*, que posteriormente ficou conhecida como Doc2Vec. Essa abordagem foi inspirada nos trabalhos de construção de representações vetoriais de palavras por meio de redes neurais. Na abordagem Doc2Vec, um rede neural de uma camada oculta é treinada para prever o conteúdo do documento. Após o treinamento, os pesos da rede são extraídos e usados como *embeddings* para representar os documentos. Os resultados obtidos por Le e Mikolov (2014) são competitivos com outros métodos estado da arte.

Motivado pela cobertura e propriedade multilíngue dos vetores NASARI *embedded*, neste trabalho foram propostos modelos de representação de documentos baseados em *embeddings* de palavras e sentidos pré-treinadas¹. Em contraste com a abordagem Doc2Vec, a construção dos modelos propostos não requer uma grande quantidade de dados para o aprendizado do modelo de representação. Essa é uma vantagem da proposta deste trabalho, pois a quantidade de documentos disponível para o treinamento de *embeddings* pode ser um fator crítico, especialmente para pequenas coleções de documentos. Por outro lado, atualmente existem *embeddings* de palavras e sentidos pré-treinadas e de boa qualidade. As principais novidades nas abordagens propostas neste trabalho são o uso de desambiguação lexical de sentidos e de vetores NASARI *embedded*, que possibilitam a sua aplicação em diferentes idiomas.

¹ Neste trabalho foram utilizadas as *word embeddings* de Mikolov *et al.* (2013a) e os vetores NASARI *embedded* de Camacho-Collados, Pilehvar e Navigli (2016).

5.3 Recursos linguísticos e ferramentas utilizadas para gerar as representações *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*

Para o desenvolvimento deste trabalho, em especial para a geração das representações propostas e posterior análise, foram utilizados os seguintes recursos linguísticos e ferramentas.

Word2Vec. Uma implementação dos modelos *Continuous Bag-of-Words* e *Skip-gram* foram disponibilizados pelos autores (MIKOLOV *et al.*, 2013a; MIKOLOV *et al.*, 2013b) sob o nome Word2Vec². Com a popularidade dessa ferramenta, Word2Vec acabou se tornando sinônimo dos modelos em si. Juntamente com a ferramenta, os autores de Word2Vec também disponibilizaram um conjunto de vetores pré-treinados, contendo a representação vetorial de 300 dimensões para 3 milhões de palavras e expressões. O modelo foi treinado em um corpus formado por um subconjunto do *Google News dataset*, contendo aproximadamente 100 bilhões de palavras. Para a geração das representações propostas neste trabalho foi utilizado esse conjunto de vetores pré-treinados.

BabelNet. BabelNet³ (NAVIGLI; PONZETTO, 2012) é uma grande base de conhecimento, definida pelos autores como uma rede semântica de larga escala e um dicionário enciclopédico multilíngue. BabelNet foi, originalmente, construída automaticamente a partir da integração do recurso léxico do idioma inglês WordNet (FELLBAUM, 1998) e da enciclopédia colaborativa Wikipedia⁴. Assim, cada *synset* da BabelNet é associado a um conjunto de sinônimos que expressam um determinado sentido (significado de um conceito, por exemplo) integrado à página da Wikipedia correspondente. Assim, BabelNet possui informações lexicográficas e enciclopédicas de uma grande quantidade de termos, bem como relações semânticas entre eles. Para aumentar a cobertura multilíngue, utilizou-se a tradução automática. Navigli e Ponzetto (2012) mostram que a cobertura da BabelNet é muito maior do que a cobertura de *wordnets* para os idiomas catalão, francês, alemão, italiano e espanhol. Além disso, em outras avaliações experimentais realizadas em tarefas de desambiguação lexical de sentidos, os autores mostram que o uso da BabelNet possibilita a obtenção de resultados do estado da arte e indicam que essa base é um recurso de grande cobertura e de alta qualidade. Com a evolução do sistema, outros recursos foram sendo integrados à BabelNet, enriquecendo as informações disponíveis nos *synsets*. Na versão atual (BabelNet 3.7), ela conta com a integração de 15 fontes externas, apresenta cerca de 14 milhões de *synsets* e cobre 271 idiomas.

² word2vec tool: <<https://code.google.com/archive/p/word2vec/>>

³ BabelNet: <<http://babelnet.org>>

⁴ Wikipedia: <<http://www.wikipedia.org>>

Babelfy. Babelfy⁵ (MORO; RAGANATO; NAVIGLI, 2014) é um sistema de desambiguação lexical de sentidos e de ligação de entidades (*entity linking*) baseado em grafos. Babelfy não faz uso de dados anotados para treinamento de modelos. O algoritmo *random walk with restart* é utilizado para percorrer a rede semântica BabelNet e criar, para cada *synset*, uma assinatura semântica (*semantic signature*), que corresponde a um conjunto de vértices altamente relacionados). Então, para um texto de entrada a ser desambiguado, o sistema identifica candidatos a sentidos de acordo com os sinônimos presentes nos *synsets*. Logo após, os candidatos de fragmentos diferentes do texto são conectados caso um esteja na assinatura semântica do outro. A seleção dos sentidos entre os candidatos para cada fragmento é dada com a análise de regiões mais densas do grafo. Babelfy alcança resultados estado da arte tanto em tarefas de desambiguação lexical de sentidos quanto em *entity linking*. Como Babelfy é altamente integrado a BabelNet, esse sistema pode ser usado para desambiguar textos escritos em diversos idiomas, não sendo exclusivamente para o inglês.

NASARI *embedded* Na Seção 3.3 (Página 79) foi apresentada a abordagem NASARI, que possui as versões *lexical*, *unified* e *embedded* para representação vetorial de sentidos, integrada à BabelNet. Conforme apresentado, a versão *embedded* aproveita as propriedades presentes na BabelNet, o conhecimento disponível na Wikipedia e *word embeddings* treinadas em grandes corpus. NASARI tem apresentado bons resultados em diversas tarefas de semântica lexical. Neste trabalho foram utilizados os vetores NASARI *embedded* pré-treinados e disponibilizados por Camacho-Collados, Pilehvar e Navigli (2016)⁶. Esses vetores foram gerados utilizando-se os mesmos vetores Word2Vec apresentados anteriormente nessa seção. Assim, as *word embeddings* Word2Vec e os vetores NASARI para os sentidos compartilham o mesmo espaço vetorial semântico, propriedade que é explorada neste trabalho. Outra propriedade importante é a integração entre NASARI, Babelfy e BabelNet, que possibilita o uso combinado de ambos, aproveitando as vantagens de cada um desses recursos.

5.4 Modelos propostos para representação de coleções de documentos: *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*

Visando aproveitar o conhecimento embutido em representações vetoriais de palavras e sentidos de palavras, neste trabalho foram propostos três modelos de representação: *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*. A proposta é fazer uso de representações vetoriais de palavras e sentidos de palavras pré-treinadas. O uso de *embeddings* pré-treinadas tem dois principais

⁵ Babelfy: <<http://babelfy.org>>

⁶ NASARI: <<http://lcl.uniroma1.it/nasari/>>

efeitos positivos na representação de coleções de textos: (i) dimensionalidade fixa, pois os documentos são representados no mesmo espaço semântico de baixa dimensionalidade das *embeddings*; e (ii) incorporação de conhecimento externo, visto que padrões descobertos nos grandes corpus utilizados no treinamento das *embeddings* são unidos aos padrões presentes na coleção de documentos que está sendo representada.

Os três modelos baseiam-se na desambiguação dos textos e uso de *embeddings*. O modelo *NASARI_{docs}* utiliza os vetores de conceitos *NASARI embedded* e o modelo *Babel2Vec* utiliza os vetores de palavras *Word2Vec*. O terceiro modelo, *NASARI+Babel2Vec*, tem como base os vetores *NASARI embedded* e os vetores *Word2Vec*. Em *NASARI+Babel2Vec*, os vetores *Word2Vec* e *NASARI embedded* são utilizados em conjunto para aproveitar as vantagens de ambas as representações. Conforme apresentado anteriormente, o conjunto *NASARI embedded* representa *synsets* da BabelNet referentes a conceitos e entidades nomeadas. Assim, essa representação é semanticamente mais rica do que a representação *Word2Vec*. No entanto, *NASARI embedded* contém apenas vetores para *synsets* com etiquetas morfossintáticas referentes a substantivos (nomes) e, portanto, não há representação de outras classes de palavras, como verbos e adjetivos, por exemplo. Essa limitação é superada com o uso combinado de *NASARI embedded* com *Word2Vec*. Tal combinação é possível dado que os dois conjuntos de vetores compartilham o mesmo espaço semântico, propriedade apresentada por [Camacho-Collados, Pilehvar e Navigli \(2016\)](#). Além disso, o uso de *NASARI embedded* contribui com a interpretabilidade dos documentos representados no espaço das *embeddings*, conforme será apresentado posteriormente na Seção 5.5.

Uma visão geral dos processos de construção das representações propostas é apresentada nas Figuras 26 e 27. A primeira etapa da construção (Figura 26) corresponde à desambiguação dos documentos e é comum aos três modelos propostos. Os processos de construção dos modelos se diferenciam na segunda etapa, que corresponde à construção das representações em si (Figuras 27a, 27b e 27c).

Figura 26 – Processo de construção das representações *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec* - Etapa 1

Fonte: Elaborada pelo autor.

Portanto, o processo de construção das as representações se inicia com a etapa de

Figura 27 – Processo de construção das representações $NASARI_{docs}$, $Babel2Vec$ e $NASARI+Babel2Vec$ - Etapa 2

(a) Etapa 2: Construção de $NASARI_{docs}$

(b) Etapa 2: Construção de $Babel2Vec$

(c) Etapa 2: Construção de $NASARI+Babel2Vec$

Fonte: Elaborada pelo autor.

Quadro 4 – Descrições dos *synsets* identificados na sentença “*The Toshiba Net book operates very well.*”

ID	Fragmento	Glosa em inglês	Glosa em português
bn:03423971n	Toshiba	<i>Toshiba is a Japanese multinational conglomerate corporation headquartered in Tokyo, Japan.</i>	Toshiba é uma empresa japonesa produtora de aparelhos eletrônicos e semicondutores.
bn:00024712n	Net	<i>A computer network consisting of a worldwide network of computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange.</i>	Internet é um sistema global de redes de computadores interligadas que utilizam o conjunto de protocolos padrão da internet para servir vários bilhões de usuários no mundo inteiro.
bn:03754555n	Net book	<i>Netbooks was a category of small, lightweight, legacy-free, and inexpensive computers that were introduced in 2007.</i>	Netbooks são uma categoria menor, mais leve e mais barata de laptops.
bn:00012059n	book	<i>A written work or composition that has been published (printed on pages bound together).</i>	Livro é um volume transportável, composto por páginas encadernadas, contendo texto manuscrito ou impresso e/ou imagens e que forma uma publicação unitária ou a parte principal de um trabalho literário, científico ou outro.
bn:00088629v	operates	<i>Perform as expected when applied.</i>	funcionar, operar - sem glosa no idioma português -
bn:00115380r	very well	<i>Quite well.</i>	- termo não disponível no idioma português -

desambiguação (Figura 26)⁷. Neste trabalho foi utilizado o sistema Babelfy, apresentado na Seção 5.3. Para um dado documento, o sistema Babefy retorna os *synsets* referentes aos termos (palavras ou expressões) desambiguados. Essa é uma etapa importante do processo, pois revela os conceitos e entidades nomeadas que são representadas pelas palavras presentes nos documentos. Na desambiguação, caso seja identificado que uma expressão formada por mais de um *token* (*multi-token expression*) corresponde a um único conceito, mais de um *synset* pode ser retornado para cada uma dessas palavras. Na maior parte das ocorrências, o *synset* mais específico é o *synset* de interesse. Portanto, o conjunto de *synsets* de cada documento é processado para manter apenas os *synsets* mais específicos para cada expressão.

Por exemplo, no Quadro 4 são apresentados os *synsets* identificados na sentença “*The Toshiba Net book operates very well.*”⁸. Nessa sentença, são retornados três *synsets* para os *tokens* “*Net*” e “*book*”, sendo dois *synsets* para os dois *tokens* individualmente e um *synset* para a expressão *multi-token* “*Net book*”. Quando expressões *multi-token* são identificadas, o *synset* mais específico (aquele identificado para a expressão que abrange o número maior de *tokens*) é selecionado para ser utilizado na construção das representações propostas. Assim, no caso do exemplo apresentado no Quadro 4, seria selecionado o *synset* correspondente a “*Net book*”.

O conjunto de documentos desambiguados é usado para construir as representações

⁷ Esse etapa de desambiguação também é realizada para a construção da representação *Uni-based*, apresentada no Capítulo 3 (Página 79).

⁸ Sentença extraída de um dos documentos da coleção SE-ABSA15, descrita na Seção 2.5.2 (Página 64).

propostas (Figuras 27a, 27b e 27c). Como o procedimento para construção da representação *NASARI+Babel2Vec* engloba o procedimento para construção das outras duas representações, este será apresentado em detalhes e, na sequência, serão descritas as modificações necessárias para a construção dos modelos *NASARI_{docs}* e *Babel2Vec*.

O procedimento para construção da representação *NASARI+Babel2Vec* é apresentado no Algoritmo 1. O processo de construção das representações recebe como entrada o conjunto de documentos desambiguados (D') e os conjuntos de vetores *NASARI embedded* (E) e *Word2Vec* (G). Para cada documento $d' \in D'$ são retornados os vetores de *NASARI embedded* para os *synsets* identificados (linha 7). Para os *synsets* que não possuem um vetor em *NASARI embedded*, é buscado o vetor em *Word2Vec* correspondente ao principal sinônimo (*lexicalization*) do *synset* (linhas 11 a 16). Após serem processados todos os *synsets* de d' , é gerada a representação do documento, que equivale ao centroide dos vetores retornados de E e G (linha 19). Após o processamento de todos os documentos, obtém-se a matriz correspondente à representação da coleção de documentos, na qual os documentos são representados no espaço de baixa dimensão (correspondente a dimensionalidade dos conjuntos E e G).

Os processos de construção das representações *NASARI_{docs}* e *Babel2Vec* são similares ao processo de construção de *NASARI+Babel2Vec*. A diferença é que para a construção de *NASARI_{docs}* são considerados apenas os vetores *NASARI embedded*; e para a construção de *Babel2Vec* são considerados apenas os vetores *Word2Vec*. Assim, o procedimento de construção de *NASARI_{docs}* corresponde ao procedimento apresentado no Algoritmo 1 com a exclusão do processamento dos vetores de *Word2Vec* (linhas 11 a 16). Já na construção da representação *Babel2Vec* as linhas de 6 a 10 no Algoritmo 1 não são consideradas, sendo buscados apenas os vetores em *Word2Vec* (linhas 12 e 13). Aqui vale ressaltar a vantagem do uso de *BabelNet* na etapa de desambiguação dos documentos. Visto que esse sistema é integrado à *BabelNet*, é possível utilizar *embeddings* construídas a partir de textos em idioma diferente ao idioma da coleção de documentos em questão. Enquanto que para documentos escritos no mesmo idioma da representação *Word2Vec*, as próprias palavras e expressões utilizadas nos documentos podem ser utilizadas para retornar os vetores *Word2Vec*, para documentos escritos em outros idiomas, pode-se utilizar um dos sinônimos do *synset* presente na *BabelNet*. Como os principais recursos linguísticos estão disponíveis para o idioma inglês, pode-se utilizar modelos de representações vetoriais de palavras e sentidos de palavras em inglês para representar documentos de qualquer um dos 271 idiomas atualmente presentes na *BabelNet*.

5.5 Análise das representações *NASARI+Babel2Vec*, *NASARI_{docs}* e *Babel2Vec*

Considerando a interpretabilidade das representações, a *bag-of-words* tem vantagem sobre as representações baseadas em *embeddings*. Enquanto os atributos da BOW normalmente

Algoritmo 1 – Construção da representação de coleções de textos *NASARI+Babel2Vec*.**Entrada:** D' , conjunto de documentos desambiguados E , conjunto de vetores NASARI *embedded* G , conjunto de vetores Word2Vec**Saída:** $M_{N \times F}$, matriz representando a coleção de documentos D , sendo N o número de documentos de D e F a dimensionalidade dos vetores E e G

```

1:  $n \leftarrow 0$ 
2:  $M \leftarrow$  empty matrix
3: para todo documento  $d' \in D'$  faça
4:    $\vec{doc} \leftarrow$  empty vector
5:   para todo synset  $s \in d'$  faça
6:     se  $s \in E$  então
7:        $\vec{v} \leftarrow$  vetor do synset  $s$  em  $E$ 
8:        $\vec{doc} \leftarrow \vec{doc} + \vec{v}$ 
9:        $n \leftarrow n + 1$ 
10:    senão
11:      se  $s \in G$  então
12:         $frag \leftarrow$  lexicalization do synset  $s$ 
13:         $\vec{v} \leftarrow$  vetor de  $frag$  em  $G$ 
14:         $\vec{doc} \leftarrow \vec{doc} + \vec{v}$ 
15:         $n \leftarrow n + 1$ 
16:      fim se
17:    fim se
18:  fim para
19:   $\vec{doc} \leftarrow \frac{\vec{doc}}{n}$ 
20:  append  $\vec{doc}$  to  $M$ 
21: fim para
22: retorna  $M$ 

```

são palavras normalizadas, os atributos das representações baseadas em *embeddings* não são interpretáveis. No entanto, graças a desambiguação e uso de NASARI *embedded*, as representações NASARI_{docs} e NASARI+Babel2Vec têm, por meio do espaço em que os documentos são representados, a interpretabilidade um pouco mais avançada do que as demais representações baseadas em *embeddings*. Tal característica é apresentada no decorrer desta seção.

Por outro lado, uma vantagem das representações propostas sobre a BOW é a baixa dimensionalidade, herdada das representações vetoriais de palavras e sentidos de palavras. A dimensionalidade fixa das *embeddings* é normalmente menor do que o número de atributos de uma *bag-of-words*. A baixa dimensionalidade na representação de coleções de documentos pode ser mais eficiente e diminuir o tempo de processamento dos algoritmos de AM. Outra propriedade importante das representações propostas é que elas incorporam conhecimento externo às coleções que estão sendo representadas, sem a necessidade de treinamento adicional. As *embeddings*

pré-treinadas são construídas com base em grandes corpora. Por exemplo, as *word embeddings* de Word2Vec utilizadas neste trabalho foram construídas com base em um corpus formado por aproximadamente 100 bilhões de palavras. Nos modelos propostos, esse conhecimento é naturalmente transmitido para as representações das coleções de documentos e, assim, podem melhorar os padrões escondidos nas mesmas.

Com o objetivo de analisar a qualidade de representações de documentos geradas a partir de *embeddings* de palavras e/ou sentidos, bem como a habilidade das mesmas em representar o conteúdo dos documentos, foi feita uma análise com uma amostra dos documentos (composta por documentos nos idiomas inglês e português). Essa análise foi realizada com base nos vizinhos mais próximos de cada documento no espaço semântico das *embeddings*. A similaridade entre os vetores dos documentos e das palavras (no caso de Word2Vec) e/ou dos *synsets* (no caso de NASARI *embedded*) foi calculada utilizando a medida Cosseno. As análises dos vetores mais próximos aos vetores dos documentos da amostra mostrou que os vizinhos, na maioria dos casos, são relacionados ao tópico principal do documento.

Na sequência, é apresentada uma análise dos modelos propostos neste trabalho, tanto em relação à proximidade dos documentos a palavras e conceitos (*synsets*) quanto na identificação de similaridade textual.

5.5.1 Proximidade a palavras e conceitos

A análise apresentada nesta seção foi realizada com os documentos selecionados aleatoriamente da coleção de documentos SE-ABSA15 (descrita na Seção 2.5.2, [Página 64](#)) e apresentados a seguir.

Documento A: *The Toshiba Net book operates very well. The only objection I have is that after you buy it the windows 7 system is a starter and charges for the upgrade.*

Documento B: *I've had my Macbook Pro since August 2009. Prior to this computer, I owned a PowerBook G4 for 6 years (quite a long time for a laptop). That was my first Apple product and since then I have been incredibly happy with every product of theirs I have bought. My MacBook Pro is no exception. On my PowerBook G4 I would never use the trackpad I would use an external mouse because I didn't like the trackpad. Since I've had this computer I've only used the trackpad because it is so nice and smooth. I also like that you can scroll down in a window using two fingers on the trackpad. The display is incredibly bright, much brighter than my PowerBook and very crisp. The computer runs very fast with no problems and the iLife software that comes with it (iPhoto, iMovie, iWeb, iTunes, GarageBand) is all very helpful as well. I also purchased iWork to go with it which has programs for word processing, spreadsheets, and presentations (similar to Microsoft Office). I like those programs better than Office and you can save your files to be completely compatible with the Office programs as well. I would*

recommend this laptop to anyone looking to get a new laptop who is willing to spend a little more money to get great quality!

Para cada modelo de representação analisado (NASARI_{docs}, NASARI+Babel2Vec e Babel2Vec), o espaço vetorial considerado para encontrar o vizinho mais próximo ao documento é o espaço formado pelos vetores utilizados na geração da representação. Portanto, para a representação Babel2Vec foi utilizado o espaço da Word2Vec. Tanto para a construção da representação quanto para essa análise, para a representação NASARI_{docs} foram considerados um subconjunto de NASARI *embedded*, formado pelos conceitos referentes a páginas da Wikipedia com pelo menos cinco *backlinks* na própria Wikipedia. Já para a representação NASARI+Babel2Vec, foi considerado o espaço formado pela união dos dois conjuntos de vetores anteriores.

Para esses três modelos também foi analisado o impacto do uso de *synsets* correspondentes ao sentido mais comum das palavras (MCS - do inglês *most common sense*) retornado pelo sistema Babelfy. Nesse sistema, o MCS de um *token* é retornado no processo de desambiguação quando o *score* de desambiguação está abaixo de um determinado limiar.

No [Quadro 5](#) são apresentados os *synsets* identificados no *Documento A* após a execução da primeira etapa de construção das representações, apresentada na [Seção 5.4](#). Para cada *synset* é apresentado o seu identificador na BabelNet, o correspondente fragmento do texto, a principal palavra no conjunto de sinônimos, a glosa (definição informal) do *synset* no idioma inglês e o *score* de desambiguação. Pode-se perceber que os primeiros oito *synsets* apresentados no [Quadro 5](#) são bem relacionados ao conteúdo do *Documento A*. Já os dois últimos *synsets* não são relacionados. Para o caso do fragmento “*buy it*”, o *synset* correto seria “*purchase - Obtain by purchase; acquire by means of a financial transaction*” (bn:00084331v); e para o fragmento “*charges*”, o *synset* correto seria “*bill - Demand payment*” (bn:00083486v). Ambos os *synsets* incorretos foram retornados como sendo o MCS do respectivo fragmento, ou seja, o *score* de desambiguação dos mesmos foi abaixo do limiar. No caso do *Documento A*, pode-se verificar que quatro *synsets* foram retornados como MCS, sendo que dois corretamente correspondem ao conteúdo do documento e dois correspondem a sentidos incorretos. Para se analisar o impacto de MCS na representação de documentos, para cada modelo analisado foram consideradas duas variações (com e sem MCS).

No [Quadro 6](#) são apresentadas as 5 palavras ou *synsets* mais próximos aos vetores que representam o *Documento A*, juntamente com as respectivas taxas de similaridade. O *Documento A* refere-se a algumas características de um *netbook* da marca Toshiba. Os *synsets* mais próximos são relacionados a computadores, que é o tópico do documento. Percebe-se que as representações parecem generalizar o conteúdo do documento, visto que conceitos mencionados no texto, como “*netbook*” e “*Windows 7*” não estão entre os cinco vizinhos mais próximos. A única exceção a esse fato é a palavra “*Toshiba*”, que aparece entre as cinco palavras mais próximas da representação Babel2Vec (sem MCS) do *Documento A*. No entanto, a similaridade entre eles é baixa (0,58). Um

Quadro 5 – Synsets identificados no Documento A

ID	Fragmento	Sinônimo principal	Glosa em inglês	Score
1	bn:03423971n	Toshiba	<i>Toshiba is a Japanese multinational conglomerate corporation headquartered in Tokyo, Japan.</i>	1,00
2	bn:03754555n	Net book	<i>Netbooks was a category of small, lightweight, legacy-free, and inexpensive computers that were introduced in 2007.</i>	1,00
3	bn:00088629v	operates	<i>Perform as expected when applied.</i>	1,00
4	bn:02615501n	windows 7	<i>Windows 7 is a personal computer operating system developed by Microsoft, a version of Windows NT.</i>	1,00
5	bn:00079241n	upgrade	<i>Software that provides better performance than an earlier version did.</i>	0,68
6	bn:15125301n	system	<i>A system is a set of interacting or interdependent components forming an integrated whole.</i>	0,63
7	bn:00115380r	very well	<i>Quite well.</i>	0
8	bn:00032373n	objection	<i>The act of expressing earnest opposition or protest.</i>	0
9	bn:00084340v	buy it	<i>Be killed or die.</i>	0
10	bn:00017789n	charges	<i>An impetuous rush toward someone or something.</i>	0

Nota – Synsets com score 0 correspondem ao MCS do fragmento.

destaque na análise do Documento A é o conceito “Vendor lock-in”, que aparece como quinto synset mais próximo das representações $NASARI_{docs}$ do Documento A, conforme apresentado no Quadro 6. Esse conceito está relacionado ao conteúdo do documento, visto que no documento o autor apresenta uma objeção ao fato de ser cobrado para fazer o *upgrade* do sistema operacional do referido *netbook*.

Pode-se verificar que o uso combinado dos vetores de $NASARI_{embedded}$ e Word2Vec tem impacto sobre a posição dos documentos no espaço vetorial. Os exemplos indicam que os 5 synsets mais próximos a $NASARI_{docs}$ tendem a ser mais genéricos do que os 5 synsets mais próximos a $NASARI+Babel2Vec$. Para o Documento A, a maioria dos vizinhos mais próximos a $NASARI+Babel2Vec$ correspondem a sistemas operacionais ou são relacionados a eles. Esse conceito é bem relacionado ao conteúdo do Documento A, visto que o sistema operacional Windows 7 é a fonte de preocupação do autor de tal documento. Vale ressaltar que, apesar do synset correspondente a Windows 7 não estar entre os 5 synsets mais próximo à representação $NASARI+Babel2Vec$ do Documento A, ele ainda assim está próximo. A similaridade entre tais vetores é de 0,78 para a versão sem MCS, e de 0,69 para a versão com MCS.

O Documento B se refere a algumas características do Macbook Pro e o compara com outro produto, o PowerBook. No Quadro 7 são apresentados os vizinhos mais próximos a esse documento, considerando os diferentes modelos de representação. Pode-se perceber que a representação $NASARI_{docs}$ desse documento é similar a conceitos relacionados à Apple. Tal representação sem MCS é similar ao synset correspondente ao produto PowerBook, que é mencionado no documento. Esse synset não está entre os 5 mais próximos a representação $NASARI_{docs}$ com MCS, porém a similaridade entre o vetor dessa representação e o vetor do

Quadro 6 – As 5 palavras ou synsets mais próximos às representações do Documento A

Sim.	Palavra ou synset
<i>NASARI_{docs} - com MCS</i>	
0,94	<i>System program: A program (as an operating system or compiler or utility program) that controls some aspect of the operation of a computer.</i>
0,94	<i>Pre-installed software: Pre-installed software is the software already installed and licensed on a computer or smartphone bought from an original equipment manufacturer.</i>
0,94	<i>Plug and play: In computing, a plug and play device or computer bus, is one with a specification that facilitates the discovery of a hardware component in a system without the need for physical device configuration or user intervention in resolving resource conflicts.</i>
0,94	<i>microcomputer: A small digital computer based on a microprocessor and designed to be used by one person at a time.</i>
0,94	<i>Vendor lock-in: In economics, vendor lock-in, also known as proprietary lock-in or customer lock-in, makes a customer dependent on a vendor for products and services, unable to use another vendor without substantial switching costs.</i>
<i>NASARI_{docs} - sem MCS</i>	
Mesmos synsets de <i>NASARI_{docs} - com MCS</i> .	
<i>NASARI+Babel2Vec - com MCS</i>	
0,77	<i>Burroughs MCP: The MCP is the proprietary operating system of the Burroughs small, medium and large systems, including the Unisys Clearpath/MCP systems.</i>
0,77	<i>XTS-400: The XTS-400 is a multi-level secure computer operating system.</i>
0,76	<i>RSTS/E: RSTS is a multi-user time-sharing operating system, developed by Digital Equipment Corporation, for the PDP-11 series of 16-bit minicomputers.</i>
0,76	<i>UNIVAC EXEC 8: EXEC 8 was UNIVAC's operating system developed for the UNIVAC 1108 in 1964.</i>
0,76	<i>System requirements: To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer.</i>
<i>NASARI+Babel2Vec - sem MCS</i>	
0,88	<i>NOS (software): NOS was an operating system with time-sharing capabilities, written by Control Data Corporation in the 1970s.</i>
0,87	<i>CDC Kronos: Kronos is an operating system with time-sharing capabilities, written by Control Data Corporation in the 1970s.</i>
0,87	<i>History of operating systems: Computer operating systems provide a set of functions needed and used by most application programs on a computer, and the linkages needed to control and synchronize computer hardware.</i>
0,87	<i>Resident monitor: A resident monitor was a piece of system software in many early computers from the 1950s to 1970s.</i>
0,87	<i>CDC SCOPE: SCOPE, an acronym for Supervisory Control Of Program Execution, was the name used by the Control Data Corporation for a number of operating system projects in the 1960s.</i>
<i>Babel2Vec - com MCS</i>	
0,54	upgrade
0,52	operates
0,51	NEC_Renesas
0,51	resells_Dish
0,51	mark_LabWindows
<i>Babel2Vec - sem MCS</i>	
0,62	upgrade
0,59	operates
0,58	Toshiba
0,57	system
0,55	systems

synset PowerBook é de 0,92, portanto, eles ainda estão bem próximos.

A representação *NASARI+Babel2Vec* do *Documento B* está próxima a conceitos mais específicos, que também são mencionados no documento: GarageBand, Numbers e Pages, que são componentes do pacote iWorks do Apple. O *synset* correspondente ao Pages, que está entre os 5 mais próximos à representação *NASARI+Babel2Vec* sem MCS, não está entre os 5 mais próximos a representação com MCS, porém é o sexto *synset* mais próximo, com similaridade de 0,85.

A representação *Babel2Vec* do *Documento B* é similar a entidades mencionadas no documento. Tal representação sem MCS é mais similar aos produtos MacBook e PowerBook, já a representação com MCS é mais similar a aspectos dos produtos, como *trackpad* e iLife Suite.

Em relação ao uso de MCS nas representações dos documentos, pode-se notar que as representações com ou sem MCS são bem similares. Por exemplo, para o *Documento A*, as representações *NASARI_{docs}* com e sem MCS possuem os mesmos 5 vizinhos mais próximos. No caso do *Documento B*, os conceitos mais relevantes que estão entre os 5 vizinhos mais próximos às representações *NASARI_{docs}* e *NASARI+Babel2Vec* sem MCS também estão próximos às respectivas representações com MCS. As similaridades entre os vetores das representações com e sem MCS, calculadas com base na distância de Cosseno para ambos os documentos, são apresentadas na [Tabela 6](#).

Conforme apresentado anteriormente, o uso de MCS pode trazer ruído a representação, visto que o MCS pode não ser o *synset* correto de acordo com o conteúdo do documento. Esse é o caso dos *synsets* “buy it” e “charge” retornados para o *Documento A* ([Quadro 5](#)). Por outro lado, o descarte de MCS limita o número de *synsets* identificados e pode impossibilitar a construção da representação do documento, especialmente no caso de documentos curtos. Por exemplo, em 15 dos 815 documentos da coleção *SE-ABSA15* não são identificados nenhum *synset* ao se descartar MCS e, portanto, esses documentos não podem ser representados. Ainda, para os casos de documentos em que poucos *synsets* são identificados, a representação pode ser impossibilitada por não haver vetores nos modelos pré-treinados para esses poucos *synsets*. Por outro lado, parte do ruído inserido pelo uso de MCS pode ser reduzido ao se considerar um subconjunto dos vetores *NASARI embedded*. Considerando-se o subconjunto formado apenas por conceitos relacionados a páginas da Wikipedia com pelo menos cinco *backlinks* na própria Wikipedia, apenas um dos quatro MCS (apresentados no [Quadro 5](#)) é considerado na representação do *Documento A*.

Um problema similar pode impossibilitar a construção da representação *NASARI_{docs}*. Como em *NASARI embedded* estão disponíveis apenas vetores para substantivos, pode não ser possível representar determinados documentos utilizando tal modelo. Isso pode afetar principalmente documentos de opinião curtos, que podem conter apenas adjetivos, como “bom”, “péssimo”, “limpo e confortável”. Por exemplo, no caso da coleção *SE-ABSA15*, não foi possível representar 6 dos 815 documentos utilizando o modelo *NASARI_{docs}*.

Quadro 7 – As 5 palavras ou synsets mais próximos às representações do Documento B

Sim.	Palavra ou synset
<i>NASARI_{docs} - com MCS</i>	
0,94	<i>Apple II series: The Apple II series is a set of home computers, one of the first highly successful mass-produced microcomputer products, designed primarily by Steve Wozniak, manufactured by Apple Computer and introduced in 1977 with the original Apple II.</i>
0,94	<i>Xerox Alto: The Xerox Alto was one of the first personal computers, a general purpose computer designed for individual use.</i>
0,93	<i>Apple IIGS: The Apple IIGS is the fifth and most powerful model in the Apple II series of personal computers produced by Apple Computer.</i>
0,93	<i>Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.</i>
0,93	<i>Apple Desktop Bus: Apple Desktop Bus is a bit-serial computer bus connecting low-speed devices to computers.</i>
<i>NASARI_{docs} - sem MCS</i>	
0,95	<i>Xerox Alto: The Xerox Alto was one of the first personal computers, a general purpose computer designed for individual use.</i>
0,95	<i>Apple II series: The Apple II series is a set of home computers, one of the first highly successful mass-produced microcomputer products, designed primarily by Steve Wozniak, manufactured by Apple Computer and introduced in 1977 with the original Apple I.</i>
0,95	<i>Apple Desktop Bus: Apple Desktop Bus is a bit-serial computer bus connecting low-speed devices to computers.</i>
0,94	<i>Apple IIGS: The Apple IIGS is the fifth and most powerful model in the Apple II series of personal computers produced by Apple Computer.</i>
0,94	<i>PowerBook: The PowerBook is a line of Macintosh laptop computers that was designed, manufactured and sold by Apple Computer, Inc. from 1991 to 2006.</i>
<i>NASARI+Babel2Vec - com MCS</i>	
0,87	<i>Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.</i>
0,85	<i>Numbers (spreadsheet): Numbers is a spreadsheet application developed by Apple Inc. as part of the iWork productivity suite alongside Keynote and Pages.</i>
0,85	<i>GarageBand: GarageBand is a software application for OS X and iOS that allows users to create music or podcasts.</i>
0,85	<i>Keyboard shortcut: In computing, a keyboard shortcut is a series of one or several keys that invoke a software or operating system operation when triggered by the user.</i>
0,85	<i>Cut, copy, and paste: In human – computer interaction, cut and paste and copy and paste are related commands that offer a user-interface interaction technique for transferring text, data, files or objects from a source to a destination.</i>
<i>NASARI+Babel2Vec - sem MCS</i>	
0,89	<i>Pages (word processor): Pages is a word processor and a page layout application developed by Apple Inc.</i>
0,89	<i>Numbers (spreadsheet): Numbers is a spreadsheet application developed by Apple Inc. as part of the iWork productivity suite alongside Keynote and Pages.</i>
0,89	<i>Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.</i>
0,89	<i>GarageBand: GarageBand is a software application for OS X and iOS that allows users to create music or podcasts.</i>
0,89	<i>MobileMe: MobileMe was a subscription-based collection of online services and software offered by Apple Inc.</i>
Continua na página seguinte.	

Quadro 7 – As 5 palavras ou *synsets* mais próximos às representações do *Documento B* (continuação)

Continuação da página anterior.	
Sim.	Palavra ou <i>synset</i>
<i>Babel2Vec</i> - com MCS	
0,74	1Gig_DIMM
0,73	MacBook_trackpad
0,73	Apple_iLife_suite
0,72	G5_Quad
0,72	Macbook
<i>Babel2Vec</i> - sem MCS	
0,75	Macbook
0,75	MacBook_Pro
0,74	MacBook
0,74	PowerBook
0,74	Macbook_Pro

Assim, a avaliação experimental na classificação de textos apresentada na Seção 5.6 foi realizada com as representações *Babel2Vec* e *NASARI+Babel2Vec* considerando os *synsets* retornados como MCS. Três motivos levaram a essa seleção de modelos a serem avaliados: (i) o descarte de MCS limita a habilidade de se representar textos curtos; (ii) o uso de um subconjunto de *NASARI embedded* pode filtrar parte dos *synsets* que inserem ruídos na representação; e (iii) o modelo de representação *NASARI_docs* é limitado a apenas substantivos.

Tabela 6 – Similaridade entre as representações de documentos com e sem MCS

	Modelo de representação	Similaridade
<i>Documento A</i>	<i>NASARI_docs</i>	1,00
	<i>NASARI+Babel2Vec</i>	0,87
	<i>Babel2Vec</i>	0,85
<i>Documento B</i>	<i>NASARI_docs</i>	0,99
	<i>NASARI+Babel2Vec</i>	0,98
	<i>Babel2Vec</i>	0,98

5.5.2 Similaridade textual

Complementando a análise apresentada na subseção anterior, diferentes modelos de representação de documentos foram avaliados na tarefa de identificação de similaridade semântica com o *dataset* de Lee, Pincombe e Welsh (2005). Esse *dataset* foi construído para possibilitar a comparação entre similaridades textuais calculadas automaticamente com similaridades atribuídas por pessoas e, assim, avaliar a capacidade de diferentes modelos em simular julgamentos feitos por humanos (LEE; PINCOMBE; WELSH, 2005). O *dataset* é composto por 50 documentos selecionados entre notícias veiculadas por um serviço de *e-mails* da *Australian Broadcasting Corporation*, cujos tamanhos variam de 51 a 126 palavras. A similaridade entre dois documentos (totalizando 1225 pares) foi avaliada por 83 estudantes universitários, atribuindo uma taxa de

similaridade variando de 1 (altamente não relacionados) a 5 (altamente relacionados). Cada par foi avaliado por pelo menos 8 pessoas (anotadores).

Na análise apresentada nesta seção, foram considerados, além dos dois modelos propostos, outros três modelos de representação da literatura: BOW, LDA e Word2Vec⁹. As similaridades entre os pares de documentos foram calculadas por meio da distância de Cosseno e a correlação entre essas similaridades e as similaridades médias atribuídas pelos anotadores foram calculadas usando os coeficientes de Pearson e de Spearman. O coeficiente de correlação de Pearson mede a correlação linear entre as duas variáveis, que no caso desta análise são a similaridade atribuída pelos anotadores e a similaridade obtida com a distância de Cosseno. Já o coeficiente de correlação de Spearman avalia os *rankings* das duas variáveis e mede o quanto o relacionamento entre as duas variáveis pode ser descrito por uma função monótona.

As correlações entre as taxas de similaridade atribuídas pelos anotadores e a similaridades calculadas utilizando cada um dos modelos de representação são apresentadas na Tabela 7. Nessa tabela, o valor identificado por *inter-rater correlation* corresponde à uma medida produzida por Lee, Pincombe e Welsh (2005) para mensurar a correlação entre as atribuições dadas pelos anotadores. O valor de *inter-rater correlation* foi calculado como sendo a correlação de Pearson entre um *rating* selecionado aleatoriamente e a média dos demais *ratings* para cada par de documentos.

Tabela 7 – Correlação entre similaridades atribuídas por anotadores e similaridades calculadas por distância de vetores

Representação	Correlação	
	Pearson	Spearman
NASARI+Babel2Vec	0,532	0,541
Babel2Vec	0,659	0,555
BOW	0,557	0,289
LDA	-0,039	-0,034
Word2Vec	0,558	0,475
<i>Inter-rater correlation</i>	0,605	-

Nota – As correlações mais altas são apresentadas em negrito.

Pode-se verificar que a representação LDA utilizando 300 dimensões não é adequada ao *dataset* em questão, visto que é um *dataset* de tamanho bem reduzido. Esse fato pode ser verificado pela baixa correlação apresentada pelo modelo de representação LDA. Os demais modelos de representação apresentam correlações mais alta em relação às taxas de similaridade fornecidas pelos anotadores. Na avaliação de Lee, Pincombe e Welsh (2005), a correlação mais alta foi de 0,6 obtida com um modelo baseado em LSA, porém utilizando um corpus estendido, composto por 364 documentos. Nenhum dos modelos avaliados alcançou o valor de *inter-rater correlation* utilizando o corpus de 50 documentos. Nesta análise, a representação que apresentou

⁹ Detalhes sobre a geração das representações são apresentados no Apêndice B.

as melhores correlações, tanto de Person quanto de Spearman, foi a representação proposta *Babel2Vec*, sendo inclusive maior do que a correlação entre anotadores (*inter-rater correlation*).

Na [Figura 28](#) são apresentadas as relações entre as taxas de similaridades atribuídas pelos anotadores e as similaridades calculadas por meio da distância de Cosseno para os modelos de representação com melhores correlações na [Tabela 7](#). Pode-se notar que a BOW é deficiente na representação das similaridades entre os documentos, especialmente para pares que os anotadores atribuíram similaridades entre 2 e 4. Quando os documentos são representados utilizando-se a BOW, os vetores tendem a ficarem mais distantes e, portanto, as similaridades tendem a ser baixas. Os outros modelos de representação apresentam um padrão melhor de relacionamento com as similaridades dos anotadores e também apresentam maiores valores de correlação de Spearman. No geral, esses modelos de representação identificam bem similaridades altas e são menos eficientes na identificação de similaridades baixas.

Figura 28 – Relações entre similaridades atribuídas por anotadores e similaridades calculadas por distância de vetores

Fonte: Elaborada pelo autor.

Nota – A coleção de documentos e as anotações de similaridade identificadas por anotadores foram disponibilizadas por [Lee, Pincombe e Welsh \(2005\)](#).

5.6 Avaliação experimental - representações baseadas em *embeddings*

As representações propostas, *Babel2Vec* e *NASARI+Babel2Vec*, foram avaliadas em cenários de classificação de textos. Essa avaliação experimental foi realizada com os objetivos de avaliar a performance de classificação em relação a diferentes níveis de complexidade semântica e o impacto do uso de *embeddings* na representação de documentos. Com essa avaliação, discute-se as seguintes perguntas: (i) Representações baseadas em *embeddings* de termos desambiguados podem levar a melhores resultados de classificação?; e (ii) A combinação de representações baseadas em *embeddings* de termos desambiguados com a BOW leva a melhores resultados de classificação?.

5.6.1 *Datasets, representações de textos e configuração experimental*

Essa avaliação experimental foi realizada utilizando-se os *datasets* das coleções *BEST sports Top 4*, *SE-ABSA15* e *BBC*, apresentados na Seção 2.5.2 (Página 64). Os modelos de representação baseados em *embeddings*, propostos e apresentados na Seção 5.4, foram avaliados em relação a quatro outros modelos da literatura. Assim, para cada *dataset*, foram geradas seis representações no modelo espaço-vetorial, listadas a seguir.

- *Babel2Vec*;
- *NASARI+Babel2Vec*;
- BOW (*baseline*);
- LDA;
- Word2Vec;
- Doc2Vec.

Por simplicidade, as representações *Babel2Vec*, *NASARI+Babel2Vec*, LDA, Word2Vec e Doc2Vec são chamadas de representações semânticas, visto que essas representações contêm algum tipo de enriquecimento semântico em relação à representação BOW.

Os experimentos foram realizados utilizando variações de parâmetros de seis algoritmos de AM. Também foi avaliada a combinação de classificadores, utilizando três estratégias para compor os votos entre os diferentes classificadores base: *Sum of Confidences* (SC), *Weighted Sum of Confidences* (WSC) e *Most Confident* (MC). Detalhes sobre a geração das representações, bem como sobre a configuração experimental adotada, são apresentados no Apêndice B.

5.6.2 Resultados - embeddings

A execução da configuração experimental adotada resultou em 104 classificadores para cada representação avaliada em cada *dataset*. A distribuição dos valores de acurácia em cada conjunto de dados é apresentada na Figura 29. As duas primeiras linhas de *box plots* (Figuras 29a, 29b e 29c) correspondem aos *datasets* do primeiro nível de complexidade semântica, ou seja, organização por tópico. Esses *datasets* apresentam os melhores valores de acurácia. Pode-se perceber que o uso de representações semânticas leva à redução do intervalo interquartil, apresentando os valores de acurácia próximos a 100%, com exceção de alguns *outliers*.

A terceira linha de *box plots* (Figuras 29d e Figure 29e) correspondem aos *datasets* do segundo nível de complexidade semântica (organização semântica) da coleção de documentos em inglês (*SE-ABSA15*). Para esses *datasets*, os valores máximos de acurácia ficam em torno de 80%. Já a distribuição dos valores de acurácia para os *datasets* que correspondem à organização semântica da coleção de documentos em português (*BS-Top4*) são apresentadas nas Figuras 29f e 29g. Para esses *datasets* os valores máximos são um pouco mais baixos, ficando abaixo de 70%.

No caso dos *datasets* correspondentes à organização por tópico e à organização semântica de documentos em inglês, o intervalo interquartil da BOW é, em geral, maior do que o intervalo interquartil das representações semânticas. Além disso, a mediana das acurácias obtidas com cada representação proposta (gráficos identificados por (5) *NASARI+Babel2Vec* e (6) *Babel2Vec*) são mais altas do que as medianas das outras representações avaliadas no caso dos *datasets* que correspondem à organização semântica da coleção de documentos em inglês (Figuras 29d e Figure 29e). No caso da organização semântica da coleção de documentos em português (Figuras 29f e 29g), os resultados das representações semânticas apresentam intervalo interquartil semelhante ao intervalo da representação BOW.

Conforme apresentado na Seção 2.2.2 (Página 44), a medida acurácia corresponde à abordagem *micro-averaging* para sumarização das medidas precisão, revocação e F1 em problemas multiclasse. Na Figura 30, é apresentada a relação entre essas medidas para cada representação. Pode-se perceber que todas as representações apresentam comportamento semelhante. Na Tabela 8 é apresentada a correlação de Pearson entre $F1^\mu$ e $F1^M$ para as representações e combinações de representações avaliadas. Pode-se perceber que os resultados de $F1^\mu$, que corresponde à medida Acurácia, apresentam correlação positiva com $F1^M$.

Como as medidas $F1^\mu$ e $F1^M$ são bem correlacionadas e apresentam comportamento semelhante considerando as diferentes representações e *datasets*, no restante desta seção as perguntas estabelecidas no início são discutidas considerando, principalmente, os resultados obtidos com a medida acurácia ($F1^\mu$). No entanto, também são apresentados os resultados de comparações entre as representações considerando a medida $F1^M$. Resultados mais detalhados para a medida $F1^M$ são apresentados no Apêndice C (Seção C.2, Página 198).

Figura 29 – Box plots das acurácias de classificação, por representação utilizada, para cada *dataset*

Fonte: Elaborada pelo autor.

Nota – Para cada *dataset*, os gráficos das representações são apresentados na seguinte ordem: (1) BOW; (2) LDA; (3) Word2Vec; (4) Doc2Vec; (5) *NASARI+Babel2Vec*; (6) *Babel2Vec*.

Figura 30 – Relação entre $F1^{\mu}$ e $F1^M$ por representação utilizada **Legenda**

- BBC
- SE-polarity
- SE-product
- SE-product-polarity
- BS-semantic
- BS-topic
- BS-topic-semantic

Fonte: Elaborada pelo autor.

Tabela 8 – Correlação de Pearson entre $F1^\mu$ e $F1^M$

	<i>BBC</i>	<i>SE-</i> <i>-product</i>	<i>SE-</i> <i>-polarity</i>	<i>SE-product-</i> <i>-polarity</i>	<i>BS-</i> <i>-topic</i>	<i>BS-</i> <i>-semantic</i>	<i>BS-topic-</i> <i>-semantic</i>
BOW	0,9781	0,9653	0,7966	0,9670	0,9995	0,9842	0,9717
LDA	0,9994	0,9661	0,7016	0,9579	0,9990	0,9853	0,9640
Word2Vec	0,9995	0,9161	0,8601	0,9322	0,9994	0,9925	0,9518
Doc2Vec	0,9966	0,8874	0,8054	0,8921	0,9961	0,9803	0,9726
<i>NASARI+Babel2Vec</i>	0,9998	0,9206	0,8724	0,9264	0,9974	0,9805	0,9700
<i>Babel2Vec</i>	0,9995	0,9260	0,8590	0,9304	0,9984	0,9835	0,9570
BOW + <i>NASARI+Babel2Vec</i> - SC	0,9843	0,9462	0,8626	0,9479	0,9990	0,9849	0,9655
BOW + <i>NASARI+Babel2Vec</i> - WSC	0,9837	0,9401	0,8631	0,9451	0,9986	0,9849	0,9653
BOW + <i>NASARI+Babel2Vec</i> - MC	0,9832	0,9463	0,8567	0,9491	0,9984	0,9831	0,9700
BOW + <i>Babel2Vec</i> - SC	0,9843	0,9493	0,8560	0,9479	0,9994	0,9869	0,9633
BOW + <i>Babel2Vec</i> - WSC	0,9838	0,9430	0,8572	0,9448	0,9991	0,9866	0,9634
BOW + <i>Babel2Vec</i> - MC	0,9834	0,9494	0,8454	0,9498	0,9990	0,9847	0,9640

Representações baseadas em *embeddings* de termos desambiguados podem levar a melhores resultados de classificação?

A [Tabela 9](#) apresenta a melhor acurácia obtida por cada algoritmo de aprendizado entre todos os parâmetros testados. A melhor acurácia foi avaliada com o objetivo de considerar o melhor caso de cada estratégia de aprendizado. A penúltima linha apresenta o *ranking* médio das representações segundo o teste de Friedman, considerando as 728 configurações experimentais testadas (104 variações de algoritmos e parâmetros e 7 *datasets*). A última linha apresenta a posição das representações de acordo com esse *ranking*.

Considerando os melhores valores de acurácia alcançados pelos algoritmos de AM, as representações propostas apresentaram resultados melhores nas coleções de textos em inglês do que na coleção de textos em português. Para essa última coleção (*BS-Top4*), a melhor acurácia obtida com o uso da BOW foi maior do que as melhores acurácias obtidas com as representações semânticas para a maioria dos algoritmos considerados. Em poucos casos o uso de representações semânticas foi melhor do que o uso da BOW. O caso em que o uso de representação semântica teve a maior diferença na acurácia para a coleção *BS-Top4* foi a aplicação da representação *NASARI+Babel2Vec* com o algoritmo NB para o *dataset BS-topic-semantic*. Nesse caso, a melhor acurácia obtida foi de 62,5246% com o uso de *NASARI+Babel2Vec*, enquanto que o melhor valor obtido com a BOW foi de 57,2537%.

Considerando as coleções de textos em inglês, as representações semânticas superaram as acurácias obtidas com a BOW na maioria das configurações testadas. As maiores diferenças foram encontradas nos *datasets* de maior complexidade semântica, ou seja, (*SE-polarity* e *SE-product-polarity*). Para esses *datasets*, a melhor acurácia obtida com *NASARI+Babel2Vec* foi maior do que a maior acurácia obtida com a BOW em 7 das 12 configurações testadas e foi maior do que a maior acurácia obtida com *Babel2Vec* também em 7 das 12 configurações. Já a melhor acurácia obtida com *Babel2Vec* foi maior do que acurácia obtida com a BOW em 8 das 12 configurações testadas.

Ainda para os *datasets SE-polarity* e *SE-product-polarity*, considerando-se apenas as acurácias mais altas de cada *dataset* (linhas em cinza da [Tabela 9](#)), a representação Word2Vec

Tabela 9 – Melhores acurácias para as coleções *BBC*, *SE-ABSA15* e *BS-Top4*

	<i>NASARI+Babel2Vec</i>	<i>Babel2Vec</i>	BOW	LDA	Word2Vec	Doc2Vec
BBC	97,3040	97,6187	96,9446	97,1252	98,0227	97,0789
IMBHN ^C	95,5510	96,2247	95,8207	94,7418	96,9884	92,3605
IMBHN ^R	94,1124	95,7300	96,9446	97,1252	95,7753	93,7983
J48	86,7430	85,8886	86,1144	<u>82,3357</u>	85,6167	53,0770
k-NN	95,8653	96,5841	95,5525	95,8227	96,8988	91,4606
NB	93,4388	95,2816	92,9439	90,0238	95,2367	95,9104
SVM	97,3040	97,6187	96,4950	96,7642	98,0227	97,0789
SE-product	99,2623	99,3857	99,1388	98,2806	99,2623	98,4011
IMBHN ^C	97,3020	99,2608	98,1587	96,1879	98,5230	96,6862
IMBHN ^R	95,9515	97,0506	<u>99,1388</u>	97,5444	96,0690	95,3327
J48	90,6745	89,3315	<u>92,2704</u>	90,7995	89,4460	76,9663
k-NN	98,4026	99,0169	<u>98,5245</u>	97,5444	98,5230	95,7016
NB	96,1939	98,4041	92,7612	86,7480	95,0918	97,9103
SVM	99,2623	99,3857	96,4408	98,2806	99,2623	98,4011
SE-polarity	85,7588	84,6507	82,8214	80,3704	86,8669	82,3427
IMBHN ^C	78,0307	79,6341	80,4908	69,5649	80,4983	72,1515
IMBHN ^R	82,8169	83,5531	82,8214	80,3704	83,0699	78,4146
J48	69,3270	<u>68,8181</u>	<u>71,5266</u>	68,7097	70,7934	65,1581
k-NN	81,7194	80,1265	77,2900	78,0337	79,6387	75,0903
NB	74,0921	71,5116	70,3071	50,9244	74,1042	63,5622
SVM	85,7588	84,6507	81,6110	79,7546	86,8669	82,3427
SE-product-polarity	81,4694	83,0638	77,8034	77,4164	83,5577	80,6173
IMBHN ^C	77,5534	79,7651	77,4345	73,3740	80,0000	69,8148
IMBHN ^R	66,7450	67,7251	73,9762	74,5980	<u>66,6245</u>	66,6245
J48	60,2394	59,9864	71,0479	<u>62,5790</u>	59,2578	49,9383
k-NN	80,1235	79,5047	75,8220	76,0690	76,1939	72,8847
NB	75,2213	70,9214	68,9521	49,9398	63,1963	69,0741
SVM	81,4694	83,0638	77,8034	77,4164	83,5577	80,6173
BS-topic	99,6429	100,0000	100,0000	100,0000	100,0000	100,0000
IMBHN ^C	97,1675	99,2980	98,9286	98,9286	97,5000	96,8103
IMBHN ^R	97,5246	98,5837	99,6429	98,9286	98,9286	97,8695
J48	80,8621	85,1601	<u>96,8227</u>	90,4064	96,1084	64,7167
k-NN	99,6429	100,0000	<u>99,6552</u>	100,0000	100,0000	99,2857
NB	96,0961	<u>96,4655</u>	<u>99,6429</u>	98,5837	97,5000	99,2857
SVM	99,6429	100,0000	100,0000	98,9286	99,6429	100,0000
BS-semantic	65,3818	65,4187	68,9532	61,4655	62,9557	57,6478
IMBHN ^C	56,2192	54,7906	<u>64,6552</u>	57,2783	55,1478	49,8892
IMBHN ^R	58,3005	63,3128	<u>68,9532</u>	61,4655	59,0148	50,2217
J48	49,4951	41,0222	<u>59,0517</u>	47,6847	52,6724	34,7044
k-NN	60,1108	61,4778	<u>65,3818</u>	60,7882	60,4926	54,4951
NB	50,5296	52,3153	<u>57,6108</u>	51,2562	45,2586	50,9113
SVM	65,3818	65,4187	63,6576	59,1010	62,9557	57,6478
BS-topic-semantic	65,7266	66,1084	66,8596	60,4680	63,6453	63,6700
IMBHN ^C	55,8744	58,6946	62,5739	59,0517	62,9187	53,7685
IMBHN ^R	43,0911	48,3867	<u>57,9926</u>	55,5049	48,4113	40,3079
J48	38,9163	35,2833	<u>55,1478</u>	45,9113	44,1872	22,9926
k-NN	58,6823	61,1453	<u>65,7512</u>	60,4310	60,4926	53,4113
NB	62,5246	59,7044	57,2537	48,7685	57,5739	54,8645
SVM	65,7266	66,1084	<u>66,8596</u>	60,4680	63,6453	63,6700
<i>Ranking médio</i>	2,9815	2,4478	3,8427	4,2150	2,8784	4,6346
<i>Posição pelo ranking médio</i>	3º	1º	4º	5º	2º	6º

Nota – Acurácias maiores que as obtidas com o *baseline* BOW (coluna cinza) são apresentadas em negrito.

A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia de cada representação para o respectivo *dataset*.

obtem os melhores resultados. No entanto, as representações propostas são mais estáveis em relação a alterações nos algoritmos de aprendizado. *NASARI+Babel2Vec* e *Babel2Vec* apresentaram medianas mais altas do que Word2Vec para esses *datasets* (Figure 29d e Figure 29e).

Os 728 resultados da avaliação experimental foram submetidos ao teste de Friedman $N \times N$ com pós-teste de Nemenyi, que rejeitou a hipótese nula (hipótese de que não há diferença estatisticamente significativa entre as representações), com nível de confiança de 95% e $p\text{-value} < 2.2 \times 10^{-16}$. Os *rankings* médios, apresentados na [Tabela 9](#), colocam a representação *Babel2Vec* na primeira posição do *ranking* e *NASARI+Babel2Vec* na terceira posição. O diagrama de diferença crítica dessa análise é apresentado na [Figura 31](#).

Considerando todas as configurações individualmente, *Babel2Vec* apresentou a melhor acurácia em 189 (259 considerando-se os empates com outras representações) das 728 configurações. Os resultados indicaram que o uso de *Babel2Vec* é adequado para o algoritmo k-NN na classificação de produtos em textos de opinião (*dataset SE-product*). Entre as 56 configurações do k-NN para *SE-product*, *Babel2Vec* foi a melhor representação em 52 configurações. Considerando o algoritmo SVM, *Babel2Vec* também foi a melhor representação para a classificação de produtos em textos de opinião. *Babel2Vec* foi a melhor representação em 34 de 36 configurações testadas para *SE-product* com o algoritmo SVM. Considerando-se os 7 *datasets*, *Babel2Vec* foi a melhor representação em 49 (111 considerando-se empates com outras representações) das 252 configurações testadas com o algoritmo SVM. *Babel2Vec* foi a segunda melhor representação para ser usada em conjunto com SVM. A melhor representação para o algoritmo SVM foi a Word2Vec, sendo a primeira no *ranking* em 74 (135 considerando-se empates) das 252 configurações testadas.

A outra representação proposta, *NASARI+Babel2Vec*, foi a melhor representação em 105 (166 considerando-se empates) das 728 configurações testadas. Os resultados indicaram que essa representação é adequada para ser usada com o algoritmo NB para os *datasets* que dependem tanto da informação de tópico quanto da informação semântica. Esse é o caso dos *datasets SE-product-polarity* e *BS-topic-semantic*. Além disso, *NASARI+Babel2Vec* também se mostrou adequada para o uso com o algoritmo k-NN nos *datasets* de organização semântica da coleção em inglês (*SE-product-polarity* e *SE-polarity*). Entre as 56 configurações do algoritmo k-NN, *NASARI+Babel2Vec* obteve o melhor resultado em 33 configurações para *SE-product-polarity* e 43 para *SE-polarity*.

O diagrama de diferença crítica, apresentado na [Figura 31](#), ilustra as diferenças entre as representações. Nesse diagrama, as representações são apresentadas de acordo com o *ranking* médio obtido considerando-se todas as configurações testadas. As representações conectadas por uma linha não apresentam diferenças estatisticamente significantes entre si. Pode-se notar que *Babel2Vec* se apresentou como a melhor representação, com diferença estatisticamente significantes para todas as demais representações. Já *NASARI+Babel2Vec* é a terceira no *ranking*, porém sem diferença estatisticamente significativa em relação à segunda representação (Word2Vec). Além disso, as duas representações propostas são superiores, com diferença estatisticamente significativa, às representações BOW, LDA e Doc2Vec. O mesmo ocorre ao se considerar a medida $F1^M$, como pode ser observado no diagrama de diferença crítica da [Figura 32](#).

Figura 31 – Diagrama de diferença crítica entre as representações avaliadas, considerando a medida Acurácia

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,2796$

Figura 32 – Diagrama de diferença crítica entre as representações avaliadas, considerando a medida $F1^M$

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,2796$

Os bons resultados obtidos com as representações geradas a partir de *embeddings* quando comparadas a BOW e LDA indicam que essas representações conseguem incorporar conhecimento obtido dos grandes corpora usados para treinar as *embeddings*. Padrões descobertos nesses corpora contribuem para uma melhor representação do conteúdo das coleções de documentos. Enquanto as representações baseadas em *embeddings* são enriquecidas com esse conhecimento externo, as representações BOW e LDA são construídas apenas com base no próprio conteúdo dos documentos.

Outro ponto importante a se considerar é a estrutura das representações. As representações semânticas comparadas nessa avaliação experimental apresentam uma dimensionalidade fixa. No caso desses experimentos, os documentos são representados por vetores com 300 dimensões. Já a dimensionalidade da BOW varia de 1.312 a 11.184 dimensões, dependendo da coleção de textos. A baixa dimensionalidade tem um impacto positivo no tempo requerido para o treinamento dos modelos de classificação.

A combinação de representações baseadas em *embeddings* de termos desambiguados com a BOW leva a melhores resultados de classificação?

Visto os resultados promissores do uso combinado da BOW com uma representação semanticamente enriquecida apresentados no Capítulo 4, neste trabalho também foi avaliada a combinação de classificadores gerados com a BOW e as representações propostas, *Babel2Vec* e *NASARI+Babel2Vec*. As distribuições de valores de acurácia obtidos para a combinação de

classificadores gerados com BOW e *NASARI+Babel2Vec* são apresentadas nas Figuras 33, 34 e 35. Já os resultados da combinação de BOW e *Babel2Vec* são apresentados nas Figuras 36, 37 e 38. Nessas figuras, para cada *dataset* são apresentados 5 gráficos, correspondentes à variação de peso do voto do classificador gerado com a BOW (pesos de 0,1 a 0,9).

Pode-se perceber que, independente da representação semântica utilizada, as distribuições das acurácias obtidas são bem semelhantes. Em relação aos pesos dados aos classificadores base, observa-se que, para as coleções de textos em inglês, os melhores resultados são obtidos com pesos menores para o classificador gerado com a BOW. Já para a coleção de textos em português, no geral, observa-se os melhores resultados quando o classificador gerado com a BOW tem o peso maior.

Na Tabela 10 são apresentados os maiores valores de acurácia obtidos pelas combinações de BOW com as duas representações semânticas propostas. Pode-se notar que, no geral, a combinação de classificadores obtém resultado melhor do que os classificadores base (resultados apresentados na Tabela 9). As maiores diferenças foram obtidas para os *datasets* que dependem tanto da informação de tópico quanto de informação semântica (*SE-product-polarity* e *BS-topic-semantic*) e o uso do algoritmo baseado em redes IMBHN^C. Para o caso de *SE-product-polarity* com a combinação de BOW e *NASARI+Babel2Vec*, a acurácia aumentou de 77,5534% (classificador gerado com *NASARI+Babel2Vec*) para 81,9723% (estratégias SC e WSC). Para o mesmo *dataset*, a combinação de BOW e *Babel2Vec* resultou em um aumento de acurácia de 79,7651% (classificador gerado com *Babel2Vec*) para 82,9464% (estratégia SC). Para o caso de *BS-topic-semantic* os aumentos foram um pouco maiores. A combinação de BOW e *NASARI+Babel2Vec* levou a um aumento de 62,5739% (classificador gerado com BOW) para 67,5369% (estratégias SC e WSC) e a combinação de BOW e *Babel2Vec* levou a um aumento para 66,4778% (estratégias SC e WSC).

Na Figura 39 são apresentados os diagramas de diferença crítica considerando as melhores representações avaliadas (*Babel2Vec*, *Word2Vec*, *NASARI+Babel2Vec* e BOW) e as combinações de BOW com *NASARI+Babel2Vec*. Para essa análise, realizou-se o teste de Friedman $N \times N$ com pós-teste de Nemenyi, que rejeitou a hipótese nula (hipótese de que não há diferença estatisticamente significativa entre as representações), com nível de confiança de 95% e $p\text{-value} < 2.2 \times 10^{-16}$. Nos diagramas da Figura 39, as representações são apresentadas de acordo com o *ranking* médio obtido considerando-se todas as 728 configurações testadas. As representações conectadas por uma linha não apresentam diferenças estatisticamente significantes entre si. Pode-se perceber que as melhores combinações são com a estratégia WSC e pesos de 0,3 ou 0,5 para o classificador gerado com BOW. Com exceção dessas duas configurações, as demais combinações de classificadores gerados com BOW e *NASARI+Babel2Vec* não apresentam diferença estatisticamente significativa para classificadores gerados com a representação *Babel2Vec*. Ao se considerar os pesos de 0,1 ou 0,9 para o classificador da BOW (Figuras 39a e 39e), a representação *Babel2Vec* é a primeira no *ranking*. Considerando-se a medida $F1^M$, conforme

Figura 33 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *NASARI+Babel2Vec*, utilizando a estratégia SC

Fonte: Elaborada pelo autor.

Figura 34 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e NASARI+Babel2Vec, utilizando a estratégia WSC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 35 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *NASARI+Babel2Vec*, utilizando a estratégia MC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 36 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *Babel2Vec*, utilizando a estratégia SC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 37 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *Babel2Vec*, utilizando a estratégia WSC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 38 – *Box plots* das acurácias de cada *dataset* para a combinação de BOW e *Babel2Vec*, utilizando a estratégia MC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Tabela 10 – Melhores acurácias para as coleções *BBC*, *SE-ABSA15* e *BS-Top4* para combinação de classificadores

	BOW + <i>NASARI+Babel2Vec</i>			BOW + <i>Babel2Vec</i>		
	SC	WSC	MC	SC	WSC	MC
BBC	97,5284	97,5286	97,3040	97,8883	97,8883	97,7534
IMBHN ^C	<u>97,4387</u>	<u>97,5286</u>	<u>97,0802</u>	<u>97,6179</u>	<u>97,6179</u>	<u>97,4381</u>
IMBHN ^R	<u>97,1698</u>	<u>97,2147</u>	96,9446	<u>97,2147</u>	<u>97,2147</u>	96,9446
J48	<u>87,3272</u>	<u>88,7195</u>	<u>87,5969</u>	<u>87,0119</u>	<u>88,0008</u>	<u>86,6968</u>
k-NN	<u>96,8545</u>	<u>96,8095</u>	<u>96,4503</u>	<u>96,8539</u>	<u>96,8539</u>	<u>96,6741</u>
NB	<u>94,7421</u>	<u>93,5285</u>	<u>95,0574</u>	95,2816	95,2816	95,2816
SVM	<u>97,5284</u>	<u>97,5284</u>	97,3040	<u>97,8883</u>	<u>97,8883</u>	<u>97,7534</u>
SE-product	99,5077	99,3857	99,3842	99,5077	99,5077	99,5062
IMBHN ^C	<u>99,3857</u>	<u>99,3857</u>	<u>99,0214</u>	<u>99,5077</u>	<u>99,5077</u>	<u>99,5062</u>
IMBHN ^R	99,1388	99,1388	99,1388	99,1388	99,1388	99,1388
J48	<u>93,0081</u>	92,2704	<u>92,5188</u>	92,2704	92,2704	92,2704
k-NN	<u>99,5077</u>	<u>99,2623</u>	<u>99,2638</u>	<u>99,3827</u>	<u>99,2623</u>	<u>99,2653</u>
NB	<u>96,4393</u>	<u>97,0521</u>	<u>97,5414</u>	98,4041	<u>98,7699</u>	<u>98,5230</u>
SVM	<u>99,3842</u>	<u>99,3842</u>	<u>99,3842</u>	99,3857	99,3857	99,3857
SE-polarity	85,7588	85,7588	85,7588	85,3914	85,3914	84,6507
IMBHN ^C	<u>82,5745</u>	<u>82,6980</u>	<u>82,4511</u>	<u>81,5959</u>	<u>81,3535</u>	<u>81,1006</u>
IMBHN ^R	<u>84,9021</u>	<u>84,9036</u>	<u>84,7787</u>	<u>85,3914</u>	<u>85,3914</u>	<u>84,1644</u>
J48	<u>73,6104</u>	<u>73,6104</u>	<u>71,6471</u>	<u>73,1180</u>	<u>73,3619</u>	71,5266
k-NN	<u>81,8383</u>	<u>81,9633</u>	<u>82,0837</u>	<u>81,7148</u>	<u>81,7148</u>	<u>81,0976</u>
NB	<u>74,5845</u>	<u>74,7079</u>	74,0921	<u>72,6227</u>	<u>72,6227</u>	71,5116
SVM	85,7588	85,7588	85,7588	84,6507	84,6507	84,6507
SE-product-polarity	82,5821	82,4556	81,8428	82,9464	82,9419	83,0638
IMBHN ^C	<u>81,9723</u>	<u>81,9723</u>	<u>79,8901</u>	<u>82,9464</u>	<u>82,8245</u>	<u>81,2376</u>
IMBHN ^R	73,7278	73,9732	73,9762	73,9717	73,9717	73,9762
J48	<u>71,1713</u>	71,0479	71,0479	<u>71,0494</u>	<u>71,0494</u>	71,0479
k-NN	<u>81,0991</u>	<u>80,9771</u>	<u>80,9741</u>	<u>80,6067</u>	<u>80,6067</u>	<u>79,8660</u>
NB	<u>75,7121</u>	<u>75,8341</u>	<u>76,6787</u>	<u>72,3893</u>	<u>71,4077</u>	<u>73,6179</u>
SVM	<u>82,5821</u>	<u>82,4556</u>	<u>81,8428</u>	82,9419	82,9419	83,0638
BS-topic	100,0000	100,0000	100,0000	100,0000	100,0000	100,0000
IMBHN ^C	<u>99,2980</u>	<u>99,2980</u>	<u>99,2857</u>	<u>99,6552</u>	<u>99,6552</u>	99,2980
IMBHN ^R	99,6429	99,6429	99,6429	<u>100,0000</u>	<u>100,0000</u>	<u>100,0000</u>
J48	96,8227	96,8227	96,8227	96,8227	96,8227	96,8227
k-NN	<u>100,0000</u>	<u>100,0000</u>	<u>100,0000</u>	100,0000	100,0000	100,0000
NB	99,6429	99,6429	99,6429	99,6429	99,6429	99,6429
SVM	100,0000	100,0000	100,0000	100,0000	100,0000	100,0000
BS-semantic	68,5961	68,5961	68,9532	69,3103	69,6675	68,9532
IMBHN ^C	<u>65,0123</u>	<u>65,0123</u>	64,6552	<u>67,5123</u>	<u>67,5123</u>	<u>65,3695</u>
IMBHN ^R	68,5961	68,5961	68,9532	<u>69,3103</u>	<u>69,6675</u>	68,9532
J48	<u>59,7537</u>	<u>59,7537</u>	59,0517	<u>59,3966</u>	<u>59,3966</u>	59,0517
k-NN	<u>66,4409</u>	<u>65,7266</u>	<u>65,7266</u>	<u>66,0961</u>	<u>66,4532</u>	<u>65,4187</u>
NB	57,6108	57,6108	<u>58,6453</u>	<u>58,3251</u>	<u>57,9680</u>	<u>59,3719</u>
SVM	<u>66,0961</u>	<u>66,0837</u>	<u>66,4286</u>	<u>67,1798</u>	<u>67,1798</u>	<u>67,5246</u>
BS-topic-semantic	67,5369	67,5369	66,8596	67,8695	67,8695	67,1921
IMBHN ^C	<u>67,5369</u>	<u>67,5369</u>	<u>63,9532</u>	<u>66,4778</u>	<u>66,4778</u>	<u>63,5714</u>
IMBHN ^R	57,2783	57,2783	57,9926	55,8498	56,9212	57,9926
J48	<u>55,5049</u>	<u>55,5049</u>	55,1478	55,1478	55,1478	55,1478
k-NN	<u>66,0961</u>	65,7512	65,7512	<u>66,0961</u>	<u>66,4532</u>	65,7512
NB	<u>63,2266</u>	<u>63,2266</u>	<u>63,2143</u>	<u>61,1084</u>	59,7044	<u>60,4187</u>
SVM	<u>67,2291</u>	<u>67,2291</u>	66,8596	<u>67,8695</u>	<u>67,8695</u>	<u>67,1921</u>

Nota – Valores em negrito correspondem a acurácias maiores do que as acurácias dos respectivos classificadores base (resultados apresentados na Tabela 9). A melhor acurácia de cada linha está sublinhada. As linhas em cinza correspondem aos melhores valores de acurácia para o respectivo dataset.

apresentado na Figura 40, observa-se esse mesmo comportamento. Além disso, analisando-se a $F1^M$, apenas a estratégia WSC com peso de 0,5 apresenta resultado superior aos demais.

Figura 39 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e *NASARI+Babel2Vec* - Acurácia

(a) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,1 para BOW

(b) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,3 para BOW

(c) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,5 para BOW

(d) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,7 para BOW

(e) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,9 para BOW

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

Figura 40 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e *NASARI+Babel2Vec* - $F1^M$

(a) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,1 para BOW

(b) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,3 para BOW

(c) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,5 para BOW

(d) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,7 para BOW

(e) Combinações de BOW e *NASARI+Babel2Vec* com peso 0,9 para BOW

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

A mesma análise foi realizada considerando-se a combinação de classificadores gerados com BOW e *Babel2Vec*. Nessa análise, o teste de Friedman $N \times N$ com pós-teste de Nemenyi também rejeitou a hipótese nula, com nível de confiança de 95% e $p\text{-value} < 2.2 \times 10^{-16}$. Os diagramas de diferença crítica considerando as melhores representações avaliadas (*Babel2Vec*, Word2Vec, *NASARI+Babel2Vec* e BOW) e as combinações de BOW com *Babel2Vec* são apresentados na Figura 41. Ao se considerar pesos iguais para os classificadores gerados com a BOW e com a *Babel2Vec* (Figura 41c), as três estratégias de combinação testadas apresentaram resultado superior aos demais com diferenças estatisticamente significantes. Já ao se considerar pesos menores para o classificador da BOW as combinações WSC e SC apresentam resultado superior aos demais (Figuras 41a e 41b). Para a medida $F1^M$, diagramas apresentados na Figura 42, pode ser observado um comportamento bem semelhante. As três estratégias de combinação apresentam resultados superiores aos demais ao se considerar pesos iguais aos classificadores. Tanto ao se analisar a acurácia quanto a $F1^M$, as combinações com peso de 0,9 para o voto do classificador gerado com a BOW (Figuras 41e e 42e) apresentam resultado inferior ao do classificador gerado com *Babel2Vec*, (representação com melhores resultados individuais).

Para analisar em conjunto a combinação de classificadores gerados com a BOW e classificadores gerados com as representações semânticas propostas (*NASARI+Babel2Vec* e *Babel2Vec*), comparou-se os resultados das quatro representações com melhores resultados individuais (*Babel2Vec*, Word2Vec, *NASARI+Babel2Vec* e BOW) e as melhores combinações (BOW + *Babel2Vec* - WSC_01, WSC_03 e WSC_05). Os diagramas de diferença crítica para as medidas Acurácia e $F1^M$ são apresentados, respectivamente, nas Figuras 43 e 44. Segundo os testes realizados, as estratégias de combinação apresentam diferença estatisticamente significativa em relação a todas as representações utilizadas individualmente. Além disso, a combinação que se apresentou como sendo a mais adequada é a combinação de classificadores gerados com BOW e *Babel2Vec*, utilizando a estratégia WSC e o peso 0,3 para o classificador da BOW.

5.7 Considerações finais

Nesse capítulo foi apresentado o trabalho realizado com o objetivo de avaliar o uso de *embeddings* na representação de coleções de documentos. Foram propostos três modelos de representação, *NASARI_docs*, *Babel2Vec* e *NASARI+Babel2Vec*, construídos com base em desambiguação lexical de sentidos e representações vetoriais de palavras e de sentidos de palavras. Tais representações podem ser facilmente geradas a partir de *embeddings* pré-treinadas.

As representações propostas apresentam as vantagens de serem projetadas no mesmo espaço das *embeddings* e de não requererem um grande volume de documentos para serem construídas. As diferenças da abordagem adotada neste trabalho para outros modelos de representação baseados em *word embeddings*, como a representação Word2Vec avaliada neste capítulo, são o uso de desambiguação lexical de sentidos e o uso de *NASARI embedded*, para o caso das

Figura 41 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e *Babel2Vec* - Acurácia

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

Figura 42 – Diagrama de diferença crítica entre as melhores representações avaliadas e as combinações de BOW e *Babel2Vec* - $F1^M$

(a) Combinações de BOW e *Babel2Vec* com peso 0,1 para BOW

(b) Combinações de BOW e *Babel2Vec* com peso 0,3 para BOW

(c) Combinações de BOW e *Babel2Vec* com peso 0,5 para BOW

(d) Combinações de BOW e *Babel2Vec* com peso 0,7 para BOW

(e) Combinações de BOW e *Babel2Vec* com peso 0,9 para BOW

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

Figura 43 – Diagrama de diferença crítica entre as melhores as representações e combinações avaliadas - Acurácia

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

Figura 44 – Diagrama de diferença crítica entre as melhores as representações e combinações avaliadas - $F1^M$

Fonte: Elaborada pelo autor.

Nota – Diferença crítica, segundo pós-teste de Nemenyi: $CD = 0,3340$.

representações $NASARI_{docs}$ e $NASARI+Babel2Vec$. Com isso, essa abordagem pode ser aplicada a textos de qualquer idioma.

Uma análise dos vetores que representam os documentos indicaram que as representações propostas geram vetores para os documentos próximos a vetores de palavras e/ou sentidos relacionados ao conteúdo dos documentos. A vantagem de $NASARI_{docs}$ e $NASARI+Babel2Vec$ é que essas representações são geradas a partir dos vetores de sentidos $NASARI_{embedded}$ e, portanto, é possível utilizar os sentidos mais próximos ao documento para interpretar a sua representação. A desvantagem de $NASARI_{docs}$ é que ela é limitada a apenas a representação de substantivos, o que pode impossibilitar a representação de determinados documentos, como por exemplo, textos de opinião contendo apenas adjetivos.

As representações $Babel2Vec$ e $NASARI+Babel2Vec$ se mostraram representações adequadas para a identificação de similaridade textual. Considerando o *dataset* de *benchmarking* de Lee, Pincombe e Welsh (2005), as representações propostas apresentaram boas correlações com similaridades atribuídas por pessoas. A correlação obtida com $Babel2Vec$ foi mais alta do que a correlação calculada entre os anotadores do *dataset* de Lee, Pincombe e Welsh (2005).

As representações propostas também foram avaliadas na classificação de textos em comparação a outros quatro modelos de representação: BOW, LDA, Word2Vec e Doc2Vec.

Assim como no capítulo anterior, a avaliação foi realizada com *datasets* de diferentes níveis de complexidade semântica. Os resultados dessa avaliação experimental indicam que os modelos propostos apresentam resultados competitivos, especialmente em cenários de maior complexidade semântica de textos em inglês. Para a coleção de textos em português, as representações semânticas não alcançaram, no geral, resultados melhores do que a BOW. Um motivo para esse fato pode ser a cobertura para o idioma português dos recursos linguísticos utilizados. Esse ainda é um desafio para a Mineração de Textos do idioma português. No entanto, os bons resultados obtidos para os textos em inglês sugerem que o uso das representações baseadas em *embeddings* são uma boa direção para futuros melhoramentos na representação de textos em português.

CONCLUSÕES

6.1 Contribuições científicas

As contribuições deste trabalho estão relacionadas às questões de pesquisa e aos objetivos apresentados na introdução desta tese. Essas questões são sintetizadas nesta seção, apresentando os resultados obtidos e as abordagens propostas e desenvolvidas.

Questão Q1 (Q1.1 a Q1.7): *Como a semântica é tratada no processo de Mineração de Textos?*

Quais são os domínios de aplicação que consideram a semântica dos textos de alguma maneira? Quais idiomas estão sendo tratados? Quais fontes externas são comumente utilizadas? Em quais tarefas a semântica aparece com mais frequência? Quais métodos e algoritmos são comumente utilizados? Como os textos são representados? Os usuários participam do processo de Mineração de Textos?

Com o objetivo de mapear os trabalhos realizados na comunidade de Mineração de Textos que consideram a semântica dos textos, foi conduzida uma revisão sistemática da literatura desse amplo tema. Esse mapeamento sistemático da literatura seguiu um protocolo bem definido, realizando a busca pelos estudos nas principais bibliotecas digitais e fontes de pesquisa da área de computação. A aplicação da expressão de busca nas cinco fontes de pesquisa resultou na identificação de 3.984 estudos, dos quais 1.693 estudos foram selecionados após verificação dos critérios de exclusão estabelecidos. Foram executados dois ciclos das fases de mapeamento sistemático, sendo que o primeiro ciclo refere-se a buscas realizadas em janeiro de 2014 e o segundo, correspondente a uma atualização do mapeamento, refere-se a buscas realizadas em fevereiro de 2016.

Os resultados obtidos com esse mapeamento confirmaram alguns indicativos prévios dos pesquisadores sobre o tema de pesquisa e apontaram outras características interessantes da área. Em relação aos domínios de aplicação (Q1.1), verificou-se a predominância de domínios

de saúde e ciências biológicas e de documentos da Web (páginas *web*, *blogs*, fóruns, redes sociais e filtragem de *e-mails*). Ao se considerar os idiomas que estão sendo tratados (Q1.2), além do idioma inglês, encontrou-se grande presença do idioma chinês nos estudos considerados. Em relação às fontes externas (Q1.3), verificou-se que WordNet e Wikipedia são as fontes externas mais utilizadas, seguidas da base de conhecimento chinesa HowNet. Considerando-se as tarefas de MT (Q1.4), a análise de sentimentos se destacou, visto que é uma área mais recente e aparece com frequência semelhante a áreas mais consolidadas como extração e recuperação de informação. Entre os métodos e algoritmos utilizados nestes estudos (Q1.5), destaca-se o tratamento da semântica latente e a aplicação de algoritmos tradicionais de AM. Em relação à representação dos textos (Q1.6), verificou-se a busca por representações no modelo espaço-vetorial semanticamente enriquecidas, seja com a construção de espaços de semântica latente ou a aplicação de métodos de PLN, além de modelos de representação baseados em redes. Os resultados do mapeamento indicam que a participação do usuário no processo de MT (Q1.7) é pouco explorada em artigos científicos, provavelmente devido à dificuldade inerente na avaliação de métodos que consideram a interação com usuários. Nesta tese, esse trabalho foi apresentado no Capítulo 2, Seção 2.4 - Página 52.

Questões Q2 e Q3: *Que tipo de problema é tradicionalmente tratado pela comunidade de pesquisa? Como a semântica interage com os diferentes problemas de classificação existentes?*

Cenários reais de aplicação da classificação automática de textos são bastante diversos. Existem diferentes fontes de dados e diferentes objetivos para a aplicação do processo de Mineração de Textos. Conforme apresentado por meio de exemplos ao longo desta tese, o tratamento adequado da semântica dos textos pode ter um grande impacto nos resultados obtidos com a aplicação do processo de MT. No entanto, tradicionalmente o problema tratado pela comunidade de pesquisa é o problema em que as classes correspondem a tópicos ou assuntos dos textos. De acordo com a definição de Sebastiani (2002), trabalho bastante conceituado na área, a tarefa de classificação automática de textos está relacionada à tarefa de detecção de tópicos, com a rotulação de textos escritos em língua natural por meio da atribuição de uma categoria temática presente em uma lista de categorias pré-definidas. Para corroborar este fato e identificar os tipos de problemas que normalmente são tratados pela comunidade de MT foram analisadas as principais coleções de *benchmarking* utilizadas em pesquisas da área. O levantamento dessas coleções foi realizado com base nos artigos considerados no mapeamento sistemático da literatura apresentado anteriormente e a partir do trabalho de Rossi, Marcacini e Rezende (2013), que apresenta 45 coleções de documentos para pesquisas em classificação e agrupamento de textos. Durante a realização do mapeamento sistemático da literatura, foi identificado o uso de 211 coleções ou fontes de textos, sendo que 10 delas aparecem em mais de 46% dos estudos cujas coleções/fontes de documentos foram identificadas. Como 7 dessas 10 coleções estão representadas entre as 45 coleções apresentadas por Rossi, Marcacini e Rezende

(2013) e essas 45 coleções são bem definidas e com *datasets* disponibilizados para a comunidade, tais coleções foram analisadas. A partir dessa análise, verificou-se que, na grande maioria dos casos, as classes correspondem a tópicos ou assuntos.

Considerando-se os aspectos semânticos dos textos e os diferentes objetivos que pode-se ter ao realizar um processo de classificação automática, ou seja, diferentes naturezas de classes, neste trabalho definiu-se dois níveis de complexidade semântica para os problemas de organização de documentos. O primeiro nível, que é chamado de organização por tópico, consiste em problemas de classificação que dependem basicamente do vocabulário. Nesse problema, cada classe possui termos bastante característicos, e, portanto, o léxico (vocabulário) possui grande relevância para representar o conteúdo dos documentos. Pode-se dizer que os documentos podem ser diferenciados em grande parte pelas palavras utilizadas. O segundo nível de complexidade semântica engloba os demais problemas de classificação de documentos. Esse segundo nível é chamado de organização semântica, no sentido de que se necessita mais do que apenas o léxico para resolvê-lo. Tais problemas requerem uma análise mais profunda, além apenas das palavras, visto que os documentos de classes distintas podem usar o mesmo vocabulário.

Apesar da identificação de um grande número de pesquisas em Mineração de Textos envolvendo a semântica de alguma maneira, não se encontrou um trabalho que discutisse essas questões, comparando os diferentes problemas de classificação automática de documentos do ponto de vista da semântica. Neste trabalho, buscou-se trabalhar essa lacuna e foram preparados *datasets* diferentes para duas coleções de documentos, representando problemas dos dois níveis de complexidade semântica definidos. Tais *datasets* possibilitaram as avaliações experimentais realizadas neste trabalho e estão disponíveis à comunidade para pesquisas futuras. Nesta tese, esse trabalho foi apresentado principalmente no Capítulo 2, Seção 2.5 - Página 58.

Questão Q4 *A representação bag-of-words é adequada a qual tipo de problema de classificação?*

Com o objetivo de avaliar a performance de classificação obtida com o uso da representação *bag-of-words* foram realizadas avaliações experimentais utilizando os *datasets* que desenvolvidos para simular problemas de diferentes níveis de complexidade semântica para uma mesma coleção de documentos. Esses *datasets*, citados no item anterior, foram apresentados nesta tese no Capítulo 2, Seção 2.5.2 - Página 64. A avaliação experimental foi realizada com três *datasets* do primeiro nível de complexidade semântica (organização por tópico) e quatro *datasets* do segundo nível (organização semântica), considerando-se diferentes algoritmos de aprendizado supervisionado. Os resultados dessa avaliação experimental indicam que a representação *bag-of-words* funciona bem para os problemas de organização por tópico, porém é deficiente em problemas de organização semântica, considerando-se a performance de classificação dada pelas medidas Acurácia e $F1^M$. Por exemplo, considere a coleção de documentos *SE-ABSA15* e os melhores resultados obtidos com o uso da representação *bag-of-words* para os seis algoritmos

de aprendizado supervisionado utilizados na avaliação experimental apresentada no Capítulo 5. Quando o objetivo é a classificação por produto, as melhores acurácias obtidas com os diferentes algoritmos de aprendizado variam entre 92,2704% a 99,1388% (Tabela 9, Página 136) e os melhores valores de $F1^M$ variam entre 0,810296 a 0,943081 (Tabela 12, Página 201). Já quando o objetivo é a classificação por polaridade da revisão, as melhores acurácias obtidas variam entre 69,3270% a 85,7588% (Tabela 9, Página 136) e os melhores valores de $F1^M$ variam entre 0,468746 a 0,558758 (Tabela 12, Página 201). O mesmo padrão de performance de classificação mais baixa em problemas de organização semântica também foi observado para a coleção de documentos em português.

Questões Q5 e Q6 *Representações semanticamente enriquecidas podem levar a melhores resultados de classificação? A combinação de representações semânticas com a bag-of-words pode levar a melhores resultados de classificação?*

Com o objetivo de desenvolver modelos de representação de coleções de documentos com base em informações semânticas de diferentes naturezas e de avaliar o potencial dessas representações semanticamente enriquecidas, neste trabalho foram exploradas diferentes abordagens em direção à incorporação da semântica na representação de textos. Assim, foram propostos sete modelos de representação de coleções de documentos: (i) *gBoED*: modelo que incorpora a semântica obtida por meio de conhecimento do domínio (apresentado no Capítulo 3, Seção 3.2 - Página 73); (ii) *Uni-based*: modelo que incorpora a semântica por meio da desambiguação lexical de sentidos e relação de hiperonímia entre conceitos (apresentado no Capítulo 3, Seção 3.3 - Página 79); (iii) *SR-based Terms* e *SR-based Sentences*: modelos que incorporam a semântica por meio de anotações de papéis semânticos (apresentados no Capítulo 4 - Página 87); (iv) *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*: modelos que incorporam a semântica por meio de desambiguação lexical de sentidos e *embeddings* de palavras e conceitos (apresentados no Capítulo 5 - Página 111). Representações de coleções de documentos geradas com os modelos foram analisadas e avaliadas na classificação automática de textos considerando os *datasets* de diferentes níveis de complexidade semântica.

Os modelos *gBoED*, *Uni-based*, *SR-based Terms* e *SR-based Sentences* apresentam atributos mais expressivos do que os atributos da *bag-of-words*, no entanto a performance de classificação é, no geral, mais baixa do que a performance obtida com a *bag-of-words*. Por apresentar resultados inferiores, porém próximos aos obtidos com a BOW, os classificadores gerados com a representação *SR-based Terms* só seriam indicados para casos em que a expressividade dos atributos tenha grande importância na utilização do modelo gerado.

Já uma análise dos vetores gerados com os modelos *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, indicou que as representações propostas geram vetores para os documentos próximos a vetores de palavras e/ou sentidos relacionados ao conteúdo dos mesmos. Uma desvantagem de representações baseadas em *embeddings*, como essas propostas, é falta de

interpretabilidade de seus atributos. Nesse ponto, os modelos *NASARI_{docs}* e *NASARI+Babel2Vec* apresentam uma vantagem em relação aos demais modelos analisados. Por serem gerados a partir dos vetores de sentidos *NASARI embedded*, é possível utilizar os sentidos mais próximos ao documento para interpretar a sua representação. No entanto, *NASARI_{docs}* é limitada à representação de substantivos, o que pode impossibilitar o seu uso na representação de determinados documentos, como os textos de opinião contendo apenas adjetivos.

As representações *Babel2Vec* e *NASARI+Babel2Vec* se mostraram adequadas para a identificação de similaridade textual, apresentando boas correlações com similaridades atribuídas por pessoas. Tais representações também se mostraram adequadas na avaliação experimental realizada com a classificação de textos, sendo que as performances de classificação obtidas com o uso das representações propostas foram superiores, com diferença estatisticamente significativa, em relação às representações BOW, LDA e Doc2Vec. Já a representação *Babel2Vec* também levou a performance superior que a representação Word2Vec. Os bons resultados obtidos com as representações geradas a partir de *embeddings* quando comparadas a BOW e LDA indicam que essas representações conseguem incorporar conhecimento obtido dos grandes corpora usados para treinar as *embeddings*. Padrões descobertos nesses corpora contribuem para uma melhor representação do conteúdo das coleções de documentos. Enquanto as representações baseadas em *embeddings* são enriquecidas com esse conhecimento externo, as representações BOW e LDA são construídas apenas com base no próprio conteúdo dos documentos.

Considerando-se a combinação de classificadores gerados com representações semanticamente enriquecidas e a tradicional *bag-of-words*, as avaliações experimentais realizadas indicam que é possível aproveitar vantagens de ambas as representações. As melhores performances de classificação foram obtidas com a combinação de classificadores gerados com a representação *Babel2Vec* com classificadores gerados com a *bag-of-words*, considerando-se pesos iguais ou maiores para o classificador gerado com *Babel2Vec*.

A principal inovação deste trabalho está na proposta e avaliação de representações de textos semanticamente enriquecidas. As propostas *gBoED*, *Uni-base*, *SR-based Terms* e *SR-based Sentences* apresentam atributos mais expressivos e possibilitam uma melhor interpretação da representação dos documentos. Já as propostas *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec* incorporam, de maneira latente, a semântica obtida de *embeddings* geradas a partir de uma grande quantidade de documentos externos. Essa propriedade tem um impacto positivo na performance de classificação, de acordo com a avaliação experimental realizada com os modelos *Babel2Vec* e *NASARI+Babel2Vec*. Em relação à avaliação experimental, destaca-se também como novidade a preparação de *datasets* que representam diferentes problemas de classificação para uma mesma coleção de documentos. Os *datasets* construídos e disponibilizados neste trabalho podem contribuir com a avaliação de pesquisas futuras em Mineração de Textos.

6.2 Publicações

Durante o desenvolvimento deste trabalho, as contribuições obtidas foram divulgadas por meio de publicação de artigos em periódicos e publicação e apresentação de artigos em conferências. Essas publicações são listadas a seguir, apresentando a relação de cada uma com este trabalho e indicando aquelas que estão diretamente relacionadas às questões de pesquisas estabelecidas.

Artigos publicados em periódicos e anais de conferências

SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. *Text mining and semantics: a systematic mapping study*. Journal of the Brazilian Computer Society, v. 23, p. 1-20, 2017.

Nesse artigo foi publicado o resultado do mapeamento sistemático da literatura desenvolvido neste trabalho e está relacionado à questão **Q1**.

SINOARA, R. A.; SCHEICHER, RICARDO B.; REZENDE, SOLANGE O. *Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity*. In: CIDM'17- Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence, Honolulu, USA, p. 2057–2064, 2017.

Nesse artigo foi publicado o trabalho referente à definição de níveis de complexidade semântica e apresenta uma prova de conceito sobre a aplicação do método LDA em cenários reais de organização de coleções de documentos. O trabalho apresentado nesse artigo está relacionado às questões **Q2, Q3 e Q4**.

SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. *Semantic Role-Based Representations in Text Classification*. In: ICPR 2016 - Proceedings of the 23rd International Conference on Pattern Recognition, Cancún, México, p. 2314-2319, 2016.

Nesse artigo foi publicada a proposta das representações *SR-based Terms* e *SR-based Sentences*. O trabalho apresentado nesse artigo está relacionado às questões **Q4, Q5 e Q6**.

SCHEICHER, R.; SINOARA, R. A.; KOGA, N.; REZENDE, S. O. *Uso de expressões do domínio na classificação automática de documentos*. In: ENIAC 2016 - Anais do XIII Encontro Nacional de Inteligência Artificial e Computacional, Recife, Brasil, p. 625-636, 2016.

Nesse artigo foi publicada a proposta da representação *gBoED*, desenvolvida neste trabalho. A implementação computacional do método e a execução da avaliação experimental foi realizada em parceria com os demais autores. A autora desta tese foi responsável pela elaboração da proposta e da configuração experimental e participou ativamente da análise dos resultados e escrita do artigo. O trabalho apresentado nesse artigo está relacionado às questões **Q4, Q5 e Q6**.

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: ICED15 - Proceedings of the 20th International Conference on Engineering Design, Milan, Italy, v. 7, p. 311-320, 2015.

Nesse artigo foi publicada a proposta da representação BoED para apoio à análise da aplicação prática de métodos de desenvolvimento de produtos em artigos dessa área. Este trabalho, realizado em parceria com outros pesquisadores, serviu como um indicativo da necessidade de incorporação da semântica na representação de coleções de documentos e também motivou o desenvolvimento da representação *gBoED*. A autora desta tese participou ativamente da elaboração da proposta, execução dos experimentos, análise dos resultados e escrita do artigo.

MARQUES, C. A. N.; MATSUNO, I. P.; **SINOARA, R. A.**; REZENDE, S. O.; ROZENFELD, H. *Comparative analysis of methods and tools applicability for product and IPSS development based on text mining techniques*. Product: Management & Development (IGDP), v. 13, p. 57-66, 2015.

Nesse artigo foi publicada uma nova aplicação do método desenvolvido e apresentado no artigo anterior. A autora desta tese participou ativamente na aplicação do método e escrita do artigo.

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: IDEAS'14 Proceedings of the 18th International Database Engineering & Applications Symposium, Porto, Portugal, p. 57-66, 2014.

Nesse artigo foi publicada uma proposta para incorporação da semântica no agrupamento hierárquico de textos por meio do uso de entidades nomeadas. Este trabalho resultou de uma exploração inicial da área pesquisa.

Resumos publicados

SILVA JÚNIOR, E. L.; **SINOARA, R. A.**; REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. *Uma ferramenta para expansão do vocabulário com base em coocorrência*. In: Resumos Expandidos da XI Mostra de estagiários e bolsistas da Embrapa Informática Agropecuária, p. 11-15, 2015.

Nesse artigo é apresentada uma ferramenta construída para representação de documentos com base em vocabulário do domínio. Esse trabalho foi desenvolvido em parceria com um aluno de iniciação científica, como parte de um projeto conjunto com a Embrapa Informática Agropecuária. A autora desta tese participou ativamente da especificação da ferramenta e acompanhamento das atividades envolvidas.

RODRIGUES, L. S.; SINOARA, R. A.; REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. *Identificação de Pontos Perceptualmente Importantes (PIP) em séries temporais de tópicos extraídos de dados textuais*. In: Resumos Expandidos da XI Mostra de estagiários e bolsistas da Embrapa Informática Agropecuária, p. 38-43, 2015.

Nesse artigo é apresentada uma ferramenta construída para identificar pontos relevantes da evolução temporal de um tópico e identificar documentos textuais que possam auxiliar a interpretar tais pontos. A autora desta tese participou de reuniões de especificação da proposta e da revisão do texto.

PARAVIA, R. P. P.; SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. *Sistema para apoio à rotulação manual de textos utilizando aprendizado de máquina*. In: SIICUSP - Anais do Simpósio Internacional de Iniciação Científica da USP, 2015.

Nesse artigo é apresentada a ferramenta RotuLABIC, desenvolvida para apoiar a rotulação manual de textos. Esse trabalho foi realizado em parceria com outros alunos de iniciação científica e de doutorado. A autora desta tese participou ativamente da elaboração da proposta, especificação da ferramenta, avaliação experimental e acompanhamento das atividades envolvidas.

MARCACINI, R. M.; SINOARA, R. A.; MATSUNO, I. P.; REZENDE, S. O. *Aprendizado não Supervisionado de WebSensors*. In: KDMiLe - Proceedings of the First Symposium on Knowledge Discovery, Mining and Learning, Workshop of Ideas and Projects, São Carlos, Brasil, p. 1-2, 2013.

Nesse artigo é apresentado o projeto de aprendizado não supervisionado de WebSensors, desenvolvido no grupo de pesquisa e em parceria com outras instituições. O trabalho apresentado nesta tese também foi desenvolvido em associação a esse projeto e a um projeto posterior e mais abrangente sobre o aprendizado de WebSensors. A autora desta tese participou do desenvolvimento desse projeto e da escrita desse artigo.

Relatórios Técnicos publicados

SINOARA, R. A.; REZENDE, S. O. *BEST sports: a Portuguese collection of documents for semantics-concerned text mining research*. Relatório Técnico N° 424, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2018.

Nesse relatório técnico é apresentada a coleção de documentos *BEST sports*, coletada e preparada para avaliações experimentais em relação aos dois níveis de complexidade semântica definidos neste trabalho. O trabalho apresentado nesse relatório está relacionado às questões **Q3**, **Q4**, **Q5** e **Q6**.

GOMES, M. M.; SINOARA, R. A.; REZENDE, S. O. *PRED.ARG: Ferramenta para gerar representações de documentos com base em papéis semânticos*. Relatório Técnico N° 425, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2018.

Nesse relatório técnico é apresentada a ferramenta PRED.ARG, que possibilita a geração das representações propostas por Persson, Johansson e Nugues (2009) e utilizadas neste trabalho na avaliação experimental apresentada no Capítulo 4. Esse trabalho foi desenvolvido em parceria com um aluno de iniciação científica. A autora desta tese foi responsável pela especificação da ferramenta e participou ativamente das atividades envolvidas. O trabalho apresentado nesse relatório está relacionado às questões Q5 e Q6.

Artigo submetido

Além dos trabalhos já publicados apresentados, o seguinte trabalho foi submetido, visando a publicação no curto prazo.

SINOARA, R. A.; CAMACHO-COLLADOS, J.; NAVIGLI, R.; REZENDE, S. O. *Knowledge-enhanced document embeddings for text classification*. [Submetido]

Nesse artigo é apresentada a proposta das representações *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*. Parte desse trabalho foi desenvolvido durante o estágio realizado na Sapienza - Università degli Studi di Roma, em parceria com autores da abordagem *NASARI*. O artigo foi submetido para revisão e publicação no periódico *Knowledge-based Systems*. O trabalho apresentado nesse artigo está relacionado às questões Q4, Q5 e Q6.

6.3 Limitações e trabalhos futuros

Nessa seção são discutidas algumas características e limitações das abordagens propostas e das avaliações experimentais realizadas, bem como direções de trabalhos futuros. Uma característica das abordagens propostas refere-se à necessidade de processamento extra dos textos para gerar as representações. No caso dos modelos *SR-based Terms* e *SR-based Sentences*, os textos devem ser processados por um sistema de anotação de papéis semânticos. No caso dos modelos *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, os documentos devem ser desambiguados. Com relação a textos muito curtos, textos de opinião ou textos de baixa qualidade, pode não ser possível a geração da representação devido a limitação na extração das informações semânticas necessárias. Esse fato pode ocorrer nos seguintes casos: para *gBoED*, caso o documento não contenha nenhum dos termos presentes nas listas utilizadas para geração da representação; para *SR-based Terms* e *SR-based Sentences*, caso não seja anotado o papel semântico de nenhum termo do documento; para *Uni-based*, *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, caso nenhum fragmento do documento seja desambiguado ou caso nenhum dos termos desambiguados esteja presente nos vetores utilizados para gerar as representações.

No sentido de explorar novas possibilidades para tratar a questão da necessidade de processamento extra e/ou a limitação na extração de informações semânticas para certos textos, pode-se investigar a aplicação da abordagem de aprendizado com informação privilegiada

(VAPNIK; VASHIST, 2009; MARCACINI, 2014). Nessa abordagem, a representação semanticamente enriquecida pode ser tratada como informação privilegiada e, portanto, pode estar disponível para apenas parte da coleção de documentos. Com isso, evita-se o processamento extra de toda a coleção e o impacto em relação a documentos que, por ventura, não possam ser representados. Durante o desenvolvimento deste projeto de doutorado, o aprendizado com informação privilegiada foi explorado utilizando representações geradas a partir de entidades nomeadas como informação privilegiada no agrupamento hierárquico de documentos, obtendo-se resultados promissores (SINOARA *et al.*, 2014).

Especificamente para o modelo *gBoED*, limitações em relação à aquisição de conhecimento do domínio e à qualidade das listas utilizadas podem impactar na qualidade da representação. Pode-se ter casos em que não é possível obter listas de identificadores de classe para todas as classes, como por exemplo no caso das representações geradas para a coleção *SE-ABSA15*, para a qual não foi possível gerar uma lista de termos para identificar a classe “neutro”. Nesse contexto, para realizar a geração automática ou a expansão das listas pode-se explorar técnicas de extração automática de termos do domínio (CONRADO *et al.*, 2014; KORKONTZELOS; KLAPFTIS; MANANDHAR, 2008).

Uma outra perspectiva que pode ser investigada é o uso da representação *gBoED* no contexto de aprendizado semissupervisionado e aprendizado ativo (AGGARWAL, 2014). A avaliação experimental apresentada nesta tese indica que a representação *gBoED*, apesar de apresentar atributos semanticamente mais ricos, não é suficiente para levar a uma boa performance de classificação quando utilizada de maneira independente. No entanto, dada a riqueza da representação *gBoED* em relação ao conhecimento do domínio, uma linha de investigação interessante é a aplicação das expressões do domínio para a rotulação de documentos para os quais apenas a representação *bag-of-words* não é suficiente para identificação do rótulo do documento com um bom nível de confiança.

Já a qualidade das representações *SR-based Terms* e *SR-based Sentences* depende da qualidade dos textos e do sistema anotador de papéis semânticos. Uma abordagem que pode ser empregada para melhorar a qualidade de textos é a aplicação de técnicas de normalização (BERTAGLIA, 2017; SILVA, 2016). Técnicas avançadas de normalização textual têm sido empregadas para identificar e tratar textos que divergem da norma culta da língua, como textos que apresentam erros ortográficos, abreviações ou gírias. Problemas como estes são comuns em conteúdo gerado por usuário e dificultam a aplicação de métodos de PLN. Considerando a qualidade da anotação de papéis semânticos, Hartmann, Duran e Aluísio (2016) analisaram dois sistemas para anotação de textos em português: o sistema de Fonseca e Rosa (2013b), utilizado neste trabalho, e o sistema de Alva-Manchego (2013), que utiliza atributos sintáticos no aprendizado. De acordo com os resultados obtidos, os autores recomendam o uso de sistemas que utilizam árvores sintáticas para anotação de textos em português e o seu treinamento com árvores sintáticas não revisadas. Assim, o uso de sistemas de anotação de papéis semânticos

com melhor desempenho pode melhorar a qualidade das representações baseadas em papéis semânticos e, conseqüentemente, os resultados da classificação automática de documentos.

No caso de *Uni-based*, *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*, a qualidade da representação depende da qualidade da desambiguação e dos vetores utilizados. Assim como para as representações anteriores, técnicas de normalização dos textos podem melhorar a qualidade de textos com muitas gírias ou erros e, conseqüentemente, melhorar os resultados da desambiguação. Analisando-se os vetores *NASARI unified* utilizados na construção da representação *Uni-based*, observou-se que, no geral, eram de boa qualidade. No entanto, foram identificadas pequenas inconsistências, como por exemplo, o *synset* “*Apple tree* - Any tree of the genus *Malus* especially those bearing firm rounded edible fruits” é uma das principais dimensões nas representações dos *synsets* relacionados a produtos da empresa Apple, como “Apple Inc.”, “MacBook Pro” e “PowerBook”. Já as representações baseadas nos vetores *NASARI embedded* e *Word2Vec* apresentaram bons resultados tanto na análise de similaridade textual quanto na classificação de textos. No entanto, em uma análise dos vetores *Word2Vec* também foi possível verificar pontos que podem ser melhorados com a realização de um pré-processamento mais minucioso antes da geração dos vetores. Por exemplo, existem vetores para as palavras “MacBook” e “Macbook”, que provavelmente poderiam ser unificadas em uma única palavra. A geração de *embeddings* está em evolução e novos vetores têm sido disponibilizados (vide <http://lcl.uniroma1.it/nasari/#two>). Os avanços nessa área, a disponibilidade de grandes corpus para treinamento e o aumento da qualidade dos vetores, têm um impacto positivo na geração das representações propostas. Além disso, para casos em que exista um grande corpus do domínio, a obtenção de vetores *NASARI* e *Word2Vec* específicos pode ser uma direção para investigações futuras. O uso desses vetores específicos do domínio pode levar a resultados ainda melhores para o uso das representações propostas, porém com um custo adicional relacionado à geração desses novos vetores.

Uma linha de trabalho futuro promissora é a investigação de diferentes contextos na geração de *embeddings* utilizadas nos modelos *NASARI_{docs}*, *Babel2Vec* e *NASARI+Babel2Vec*. Seguindo o trabalho de Levy e Goldberg (2014), pode-se investigar o uso de anotação de papéis semânticos como contexto para a geração de *embeddings* e o impacto dessas novas *embeddings* na representação de coleções de documentos, bem como na performance de classificação desses documentos.

As avaliações experimentais realizadas apresentam uma limitação relacionada ao número de coleções de documentos utilizadas. Nesse sentido, um importante trabalho futuro é a seleção, coleta e preparação de novas coleções de documentos que possibilitem a criação de *datasets* que representem tanto os problemas de organização por tópico quanto os de organização semântica. Com uma quantidade maior de *datasets*, poderá ser realizada uma avaliação mais abrangente das abordagens propostas. Ainda em relação às avaliações experimentais, neste trabalho procurou-se padronizar a geração das representações avaliadas visando manter certa igualdade de configurações entre as representações. Por exemplo, padronizou-se o uso da medida TF e o número de

300 dimensões. No entanto, existem diversas outras possibilidades de configurações para gerar as representações, que podem ser exploradas em trabalhos futuros. Um outro ponto relacionado à avaliação experimental refere-se à geração da representação *gBoED*. Por limitações dos domínios, utilizou-se apenas listas relacionadas a problemas de maior complexidade semântica. Em trabalhos futuros, pode-se buscar conhecimento mais aprofundado sobre os domínios e gerar listas específicas para cada *dataset*.

Apesar de neste trabalho o foco ter sido em classificação automática de textos, os resultados obtidos podem ser aplicados em outras tarefas de Mineração de Textos. Assim, em relação à expansão dessa pesquisa, tem-se como trabalho futuro a evolução das abordagens propostas considerando-se aspectos semânticos na representação de documentos para outras tarefas de MT. Entre essas tarefas, pode-se citar o aprendizado não supervisionado, aprendizado baseado em uma única classe (*one-class classification*), recuperação e extração de informação, além de diferentes tarefas de análise de sentimentos e sistemas de recomendação (BELLINGER *et al.*, 2017; SALEIRO *et al.*, 2017; MATSUNO *et al.*, 2017; MANZATO *et al.*, 2016). Nesse sentido, pretende-se inicialmente investigar a aplicação das abordagens propostas no aprendizado baseado em uma única classe para mineração de laudos de exames médicos no contexto do projeto MIVisBD - Mineração, Indexação e Visualização de Big Data no Contexto de Sistemas de Apoio a Decisão Clínica (Processo FAPESP no. 2016/17078-0).

REFERÊNCIAS

- AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2014. Citado nas páginas 48 e 164.
- AGGARWAL, C. C.; ZHAI, C. (Ed.). **Mining Text Data**. [S.l.]: Springer, 2012. Citado nas páginas 25, 27, 29, 38, 42, 44, 45 e 52.
- AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: AGGARWAL, C. C.; ZHAI, C. (Ed.). **Mining Text Data**. [S.l.]: Springer, 2012. cap. 4, p. 77–128. Citado na página 25.
- AGIRRE, E.; BANEJA, C.; CARDIE, C.; CER, D.; DIAB, M.; GONZALEZ-AGIRRE, A.; GUO, W.; LOPEZ-GAZPIO, I.; MARITXALAR, M.; MIHALCEA, R.; RIGAU, G.; URIA, L.; WIEBE, J. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In: **SemEval 2015: Proceedings of the 9th international workshop on semantic evaluation**. [S.l.: s.n.], 2015. p. 252–263. Citado na página 51.
- AGIRRE, E.; EDMONDS, P. G. **Word sense disambiguation: Algorithms and applications**. [S.l.]: Springer, 2007. v. 33. Citado na página 51.
- ALVA-MANCHEGO, F. E. **Anotação automática semissupervisionada de papéis semânticos para o português do Brasil**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional. Citado nas páginas 88, 96 e 164.
- AMARAL, D. O. F. do; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014. Citado na página 50.
- ARAÚJO, I. L. Por uma concepção semântico-pragmática da linguagem. **Revista Virtual de Estudos da Linguagem**, v. 5, n. 8, p. 1–26, 2007. Citado na página 48.
- BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: **LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2010. v. 10, n. 2010, p. 2200–2204. Citado na página 61.
- BALAGE FILHO, P. P.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: **STIL 2013: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. p. 215–219. Citado na página 61.
- BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: **ACL 2014: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. p. 238–247. Citado nas páginas 111, 112 e 113.

- BEKKERMAN, R.; RAGHAVAN, H.; ALLAN, J.; EGUCHI, K. Interactive clustering of text collections according to a user-specified criterion. In: **IJCAI'07: Proceedings of the 20th International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2007. p. 684–689. Citado na página [87](#).
- BELLINGER, C.; SHARMA, S.; ZAIANE, O. R.; JAPKOWICZ, N. Sampling a longer life: Binary versus one-class classification revisited. In: TORGO, L.; KRAWCZYK, B.; BRANCO, P.; MONIZ, N. (Ed.). **Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications**. [S.l.]: PMLR, 2017. (Proceedings of Machine Learning Research, v. 74), p. 64–78. Citado na página [166](#).
- BERTAGLIA, T. F. C. **Normalização textual de conteúdo gerado por usuário**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional. Citado na página [164](#).
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003. Citado nas páginas [27](#), [42](#) e [112](#).
- BONIAL, C.; BONN, J.; CONGER, K.; HWANG, J. D.; PALMER, M. Propbank: Semantics of new predicate types. In: **LREC'14: Proceedings of the Ninth International Conference on Language Resources and Evaluation**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 3013–3019. Citado na página [96](#).
- 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**, Elsevier, v. 240, p. 36–64, 2016. Citado nas páginas [73](#), [80](#), [81](#), [82](#), [83](#), [112](#), [114](#), [116](#), [117](#), [189](#) e [191](#).
- CARVALHO, D. S.; FREITAS, A.; SILVA, J. C. P. da. Graphia: Extracting contextual relation graphs from text. In: **ESWC 2013: Proceedings of the 10th Extended Semantic Web Conference**. [S.l.]: Springer Berlin Heidelberg, 2013. v. 7955, p. 236–241. Citado na página [44](#).
- CEVIKALP, H.; POLIKAR, R. Local classifier weighting by quadratic programming. **IEEE Transactions on Neural Networks**, v. 19, n. 10, p. 1832–1838, 2008. Citado na página [46](#).
- COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. Natural language processing (almost) from scratch. **Journal of Machine Learning Research**, p. 2461–2505, 2011. Citado nas páginas [88](#), [91](#), [94](#), [189](#) e [190](#).
- CONRADO, M. da S.; FELIPPO, A. D.; PARDO, T. A. S.; REZENDE, S. O. A survey of automatic term extraction for brazilian portuguese. **Journal of the Brazilian Computer Society**, v. 20, n. 1, p. 12, 2014. Citado nas páginas [26](#) e [164](#).
- CONRADO, M. da S.; GUTIÉRREZ, V. A. L.; REZENDE, S. O. Evaluation of normalization techniques in text classification for portuguese. In: **ICCSA 2012: Proceedings of the 12th International Conference on Computational Science and Its Applications**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 618–630. Citado na página [42](#).
- 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. 130–161. Citado na página [113](#).

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006. Citado na página 194.

DIETTERICH, T. G. Ensemble methods in machine learning. In: **MCS'00: Proceedings of the First International Workshop on Multiple Classifier Systems**. [S.l.: s.n.], 2000. p. 1–15. Citado na página 46.

DOAN, S.; KAWAZOE, A.; CONWAY, M.; COLLIER, N. Towards role-based filtering of disease outbreak reports. **Journal of Biomedical Informatics**, Elsevier, v. 42, n. 5, p. 773–780, 2009. Citado na página 87.

DURAN, M. S.; ALUÍSIO, S. M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: **LREC-2012: Proceedings of the Eighth International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2012. p. 1862–1867. Citado na página 88.

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.]: Editora Manole, 2003. cap. 13, p. 337–370. Citado na página 26.

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 na página 48.

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [S.l.]: MIT Press, 1998. Citado nas páginas 80 e 115.

FLEKOVA, L.; GUREVYCH, I. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In: **ACL 2016: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2016. p. 2029–2041. Citado na página 114.

FONSECA, E.; SANTOS, L. B. dos; CRISCUOLO, M.; ALUÍSIO, S. Visão geral da avaliação de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, 2016. Citado na página 51.

FONSECA, E. R. **Uma abordagem conexcionista para anotação de papéis semânticos**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional. Citado na página 96.

FONSECA, E. R.; ROSA, J. L. G. An architecture for semantic role labeling on portuguese. In: **PROPOR 2012: Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language**. [S.l.]: Springer Berlin Heidelberg, 2012. p. 204–209. Citado na página 51.

_____. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **STIL 2013: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. p. 98–107. Citado na página 190.

_____. A two-step convolutional neural network approach for semantic role labeling. In: **IJCNN 2013: Proceedings of the International Joint Conference on Neural Networks**. [S.l.: s.n.], 2013. p. 2955–2961. Citado nas páginas 88, 89, 164 e 190.

FOSSILE, D. K. Semântica & pragmática: Campos independentes. **VOOS Revista Polidisciplinar Eletrônica da Faculdade Guairacá**, v. 2, p. 40–58, 2009. Citado na página 48.

- GABRILOVICH, E.; MARKOVITCH, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: **IJCAI-07: Proceedings of the 20th International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2007. p. 1606–1611. Citado na página [43](#).
- GARRIDO, A. L.; BUEY, M. G.; ESCUDERO, S.; ILARRI, S.; MENA, E.; SILVEIRA, S. B. Tm-gen: A topic map generator from text documents. In: **ICTAI 2013: Proceedings of IEEE 25th International Conference on Tools with Artificial Intelligence**. [S.l.: s.n.], 2013. p. 735–740. Citado na página [44](#).
- GREENE, D.; CUNNINGHAM, P. Practical solutions to the problem of diagonal dominance in kernel document clustering. In: **ICML'06: Proceedings of the 23rd International Conference on Machine learning**. [S.l.]: ACM Press, 2006. p. 377–384. Citado na página [66](#).
- GRIMES, S. **Text Analytics 2014: User Perspectives on Solutions and Providers**. 2014. Alta Plana Corporation. Citado na página [29](#).
- GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **COLING 96: Proceedings of the 16th International Conference on Computational Linguistics**. [S.l.: s.n.], 1996. v. 96, p. 466–471. Citado na página [50](#).
- GROBELNIK, M. Many faces of text processing. In: **WIMS'11: Proceedings of the International Conference on Web Intelligence, Mining and Semantics**. [S.l.: s.n.], 2011. p. 5. Citado na página [37](#).
- GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157–1182, 2003. Citado na página [42](#).
- HAN, J. Mining heterogeneous information networks: the next frontier. In: **KDD'12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2012. p. 1–3. Citado na página [44](#).
- HARISPE, S.; RANWEZ, S.; JANAQI, S.; MONTMAIN, J. Semantic similarity from natural language and ontology analysis. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 8, n. 1, p. 1–254, 2015. Citado na página [51](#).
- HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In: **PROPOR 2016: Proceedings of the 12th International Conference on the Computational Processing of Portuguese**. [S.l.]: Springer International Publishing, 2016. p. 202–212. Citado nas páginas [89](#) e [164](#).
- HU, M.; LIU, B. Mining and summarizing customer reviews. In: **Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2004. p. 168–177. Citado na página [189](#).
- HU, X.; ZHANG, X.; LU, C.; PARK, E. K.; ZHOU, X. Exploiting wikipedia as external knowledge for document clustering. In: **KDD'09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2009. p. 389–396. Citado na página [28](#).
- IACOBACCI, I.; PILEHVAR, M. T.; NAVIGLI, R. Senseembed: Learning sense embeddings for word and relational similarity. In: **ACL 2015: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2015. p. 95–105. Citado na página [114](#).

- JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010. Citado na página [27](#).
- JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition**. 2nd. ed. [S.l.]: Prentice Hall, Pearson Education International, 2009. Citado na página [88](#).
- KHAN, A.; SALIM, N.; KUMAR, Y. J. A framework for multi-document abstractive summarization based on semantic role labelling. **Applied Soft Computing**, v. 30, p. 737 – 747, 2015. Citado na página [89](#).
- KIM, H. joon; HONG, K. joo; CHANG, J. Y. Semantically enriching text representation model for document clustering. In: **SAC '15: Proceedings of the 30th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2015. p. 922–925. Citado na página [28](#).
- KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007. EBSE Technical Report EBSE-2007-01. Keele University and Durham University Joint Report. Citado na página [53](#).
- KORKONTZELOS, I.; KLAPAFTIS, I. P.; MANANDHAR, S. Reviewing and evaluating automatic term recognition techniques. In: **Advances in Natural Language Processing**. [S.l.]: Springer Berlin Heidelberg, 2008. p. 248–259. Citado na página [164](#).
- KUHN, A.; DUCASSE, S.; GÍRBA, T. Semantic clustering: Identifying topics in source code. **Information and Software Technology**, v. 49, p. 230–243, 2007. Citado na página [42](#).
- LAU, J. H.; BALDWIN, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. In: **Proceedings of the 1st Workshop on Representation Learning for NLP**. [S.l.: s.n.], 2016. p. 78–86. Citado na página [43](#).
- LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **ICML-14: Proceedings of the 31st International Conference on Machine Learning**. [S.l.: s.n.], 2014. p. 1188–1196. Citado nas páginas [28](#), [33](#), [43](#) e [114](#).
- LEE, H.; YOON, Y. Engineering doc2vec for automatic classification of product descriptions on o2o applications. **Electronic Commerce Research**, p. 1–24, 2017. Citado na página [43](#).
- LEE, M. D.; PINCOMBE, B.; WELSH, M. B. An empirical evaluation of models of text document similarity. In: **Proceedings of the 27th Annual Conference of the Cognitive Science Society**. [S.l.: s.n.], 2005. p. 1254–1259. Citado nas páginas [51](#), [128](#), [129](#), [130](#) e [152](#).
- LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: **ACL 2014: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. p. 302–308. Citado nas páginas [28](#), [29](#), [111](#), [113](#) e [165](#).
- LIANG, H.; FOTHERGILL, R.; BALDWIN, T. Rosemerry: A baseline message-level sentiment classification system. In: **SemEval 2015: Proceedings of the 9th International Workshop on Semantic Evaluation**. [S.l.: s.n.], 2015. p. 551–555. Citado na página [43](#).
- LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan & Claypool Publishers, 2012. Citado na página [61](#).

- LIU, J.; WANG, J.; WANG, C. A text network representation model. In: **FSKD'08: Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery**. [S.l.: s.n.], 2008. v. 4, p. 150–154. Citado na página 44.
- LIU, Z.; LI, M.; LIU, Y.; PONRAJ, M. Performance evaluation of latent dirichlet allocation in text mining. In: **FSKD 2011: Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery**. [S.l.: s.n.], 2011. v. 4, p. 2695–2698. Citado nas páginas 27 e 113.
- LLORENS, H.; SAQUETE, E.; NAVARRO-COLORADO, B. Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. **Information Processing & Management**, Elsevier, v. 49, n. 1, p. 179–197, 2013. Citado na página 89.
- LU, Q.; GETOOR, L. Link-based classification. In: **ICML'2003: Proceedings of the International Conference on Machine Learning**. [S.l.: s.n.], 2003. p. 496–503. Citado na página 44.
- LU, Y.; MEI, Q.; ZHAI, C. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. **Information Retrieval**, v. 14, n. 2, p. 178–203, 2011. Citado nas páginas 27 e 113.
- MANZATO, M. G.; DOMINGUES, M. A.; FORTES, A. C.; SUNDERMANN, C. V.; D'ADDIO, R. M.; CONRADO, M. S.; REZENDE, S. O.; PIMENTEL, M. G. C. Mining unstructured content for recommender systems: an ensemble approach. **Information Retrieval Journal**, v. 19, n. 4, p. 378–415, 2016. Citado na página 166.
- MARCACINI, R. M. **Aprendizado de máquina com informação privilegiada: abordagens para agrupamento hierárquico de textos**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional. Citado na página 164.
- MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a large annotated corpus of english: The penn treebank. **Computational Linguistics**, MIT Press, Cambridge, MA, USA, v. 19, n. 2, p. 313–330, 1993. Citado na página 92.
- MARKOV, I.; GÓMEZ-ADORNO, H.; POSADAS-DURÁN, J.-P.; SIDOROV, G.; GELBUKH, A. Author profiling with doc2vec neural network-based document embeddings. In: **MICAI 2016: Proceedings of the 15th Mexican International Conference on Artificial Intelligence**. [S.l.]: Springer International Publishing, 2016. p. 117–131. Citado na página 43.
- MARQUES, C. A. N.; MATSUNO, I. P.; SINOARA, R. A.; REZENDE, S. O.; ROZENFELD, H. An exploratory study to evaluate the practical application of pss methods and tools based on text mining. In: **Proceedings of the 20th International Conference on Engineering Design**. [S.l.: s.n.], 2015. p. 7–311–7–320. Citado nas páginas 73, 74, 75 e 77.
- MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Aspect-based sentiment analysis using semi-supervised learning in bipartite heterogeneous networks. **Journal of Information and Data Management**, v. 7, n. 2, p. 141, 2017. Citado na página 166.
- MCCALLUM, A. K. **MALLET: A Machine Learning for Language Toolkit**. 2002. Disponível em: <<http://mallet.cs.umass.edu>>. Acesso em: 24/01/2018. Citado na página 187.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: **ICLR Workshop**. [S.l.: s.n.], 2013. Citado nas páginas 28, 43, 113, 114 e 115.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems**. USA: Curran Associates Inc., 2013. p. 3111–3119. Citado nas páginas 33 e 115.

MINER, G.; ELDER, J.; HILL, T.; NISBET, R.; DELEN, D.; FAST, A. **Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications**. 1st. ed. [S.l.]: Academic Press, 2012. Citado na página 38.

MISHRA, M.; HUAN, J.; BLEIK, S.; SONG, M. Biomedical text categorization with concept graph representations using a controlled vocabulary. In: **BIOKDD'12: Proceedings of the 11th International Workshop on Data Mining in Bioinformatics**. [S.l.: s.n.], 2012. p. 26–32. Citado na página 44.

MITCHELL, T. M. **Machine Learning**. New York, NY, USA: McGraw-Hill, Inc., 1997. Citado na página 48.

MÜLLER, A. L. de P.; VIOTTI, E. de C. Semântica formal. In: **Introdução à Lingüística II: Princípios de Análise**. [S.l.]: Editora Contexto, 2010. p. 137–159. Citado na página 50.

MONTES-Y-GOMEZ, M.; GELBUKH, A.; LOPEZ-LOPEZ, A.; BAEZA-YATES, R. Text mining with conceptual graphs. In: **SMC 2001: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics**. [S.l.: s.n.], 2001. v. 2, p. 898–903. Citado na página 44.

MORO, A.; CECCONI, F.; NAVIGLI, R. Multilingual word sense disambiguation and entity linking for everybody. In: **ISWC-PD'14: Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272**. Aachen, Germany, Germany: CEUR-WS.org, 2014. p. 25–28. Citado nas páginas 190 e 195.

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 nas páginas 51, 116, 187 e 195.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. **Artificial Intelligence**, Elsevier, v. 193, p. 217–250, 2012. Citado nas páginas 81 e 115.

NÓBREGA, F. A. A.; PARDO, T. A. S. General purpose word sense disambiguation methods for nouns in portuguese. In: **PROPOR 2014: Proceedings of 11th International Conference on Computational Processing of the Portuguese Language**. [S.l.]: Springer International Publishing, 2014. v. 8775, p. 94–101. Citado na página 51.

NOGUEIRA, B. M.; REZENDE, S. O. Dois novos métodos para seleção não-supervisionada de atributos em mineração de textos. In: **CLEI'09: Anais da XXXV Conferencia Latinoamericana de Informática**. [S.l.: s.n.], 2009. p. 1–10. Citado nas páginas 26 e 42.

- OCHOA, J. L.; VALENCIA-GARCÍA, R.; PEREZ-SOLTERO, A.; BARCELÓ-VALENZUELA, M. A semantic role labelling-based framework for learning ontologies from spanish documents. **Expert Systems with Applications**, v. 40, n. 6, p. 2058–2068, 2013. Citado nas páginas 87 e 89.
- OLIVEIRA, R. P. Semântica. In: **Introdução à linguística: domínios e fronteira, volume 2**. [S.l.]: Cortez, 2012. p. 23–54. Citado na página 48.
- PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. **Computational Linguistics**, MIT Press, v. 31, n. 1, p. 71–106, 2005. Citado na página 88.
- PALMER, M.; GILDEA, D.; XUE, N. **Semantic Role Labeling**. [S.l.]: Morgan & Caypool Publishers, 2010. Citado nas páginas 51 e 87.
- PARAVIA, R. de P. P.; SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. Sistema para apoio à rotulação manual de textos utilizando aprendizado de máquina. In: **SIICUSP: Anais do 23o. Simpósio Internacional de Iniciação Científica da USP**. [S.l.: s.n.], 2015. Citado nas páginas 33 e 179.
- PERSSON, J.; JOHANSSON, R.; NUGUES, P. Text categorization using predicate-argument structures. In: **NODALIDA 2009: Proceedings of the 17th Nordic Conference of Computational Linguistics**. [S.l.: s.n.], 2009. p. 142–149. Citado nas páginas 33, 87, 89, 90, 97, 98, 103, 107, 163, 186 e 187.
- PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In: **EASE 2008: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering**. Swinton, UK, UK: British Computer Society, 2008. p. 68–77. Citado na página 53.
- PIETROFORTE, A. V. S. Semântica lexical. In: **Introdução à Lingüística II: Princípios de Análise**. [S.l.]: Editora Contexto, 2010. p. 111–135. Citado na página 49.
- PILEHVAR, M. T.; NAVIGLI, R. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. **Artificial Intelligence**, v. 228, p. 95 – 128, 2015. Citado na página 51.
- PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; MANANDHAR, S.; ANDROUTSOPOULOS, I. Semeval-2015 task 12: Aspect based sentiment analysis. In: **SemEval 2015: Proceedings of the 9th International Workshop on Semantic Evaluation**. [S.l.: s.n.], 2015. p. 486–495. Citado nas páginas 65 e 189.
- REINSEL, D.; GANTZ, J.; RYDNING, J. **Data Age 2025: The Evolution of Data to Life-Critical**. 2017. IDC White Paper. Disponível em: <<https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>>. Acesso em: 07/03/2018. Citado na página 25.
- REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicações**. [S.l.]: Editora Manole, 2003. Citado na página 25.
- RIEMER, N. **Introducing Semantics**. [S.l.]: Cambridge University Press, 2010. (Cambridge Introductions to Language and Linguistics). Citado nas páginas 48 e 58.

- ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional. Citado nas páginas [40](#), [41](#), [42](#), [45](#), [47](#), [113](#), [191](#) e [194](#).
- 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: **ICDM'2012: Proceedings of the 12th IEEE International Conference on Data Mining**. Bruxelas: [s.n.], 2012. v. 1, p. 1086–1091. Citado na página [44](#).
- ROSSI, R. G.; LOPES, A. de A.; FALEIROS, T. de P.; REZENDE, S. O. Inductive model generation for text classification using a bipartite heterogeneous network. **Journal of Computer Science and Technology**, v. 29, n. 3, p. 361–375, 2014. Citado nas páginas [45](#), [94](#), [186](#) e [191](#).
- ROSSI, R. G.; LOPES, A. de A.; REZENDE, S. O. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. **Information Processing & Management**, v. 52, n. 2, p. 217–257, 2016. Citado nas páginas [45](#) e [191](#).
- ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. **Benchmarking Text Collections for Classification and Clustering Tasks**. [S.l.], 2013. Relatório Técnico 395, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado nas páginas [15](#), [61](#), [62](#), [63](#), [156](#) e [157](#).
- ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53 – 65, 1987. Citado na página [66](#).
- SALEIRO, P.; MILIC-FRAYLING, N.; RODRIGUES, E. M.; SOARES, C. Relink: A research framework and test collection for entity-relationship retrieval. In: **SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.]: ACM, 2017. p. 1273–1276. Citado na página [166](#).
- SCHEICHER, R. B.; SINOARA, R. A.; KOGA, N. J.; REZENDE, S. O. Uso de expressões do domínio na classificação automática de documentos. In: **ENIAC 2016: Anais do XIII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.: s.n.], 2016. p. 625–636. Citado nas páginas [75](#), [76](#) e [78](#).
- SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. Citado nas páginas [25](#), [45](#), [46](#), [47](#), [48](#), [60](#) e [156](#).
- SHEHATA, S.; KARRAY, F.; KAMEL, M. S. An efficient model for enhancing text categorization using sentence semantics. **Computational Intelligence**, Wiley Online Library, v. 26, n. 3, p. 215–231, 2010. Citado nas páginas [87](#), [89](#) e [90](#).
- SILVA, T. P. da. **Normalização textual e indexação semântica aplicadas da filtragem de SMS spam**. Dissertação (Mestrado) — Centro de Ciências em Gestão e Tecnologia, Universidade Federal de São Carlos, 2016. Programa de Pós-Graduação em Ciências de Computação. Citado na página [164](#).
- SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. **Journal of the Brazilian Computer Society**, v. 23, n. 9, p. 1–20, 2017. Citado nas páginas [27](#), [39](#), [53](#), [55](#), [56](#) e [57](#).

- 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 64.
- SINOARA, R. A.; ROSSI, R. G.; REZENDE, S. O. Semantic role-based representations in text classification. In: **ICPR 2016: Proceedings of the 23rd International Conference on Pattern Recognition**. [S.l.: s.n.], 2016. p. 2314–2319. Citado na página 28.
- SINOARA, R. A.; SCHEICHER, R. B.; REZENDE, S. O. Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity. In: **CIDM'17: Proceedings of the 2017 IEEE Symposium on Computational Intelligence and Data Mining**. [S.l.: s.n.], 2017. p. 2057–2064. Citado na página 69.
- 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: **IDEAS'14: Proceedings of the 18th International Database Engineering & Applications Symposium**. [S.l.]: ACM, 2014. p. 57–66. Citado nas páginas 28, 87, 164 e 186.
- SOLÉ, R. V.; MURTRA, B. C.; VALVERDE, S.; STEELS, L. Language networks: their structure, function and evolution. **Trends in Cognitive Sciences**, p. 1–9, 2005. Citado na página 44.
- SPANAKIS, G.; SIOLAS, G.; STAFYLOPATIS, A. Exploiting wikipedia knowledge for conceptual hierarchical clustering of documents. **Computer Journal**, v. 55, n. 3, p. 299–312, 2012. Citado nas páginas 28, 43 e 87.
- STAVRIANOU, A.; ANDRITSOS, P.; NICOLOYANNIS, N. Overview and semantic issues of text mining. **SIGMOD Record**, v. 36, n. 3, p. 23–34, 2007. Citado na página 44.
- SUN, Y.; HAN, J. **Mining Heterogeneous Information Networks: Principles and Methodologies**. [S.l.]: Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Data Mining and Knowledge Discovery). Citado na página 44.
- TURNER, V.; GANTZ, J. F.; REINSEL, D.; MINTON, S. **The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things**. 2014. IDC White Paper. Disponível em: <<http://www.emc.com/leadership/digital-universe/index.htm>>. Acesso em: 07/12/2017. Citado na página 25.
- TURNER, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **Journal of Artificial Intelligence Research**, AI Access Foundation, USA, v. 37, n. 1, p. 141–188, 2010. Citado nas páginas 29, 30, 40, 44, 68, 111, 112 e 193.
- VAPNIK, V.; VASHIST, A. A new learning paradigm: Learning using privileged information. **Neural Networks**, v. 22, n. 5-6, p. 544–557, 2009. Citado na página 164.
- VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining**, John Wiley & Sons, Inc., New York, NY, USA, v. 3, n. 4, p. 209–235, 2010. Citado na página 66.
- VIEIRA, R.; GONÇALVES, P. N.; SOUZA, J. G. C. de. Processamento computacional de anáfora e correferência. **Revista de Estudos da Linguagem**, v. 16, n. 1, 2008. Citado na página 51.
- WIDDOWS, D. **Geometry and Meaning**. [S.l.]: CSLI Publications, 2004. Citado nas páginas 29 e 69.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2nd. ed. [S.l.]: Morgan Kaufmann, 2005. Citado na página [191](#).

WU, J.; DANG, Y.; PAN, D.; XUAN, Z.; LIU, Q. Textual knowledge representation through the semantic-based graph structure in clustering applications. In: **HICSS 2010: Proceedings of the 43rd Hawaii International Conference on System Sciences**. [S.l.: s.n.], 2010. p. 1–8. Citado na página [44](#).

WU, J.; XUAN, Z.; PAN, D. Enhancing text representation for classification tasks with semantic graph structures. **International Journal of Innovative Computing, Information and Control (ICIC)**, v. 7, n. 5, 2011. Citado na página [44](#).

ZHENG, W.; AN, L.; XU, Z. Dimensionality reduction by combining category information and latent semantic index for text categorization. **Journal of Information and Computational Science**, v. 10, n. 8, p. 2463–2469, 2013. Citado na página [26](#).

ZRIGUI, M.; AYADI, R.; MARS, M.; MARAOUI, M. Arabic text classification framework based on latent dirichlet allocation. **Journal of Computing and Information Technology**, SRCE-Sveučilišni računski centar, v. 20, n. 2, p. 125–140, 2012. Citado na página [113](#).

PROCESSO DE ROTULAÇÃO DA COLEÇÃO *BEST SPORTS - TOP 4*

Com o objetivo de viabilizar a avaliação das propostas deste trabalho em diferentes níveis de complexidade semântica, foi realizado um processo de rotulação manual da coleção de documentos *BEST sports - Top 4* (*BS-Top4*). Para isso, um novo objetivo de classificação da coleção de documentos foi definido e, a partir desse objetivo, foi desenvolvido o guia de rotulação. Esse guia, apresentado na [Figura 45](#), contém a especificação da tarefa e instruções para sua execução. O processo de rotulação manual da coleção *BS-Top4* foi realizado na ferramenta RotuLABIC¹ ([PARAVIA et al., 2015](#)). Os parâmetros de configuração do processo de rotulação na ferramenta são apresentados na [Figura 46](#). O processo foi realizado por um anotador, seguindo o guia de rotulação.

¹ RotuLABIC: <http://labic.icmc.usp.br/material/14>

Figura 45 – Guia para rotulação da coleção *BS-Top4*

GUIA PARA ROTULAÇÃO DE TEXTOS - BESTSPORTS TOP 4

Base de Textos: Bestsports - Top4

Total de documentos: 283

Essa é uma base de textos formada pela 4 classes com maior número de notícias da coleção Best sports:

- ESP_FORMULA_1: 91 notícias
- ESP_FUTEBOL: 68 notícias
- ESP_MOTOVELOCIDADE: 64 notícias
- ESP_TENIS: 60 notícias

O objetivo desse processo de rotulação é classificar os documentos em relação à quem ganhou a competição relatada na notícia, sendo que estamos interessados nos atletas brasileiros. Assim, as notícias devem ser classificadas de acordo com o esporte relatado (Formula 1, Futebol, Motovelocidade ou Tênis) e se há relato sobre a vitória de um atleta brasileiro.

Rótulos:

- ESP_FORMULA_1-BRvenceu

Caso piloto brasileiro tenha ficado entre os três primeiros colocados.

- ESP_FORMULA_1-BRnaoVenceu

Caso piloto brasileiro não tenha ficado entre os três primeiros colocados.

- ESP_FORMULA_1-BRnaoCitado

Caso nenhum piloto brasileiro tenha sido citado na notícia.

- ESP_FORMULA_1-Ruido

Caso a notícia não relate uma etapa do campeonato.

- ESP_FUTEBOL-BRvenceu

Caso a equipe brasileira tenha vencido uma partida ou ficado entre os três primeiros colocados.

- ESP_FUTEBOL-BRnaoVenceu

Caso a equipe brasileira tenha perdido uma partida.

- ESP_FUTEBOL-BRnaoCitado

Caso a equipe brasileira não tenha sido citada na notícia.

- ESP_FUTEBOL-Ruido

Caso a notícia não relate o resultado de uma partida ou campeonato.

- ESP_MOTOVELOCIDADE-BRvenceu

Caso piloto brasileiro tenha ficado entre os três primeiros colocados.

- ESP_MOTOVELOCIDADE-BRnaoVenceu

Caso piloto brasileiro não tenha ficado entre os três primeiros colocados.

- ESP_MOTOVELOCIDADE-BRnaoCitado

Caso nenhum piloto brasileiro tenha sido citado na notícia.

- ESP_MOTOVELOCIDADE-Ruido

Caso a notícia não relate uma etapa do campeonato.

- ESP_TENIS-BRvenceu

Caso atleta brasileiro tenha vencido uma partida ou ficado entre os três primeiros colocados.

- ESP_TENIS-BRnaoVenceu

Caso atleta brasileiro não tenha vencido ou não tenha ficado entre os três primeiros colocados.

- ESP_TENIS-BRnaoCitado

Caso nenhum atleta brasileiro tenha sido citado na notícia.

- ESP_TENIS-Ruido

Caso a notícia não relate uma partida ou campeonato.

Fonte: Elaborada pelo autor.

Figura 46 – Configuração do processo de rotulação da coleção *BS-Top4* na ferramenta RotuLABIC

Sistema de Apoio à Rotulação Manual de Textos -- Olá, RobertaSinoara! Perfil Manual Sobre Sair

Criar Novo Processo de Rotulação

Nome do Processo de Rotulação

Anexar instruções para os rotuladores

Anexar documentos a serem rotulados

Multirótulo ☐ Sim ☒ Não

Tipo de rótulos ☒ Fixados ☐ Permitir Sugestões

Taxa mínima de concordância de rótulos (documentos)

Adicionar opções de rótulos

	Adicionar
ESP_FORMULA_1-BRvenc	<input type="button" value="✖"/>
ESP_FORMULA_1-BRnao	<input type="button" value="✖"/>
ESP_FORMULA_1-BRnao	<input type="button" value="✖"/>
ESP_FORMULA_1-Ruido	<input type="button" value="✖"/>
ESP_FUTEBOL-BRvenceu	<input type="button" value="✖"/>
ESP_FUTEBOL-BRnaoVer	<input type="button" value="✖"/>
ESP_FUTEBOL-BRnaoCita	<input type="button" value="✖"/>
ESP_FUTEBOL-Ruido	<input type="button" value="✖"/>
ESP_MOTOVELOCIDADE	<input type="button" value="✖"/>
ESP_MOTOVELOCIDADE	<input type="button" value="✖"/>
ESP_MOTOVELOCIDADE	<input type="button" value="✖"/>
ESP_MOTOVELOCIDADE	<input type="button" value="✖"/>
ESP_TENIS-BRvenceu	<input type="button" value="✖"/>
ESP_TENIS-BRnaoVenceu	<input type="button" value="✖"/>
ESP_TENIS-BRnaoCitado	<input type="button" value="✖"/>
ESP_TENIS-Ruido	<input type="button" value="✖"/>

Algoritmo para sugestão de rótulos

Idioma dos documentos

Taxa para reset (algoritmo transdutivo)

Fonte: Elaborada pelo autor.

O processo foi documentado conforme pontos de incertezas surgiram e foram discutidos com um especialista do domínio. A seguir são apresentadas as decisões tomadas e que foram seguidas durante o processo de rotulação, complementando o guia apresentado na [Figura 45](#).

Para documentos de Fórmula 1:

- qualquer posição no pódio foi considerada uma vitória;
- documentos com informações sobre a temporada e apresentação de carros foram rotulados como ESP_FORMULA_1-Ruido;

- a definição do *grid* de largada (notícias sobre treinos oficiais) foi considerada como um resultado de competição;
- notícias sobre testes coletivos (sem classificação) foram rotuladas como ESP_FORMULA-1-Ruido.

Para documentos de Futebol:

- o interesse está em resultado de partidas;
- notícias que reportam empates de equipe brasileira foram rotulados como ESP_FUTEBOL-Ruido;
- colocações no campeonato não são consideradas vitórias ou derrotas;
- notícias sobre *Ranking* FIFA foram rotuladas como ESP_FUTEBOL-Ruido.

Para documentos de Motovelocidade:

- segue as mesmas definições de documentos de *Fórmula 1*;
- notícias sobre *pole* provisória foram rotuladas como ESP_MOTOVELOCIDADE-Ruido.

Para documentos de Tênis:

- o interesse está em resultado de partidas;
- documentos que reportam tanto vitória quanto derrota de brasileiro(s) foram rotulados como ESP_TENIS-Ruido.

A distribuição dos documentos nas 16 classes definidas é apresentada na [Tabela 11](#).

A partir dos documentos rotulados e exportados da ferramenta RotuLABIC, foram gerados dois *datasets* alternativos para a coleção de textos: *BS-semantic* e *BS-topic-semantic*. Tais *datasets* são descritos na Seção 2.5.2 (Página 64). Após a rotulação, a classe “Ruido” foi renomeada para “Não definido” por ser uma nomenclatura considerada mais adequada aos textos que descreve. O rótulo “Não definido” refere-se a documentos que não reportam os resultados de uma competição, reportam empates ou reportam tanto vitória quanto derrota de atletas brasileiros.

Tabela 11 – Distribuição dos documentos da coleção *BEST sports - Top 4*

Classe	# Documentos
ESP_FORMULA_1-BRvenceu	48
ESP_FORMULA_1-BRnaoVenceu	29
ESP_FORMULA_1-BRnaoCitado	0
ESP_FORMULA_1-Ruido	14
ESP_FUTEBOL-BRvenceu	29
ESP_FUTEBOL-BRnaoVenceu	12
ESP_FUTEBOL-BRnaoCitado	8
ESP_FUTEBOL-Ruido	19
ESP_MOTOVELOCIDADE-BRvenceu	5
ESP_MOTOVELOCIDADE-BRnaoVenceu	28
ESP_MOTOVELOCIDADE-BRnaoCitado	12
ESP_MOTOVELOCIDADE-Ruido	19
ESP_TENIS-BRvenceu	11
ESP_TENIS-BRnaoVenceu	19
ESP_TENIS-BRnaoCitado	17
ESP_TENIS-Ruido	13

CONFIGURAÇÃO EXPERIMENTAL

Os modelos de representação de textos considerando informações semânticas propostos neste trabalho foram avaliados segundo a performance de classificação de documentos. Foram considerados diferentes algoritmos de Aprendizado de Máquina, com o objetivo de avaliar as representações em configurações diversificadas. Também foi considerada a combinação de classificadores gerados com a BOW e com as representações propostas, visando unir as vantagens obtidas com ambas as representações. As avaliações experimentais realizadas foram padronizadas, visando a comparação de diferentes modelos de representação de textos.

Nas próximas seções são apresentados os detalhes da configuração experimental adotada. É descrito como as representações foram geradas e são apresentados os algoritmos de Aprendizado de Máquina e seus parâmetros, a abordagem de combinação de classificadores e os critérios de avaliação utilizados.

B.1 Geração das representações

Antes da geração das representações, a coleção de documentos *BS-Top4*, apresentados na Seção 2.5.2 (Página 64), passou por um processo de limpeza, no qual foi realizado: (i) inclusão de ponto final no título das notícias; (ii) substituição do caractere “-”, que separava duas frases no título de algumas notícias, por ponto final; (iii) exclusão do código HTML “
”, que era encontrado em alguns documentos; e (iv) remoção de alguns caracteres especiais, como aspas e °. Para as demais coleções de documentos, não foi necessário executar esse processo de limpeza.

Nessa seção são apresentados os detalhes da geração das representações consideradas neste trabalho, tais como técnicas de pré-processamento específicas para a construção de cada representação e ferramentas auxiliares utilizadas para geração dos atributos. As representações estão divididas em duas subseções: representações de comparação e representações propostas.

B.1.1 Representações de comparação (baselines)

As seguintes representações da literatura foram utilizadas como modelos de comparação nas avaliações experimentais apresentadas neste trabalho.

- Capítulos 3, 4 e 5: BOW.
- Capítulo 4: pred.Arg1 e BOW+pred.Arg1.
- Capítulo 5: LDA, Word2Vec e Doc2Vec.

BOW

A representação *bag-of-words* (**BOW**) foi utilizada como *baseline* em todas as avaliações experimentais. As seguintes técnicas de pré-processamento foram aplicadas para a geração dessa representação.

- Remoção de acentos da coleção *BS-Top4* (documentos em português).
- Remoção de *stopwords*. Para as coleções com textos em inglês foi utilizada a *stoplist* para inglês da ferramenta Text Categorization¹. Para a coleção *BS-Top4* foi utilizada a *stoplist* para português da ferramenta Text Categorization, acrescida de *stopwords* do domínio, que foram identificadas em um trabalho anterior (SINOARA *et al.*, 2014).
- Remoção de números e caracteres especiais.
- Radicalização dos termos. Para as coleções com textos em inglês, foi utilizado o algoritmo de Porter². Para as coleções com textos em português, foi utilizado o algoritmo RSLP³.
- Remoção de termos presentes em apenas um documento.
- Uso da medida TF para cálculo do peso dos termos. Essa medida foi selecionada com base nos resultados apresentados por Rossi *et al.* (2014), que compara as medidas TF e TF-IDF utilizando 14 coleções de textos e seis algoritmos de classificação. Em tal avaliação experimental, a maioria dos algoritmos obtiveram melhores resultados com o uso de TF.

pred.Arg1 e BOW+pred.Arg1

As representações **pred.Arg1** e **BOW+pred.Arg1** são implementações de modelos de representação baseados em papéis semânticos propostos por Persson, Johansson e Nugues (2009)

¹ Text Categorization tool: <http://sites.labc.icmc.usp.br/ragero/thesis/text_categorization_tool/>

² Porter Stemmer: <<http://tartarus.org/~martin/PorterStemmer/>>

³ RSLP Stemmer: <<http://www.inf.ufrgs.br/~viviane/rsdp/index.htm>>

e apresentados na Seção 4.2 (Página 89). Tais representações foram construídas com apoio da ferramenta PRED.ARG, desenvolvida no LABIC⁴.

A representação chamada nesta tese de **pred.Arg1** corresponde à representação identificada pelo código c010010 em (PERSSON; JOHANSSON; NUGUES, 2009), que é a união do conjunto de atributos *Predicates* (código c010000) com o conjunto de atributos *Argument 1* (c000010). De maneira análoga, a representação chamada de **BOW+pred.Arg1** nesta tese corresponde à representação identificada por c110010, sendo a união da *bag-of-words* com os conjuntos de atributos da representação **pred.Arg1**.

As representações foram geradas seguindo as indicações de Persson, Johansson e Nuges (2009), porém utilizando as mesmas ferramentas para a desambiguação e anotação de papéis semânticos utilizadas para gerar as representações propostas neste trabalho. Assim, os textos foram desambiguados utilizando o sistema Babelfy (MORO; RAGANATO; NAVIGLI, 2014), conforme o procedimento descrito na Seção B.1.2. Os rótulos de papéis semânticos utilizados foram os mesmos das representações *SR-based Terms* e *SR-based Sentences* (Seção B.1.2). A representação **BOW** descrita na Seção B.1.1 foi utilizada para a geração da representação **BOW+pred.Arg1**. Os pesos dos atributos foram calculados utilizando a medida TF.

LDA

A representação **LDA** foi gerada utilizando a implementação da ferramenta Mallet⁵ (MCCALLUM, 2002). Essa ferramenta implementa o método *Gibbs Sampling* para inferir a probabilidade *a posteriori* no modelo LDA e também possui métodos para estimar automaticamente os parâmetros α e β . Foram utilizadas as mesmas *stoplists* usadas para gerar a **BOW**.

Para gerar a representação foram geradas 100 soluções de 300 tópicos para cada coleção de documento. A matriz documento-tópico de uma dessas soluções, selecionada aleatoriamente, foi utilizada como representação da coleção. Utilizou-se 300 tópicos para a representação **LDA** por ser o mesmo número de dimensões das demais representações avaliadas que possuem número fixo de dimensões.

Word2Vec

Na representação **Word2Vec**, cada documento é representado pelo centróide dos vetores de suas palavras ou expressões. Foi utilizado o mesmo modelo Word2Vec pré-treinado utilizado na geração das representações *Babel2Vec* e *NASARI+Babel2Vec*, descrito na Seção B.1.2. Para a construção dessa representação utilizou-se a opção “Bag of Vectors” da ferramenta *Text Representations*⁶. Para cada coleção de textos, o parâmetro *n-grams* foi configurado com o valor

⁴ PRED.ARG: <<http://labic.icmc.usp.br/material/16>>

⁵ Mallet: <<http://mallet.cs.umass.edu>>

⁶ *Text Representations*: <<https://github.com/joao4ntunes/text-mining/>>

máximo de palavras contidas em expressões que foi encontrado na *embeddings*, sendo 3 para *BS-Top4*, 4 para *SE-ABSA15* e 4 para *BBC*.

Doc2Vec

A representação **Doc2Vec** foi gerada utilizando a biblioteca Gensim⁷, por meio da ferramenta Doc2Vec tools⁸. Foi treinado um modelo para cada idioma, ambos contendo 300 dimensões (mesmo valor utilizado nas demais representações com dimensionalidade fixa que foram geradas nesse trabalho).

Para esse treinamento utilizou-se a Wikipedia, *dumps* de 20 novembro de 2017^{9,10}. A Wikipedia foi selecionada como o corpus para treinamento do modelo por não haver disponível modelos Doc2Vec pré-treinados, como existem para Word2Vec. Além disso, além de conter versões tanto para inglês como para português, a Wikipedia também foi utilizada na construção do conjunto de vetores NASARI *embedded* utilizado nesse trabalho (Seção B.1.2).

B.1.2 Representações propostas

Neste trabalho foram propostas e avaliadas as seguintes representações.

- Capítulo 3: *Uni-based* e *gBoED*.
- Capítulo 4: *SR-based Terms* e *SR-based Sentences*.
- Capítulo 5: *Babel2Vec* e *NASARI+Babel2Vec*.

Os detalhes da geração dessas representações para as avaliações experimentais são apresentados a seguir.

gBoED

Para a geração da representação ***gBoED*** foi necessário definir as listas de termos do domínio e de identificadores de classe para as duas coleções de documentos utilizadas na avaliação experimental. Para a coleção *BS-Top4*, considerou-se o cenário de classificação baseado no desempenho de atletas brasileiros por ser um cenário do segundo nível de complexidade semântica, conforme definido na Seção 2.5 (Página 58). Assim, um usuário especialista do domínio forneceu as seguintes listas para a coleção *BS-Top4*.

- *Domain_Keywords*: lista de nomes de atletas brasileiros.

⁷ Gensim: <<https://radimrehurek.com/gensim/>>

⁸ Doc2Vec tools: <<http://labic.icmc.usp.br/material/18>>

⁹ Wikipedia - inglês: <<https://dumps.wikimedia.org/enwiki/>>

¹⁰ Wikipedia - português: <<https://dumps.wikimedia.org/ptwiki/>>

- *Class_Keywords1*: lista de verbos normalmente utilizados para expressar vitórias.
- *Class_Keywords2*: lista de verbos normalmente utilizados para expressar derrotas.

Já para a coleção *SE-ABSA15*, foram utilizadas as seguintes listas de termos do domínio e de identificadores de classe para a construção da *gBOED*.

- *Domain_Keywords*: lista de aspectos de hotéis, restaurantes e *laptops* extraídos dos textos de opinião da coleção *SE-ABSA15*, disponibilizados por Pontiki *et al.* (2015).
- *Class_Keywords1*: lista de palavras positivas para o idioma inglês, originalmente utilizada no trabalho de Hu e Liu (2004).
- *Class_Keywords2*: lista de palavras negativas para o idioma inglês, originalmente utilizada no trabalho de Hu e Liu (2004)¹¹.

Para as duas coleções de documentos, as expressões do domínio foram construídas com os termos das listas radicalizados. Para essa radicalização foram utilizados os mesmos algoritmos utilizados para radicalizar os termos da representação **BOW**. Os pesos dos atributos foram calculados utilizando a medida TF.

Uni-based

Para a geração das representações *Uni-based* as coleções de documentos foram desambiguadas utilizando a ferramenta Babelify, seguindo as mesmas configurações usadas para gerar as representações *NASARI+Babel2Vec* e *Babel2Vec*, descritas também nessa seção. Além disso, foi utilizada a versão do idioma inglês do conjunto de vetores *NASARI unified*¹², disponibilizado por Camacho-Collados, Pilehvar e Navigli (2016).

SR-based Terms e SR-based Sentences

Para a geração das representações *SR-based Terms* e *SR-based Sentences* é necessário processar os documentos para se obter as anotações de classes morfofossintáticas e de papéis semânticos. Para os *datasets* da coleção *SE-ABSA15*, as anotações foram obtidas com a ferramenta SENNA¹³. Essa ferramenta, específica para o processamento de textos em inglês, implementa o aprendizado baseado em redes neurais proposto por Collobert *et al.* (2011) para a resolução de diferentes tarefas de PLN, como anotação de classes morfofossintáticas, reconhecimento de entidades nomeadas e anotação de papéis semânticos. Os resultados reportados no trabalho de

¹¹ As listas de palavras positivas e negativas foram obtidas em: <<http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>>

¹² Os vetores *NASARI unified* pré-treinado foi obtido em: <<http://lcl.uniroma1.it/nasari/>>

¹³ SENNA: <<https://ronan.collobert.com/senna/>>

Collobert *et al.* (2011) indicam que o sistema de anotação de papéis semânticos apresenta resultados próximos ao estado da arte. Segundo os autores, o método implementado na ferramenta SENNA obteve F1 de 75,49% no conjunto de testes do CoNLL 2005, enquanto o sistema estado da arte apresentou F1 de 77,92% para a mesma tarefa. Apesar de obter resultado um pouco inferior em relação à medida F1, SENNA apresenta um tempo de processamento menor e menos consumo de memória (COLLOBERT *et al.*, 2011).

A ferramenta SENNA foi utilizada com a opção *-posvbs*, configurando assim o rotulador de papéis semânticos para considerar os verbos identificados pelo rotulador de classes morfossintáticas e não ignorar certos verbos, como o verbo “to be”. As classes morfossintáticas consideradas para essa coleção foram adjetivos, nomes, pronomes e verbos.

As anotações de classes morfossintáticas e de papéis semânticos dos *datasets* da coleção *BEST sports Top 4* foram obtidas com a ferramenta *nlpnet*¹⁴ (FONSECA; ROSA, 2013b; FONSECA; ROSA, 2013a). Essa ferramenta foi desenvolvida seguindo a mesma abordagem da ferramenta SENNA, utilizada para os textos em inglês. Neste trabalho, foram utilizados os modelos pré-treinados disponibilizados juntamente com ferramenta: POS (Portuguese) e SRL (Portuguese). As classes morfossintáticas utilizadas para essa coleção de textos foram nomes, pronomes e verbos.

Os termos foram radicalizados utilizando implementações dos mesmos algoritmos utilizados para gerar a BOW. Assim como na representação BOW, a medida TF também foi utilizada para calcular o peso dos atributos.

Babel2Vec* e *NASARI+Babel2Vec

Para a geração das representações ***Babel2Vec*** e ***NASARI+Babel2Vec*** é necessário processar os documentos para se obter os termos desambiguados. Todas as coleções de documentos utilizadas nas avaliações experimentais foram desambiguadas utilizando a API do sistema Babelfy¹⁵ (MORO; CECCONI; NAVIGLI, 2014), descrito na Seção 5.3 (Página 115). Os seguintes parâmetros foram configurados para processamento das coleções de documentos.

- *Language*. O idioma foi configurado de acordo com o idioma da coleção de documentos sendo processada.
- *Matching Type*. Foi utilizada a opção “EXACT_MATCHING” para os textos em inglês e “PARTIAL_MATCHING” para os textos em português. Tal opção foi tomada com base na observação da desambiguação realizada pelo sistema Babelfy para dois documentos selecionados aleatoriamente. A análise realizada é apresentada na Seção C.1 do Apêndice C (Página 195).

¹⁴ *nlpnet* tool: <<http://nilc.icmc.usp.br/nlpnet/>>

¹⁵ Babelfy API: <<http://babelfy.org/guide>>

- *Threshold*. Para favorecer o retorno da desambiguação em relação ao sentido mais comum dos fragmentos (*most common sense*), o limiar para o score de desambiguação foi configurado para 0,6, valor um pouco abaixo do padrão do sistema (0,7).

A construção das representações **Babel2Vec** e **NASARI+Babel2Vec** foi realizada conforme o processo apresentado na Seção 5.4 (Página 116). Além dos documentos desambiguados, a construção das representações (apresentada no Algoritmo 1, Página 121) também recebe como entrada *embeddings* pré-treinadas.

Para a geração da representação **Babel2Vec** foi utilizado o modelo para palavras e expressões disponibilizado juntamente com a ferramenta Word2Vec¹⁶. Tal modelo foi treinado com documentos do corpus *Google News*, composto por cerca de 100 bilhões de palavras. O modelo é formado por 3 milhões de vetores de 300 dimensões.

Para a geração da representação **NASARI+Babel2Vec**, além do modelo anterior, também foi utilizado um subconjunto do modelo pré-treinado NASARI *embedded*¹⁷. O subconjunto é formado apenas por conceitos relacionados a páginas da Wikipedia com pelo menos cinco *backlinks* na própria Wikipedia. O modelo NASARI *embedded*, versão para o idioma inglês, que foi treinado usando o mesmo modelo Word2Vec pré-treinado com documentos do corpus *Google News*. Assim, os dois modelos utilizados compartilham o mesmo espaço semântico (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016) de 300 dimensões.

B.2 Algoritmos de aprendizado supervisionado

Neste trabalho foram aplicados sete algoritmos de classificação indutiva, tradicionais e estado da arte, sendo cinco algoritmos da biblioteca Weka (WITTEN; FRANK, 2005) e dois algoritmos baseados em redes bipartidas (ROSSI *et al.*, 2014; ROSSI; LOPES; REZENDE, 2016). Apesar das representações avaliadas neste trabalhos sejam representações no modelo espaço-vetorial, a utilização de algoritmos baseados em redes bipartidas também é possível. As redes bipartidas podem ser obtidas por mapeamento direto de representações documento-termo, ou seja, representações no modelo espaço-vetorial (ROSSI; LOPES; REZENDE, 2016).

Com o objetivo de se avaliar as representações em cenários bem diversificados, buscou-se diferentes variações dos parâmetros utilizados para cada algoritmo. Os valores selecionados para os parâmetros dos algoritmos foram baseados nos valores utilizados na extensa avaliação experimental de classificação indutiva supervisionada realizada por Rossi (2016). Assim, a seleção de parâmetros foi realizada visando a diversificação dos modelos e o uso de configurações utilizadas em outros trabalhos da literatura e que apresentam bons resultados. Os algoritmos e os parâmetros utilizados nas avaliações experimentais deste trabalho são apresentados a seguir.

¹⁶ O modelo Word2Vec pré-treinado foi obtido em: <<https://code.google.com/archive/p/word2vec/>>

¹⁷ O modelo NASARI *embedded* pré-treinado foi obtido em: <<http://lcl.uniroma1.it/nasari/>>

- *Naive Bayes* (NB)

- *Multinomial Naive Bayes* (MNB)

Por não ser compatível com atributos com valores negativos, esse algoritmo foi utilizado apenas nos experimentos apresentados nos Capítulos 3 e 4.

- C4.5

Foi utilizada a implementação da ferramenta Weka (J48). Foram utilizados os níveis de confiança 0, 15, 0,20 e 0,25.

- *Support Vector Machine* (SVM)

Foi utilizado o algoritmo *Sequential Minimal Optimization* (SMO), implementação da ferramenta Weka. Foram utilizados três tipos de *kernel*: *linear*, *polynomial* (com expoente = 2) e (*radial basis function*). Os valores de C utilizados para cada tipo de *kernel* foram: 0, 10^{-5} , 10^{-4} , 10^{-3} , 10^{-2} , 10^{-1} , 10^0 , 10^1 , 10^2 , 10^3 , 10^4 e 10^5 .

- *k-Nearest Neighbors* (k-NN)

Foi utilizada a implementação da ferramenta Weka (IBk). O algoritmo foi utilizado com e sem o voto ponderado pela distância entre os exemplos. Foram aplicadas duas medidas de distância: Cosseno e Euclidiana. Os valores de k utilizados foram: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45 e 55.

- IMBHN^C

Versão do algoritmo *Inductive Model based on Bipartite Heterogeneous Networks*, algoritmo baseado em redes heterogêneas bipartidas. Essa versão do algoritmo altera as informações de classe dos termos caso as informações de classe correntes produzam um erro de classificação. Foram utilizadas as taxas de correção de erro de 0,01, 0,05, 0,1 e 0,5. O número máximo de iterações foi definido em 1000 e o critério de parada foi definido como sendo o limiar de 0,01 para o erro quadrático médio.

- IMBHN^R

Versão do algoritmo *Inductive Model based on Bipartite Heterogeneous Networks*, algoritmo baseado em redes heterogêneas bipartidas. Essa versão do algoritmo realiza uma regressão para induzir as informações de classe dos termos. Foram utilizadas as taxas de correção de erro de 0,01, 0,05, 0,1 e 0,5. O número máximo de iterações foi definido em 1000 e o critério de parada foi definido como sendo o limiar de 0,01 para o erro quadrático médio.

B.3 Combinação de classificadores

A combinação de representações propostas com a BOW também foi avaliada neste trabalho, visto que os atributos semânticos, apesar de mais ricos semanticamente do que palavras isoladas, são geralmente menos frequentes no conjunto de dados. Consequentemente, as matrizes das representações semânticas são geralmente mais esparsas do que a **BOW**, o que pode afetar a performance de classificação dos algoritmos de AM. Seria necessário um volume muito maior de textos para que os atributos ocorressem com mais frequência no corpus, tornando a matriz menos esparsa e, de acordo com a hipótese de semântica estatística¹⁸, mais útil (TURNERY; PANTEL, 2010).

O uso combinado de duas representações foi realizado por meio de *ensemble* de classificadores. Dois classificadores base foram gerados pelo mesmo algoritmo e com as mesmas configurações de parâmetros; cada um utilizando uma representação a ser combinada. O primeiro classificador base (c_1) foi gerado utilizando a representação **BOW** e o segundo (c_2) foi gerado utilizando uma das representações propostas sendo avaliada. Para classificar um novo exemplo, a combinação dos resultados dos classificadores base foi realizada com três estratégias diferentes:

Most Confident (MC). O novo exemplo recebe o rótulo atribuído pelo classificador que apresenta a maior confiança de classificação (maior probabilidade do exemplo pertencer à classe atribuída).

Sum of Confidences (SC). Para se determinar o rótulo do novo exemplo, é feita a soma das confianças de classificação (probabilidades do exemplo pertencer a cada classe) dadas pelos classificadores c_1 e c_2 . O rótulo que apresenta a maior confiança é atribuído ao exemplo.

Weighted Sum of Confidences (WSC). Assim como na abordagem SC, é feita a soma das probabilidades do exemplo pertencer a cada classe dadas por c_1 e c_2 , porém a soma é ponderada pela acurácia do classificador no conjunto de treinamento. O rótulo que apresenta a maior confiança é atribuído ao exemplo.

Nas três estratégias, foi considerada uma variação de pesos para o voto de cada classificador base. Considerando que w_{c_1} é o peso estabelecido para o voto do classificador c_1 e w_{c_2} é o peso do voto do classificador c_2 , foram considerados os pesos $w_{c_1} = \{0, 1, 0, 3, 0, 5, 0, 7, 0, 9\}$ e $w_{c_2} = 1 - w_{c_1}$. Tais pesos foram considerados com o objetivo de avaliar a performance de classificação considerando-se diferentes distribuições de peso entre os classificadores base.

¹⁸ A hipótese de semântica estatística (*statistical semantics hypothesis*) diz que padrões estatísticos do uso humano de palavras podem ser usados para descobrir o que as pessoas querem dizer (TURNERY; PANTEL, 2010).

B.4 Critérios de avaliação

A performance dos classificadores foram avaliadas por meio das medidas Acurácia (equivalente a Micro-F1) e Macro-F1 (descritas na Seção 2.2.2, Página 44). As medidas foram obtidas por meio de *10-fold cross-validation*. Todos os algoritmos foram avaliados segundo as mesmas partições dos dados no processo de *10-fold cross-validation*. Para avaliar a significância estatística das diferenças entre as representações, os resultados foram submetidos ao teste de Friedman $N \times N$ com pós-teste de Nemenyi (DEMSAR, 2006), considerando nível de confiança de 95%.

As avaliações experimentais foram executadas por meio da ferramenta Text Categorization¹⁹, disponibilizada por (ROSSI, 2016). Os testes de significância estatística foram realizados com o auxílio da ferramenta KEEL²⁰ e pacote *Statistical Comparison of Multiple Algorithms in Multiple Problems* (scmamp)²¹ do ambiente R.

¹⁹ Text Categorization tool: <http://sites.labic.icmc.usp.br/ragero/thesis/text_categorization_tool/>

²⁰ KEEL: <<http://www.keel.es/>>

²¹ scmamp: <<https://cran.r-project.org/web/packages/scmamp/index.html>>

ANÁLISES E RESULTADOS EXTRAS

Nas próximas seções são apresentadas análises extras realizadas durante o desenvolvimento desse trabalho, bem como alguns resultados complementares.

C.1 Análise do parâmetro *Matching Type* do sistema Babelfy

Nessa seção é apresentada uma breve análise realizada para se identificar diferenças no comportamento do sistema Babelfy em relação a variação do parâmetro *Matching Type*. Esse parâmetro define como os sentidos candidatos são selecionados (MORO; CECCONI; NAVIGLI, 2014) e pode ser configurado de duas maneiras: EXACT_MATCHING ou PARTIAL_MATCHING. Utilizando EXACT_MATCHING, o fragmento deve coincidir exatamente com um dos sinônimos do *synset*. Já a opção PARTIAL_MATCHING possibilita que apenas uma substring do fragmento coincida com os sinônimos do *synset*. Essa heurística de *partial matching* é definida em Moro, Raganato e Navigli (2014).

Nos Quadros 8 e 9 são apresentadas as saídas da desambiguação dos dois documentos apresentados a seguir. Os documentos foram selecionados aleatoriamente, sendo que o documento escrito em português pertence à coleção *BS-Top4* e o documento escrito em inglês pertence à coleção *SE-ABSA15*.

Documento em português: *Barrichello é pole em Suzuka e lidera trio sul-americano com Montoya em 2 e Da Matta em 3. Três pilotos sul-americanos garantiram as três primeiras colocações do grid de largada do GP do Japão, última etapa do Campeonato Mundial de Fórmula Um 2003. Com garoa na pista de Suzuka, o brasileiro Rubens Barrichello marcou o tempo de 1m31s713 e conquistou sua nona pole position da carreira. O colombiano Juan Pablo Montoya fez o segundo tempo e o brasileiro Cristiano da Matta conquistou sua melhor posição*

de largada na Fórmula 1, terceiro lugar. A chuva apertou no final e ajudou a garantir o trio sul-americano na ponta. Entre os últimos a entrar na pista, o finlandês Kimi Raikkonen ficou com o oitavo tempo e Michael Schumacher foi apenas o 14. A previsão é de mais chuva para a corrida de domingo.

Documento em inglês: *Very disappointing. Lots of services promised and not provided. The shuttle wasn't running, and the restaurant was closed down - both of those explained to me with because business is low right now around the holidays. So I guess you should make sure it's going to be busy when you go! The business center has been taken apart - the front desk staff did not know if it was going to be back, since they had no idea it had been taken apart and were quite shocked to find out, when they tried to let me in to use it. Vending machines were out of everything except in the lobby. Staff was very friendly and a bit embarrassed about everything that wasn't available. The room itself was fine.*

Nesses quadros, para cada método (EXACT_MATCHING ou PARTIAL_MATCHING), são apresentadas as seguintes informações:

- fragmento do texto que foi desambiguado (coluna “frag”);
- identificador do *synset* na base BabelNet (coluna “id”);
- fonte do retorno (coluna “src”), sendo o próprio método de desambiguação do Babelfy (BFY) ou o sentido mais comum para o fragmento (MCS);
- *score* de desambiguação (coluna “score”);
- indicação se o fragmento foi corretamente desambiguado (coluna “✓”); e
- indicação se o mesmo *synset* foi retornado pelos dois métodos (coluna “=”).

Como pode ser observado no [Quadro 8](#), para o texto escrito em português o uso de EXACT_MATCHING levou à identificação de 76 *synsets*. Analisando tais *synsets* e confrontando-os com o conteúdo do texto, verificou-se que metade dos *synsets* foram corretamente identificados. Já o uso de PARTIAL_MATCHING levou à identificação de 81 *synsets*, sendo 50 (61,73%) corretamente identificados. Comparando os dois métodos, 58 *synsets* foram identificados igualmente nos dois testes, sendo que, desses 58 *synsets*, 33 foram corretamente identificados. Verificou-se que, para esse caso, o método PARTIAL_MATCHING obteve resultados melhores, com um número maior de *synsets* identificados e também um número maior de acertos.

Para o texto escrito em inglês ([Quadro 9](#)), o uso de EXACT_MATCHING levou à identificação de 47 *synsets*, sendo que 28 (59,57%) foram corretamente identificados. Já o uso de PARTIAL_MATCHING resultou em 46 *synsets*, sendo 24 (52,17%) corretos. Comparando os dois métodos, 23 *synsets* foram identificados igualmente nos dois testes, sendo que, desses

Quadro 8 – Comparação da desambiguação para o documento em português

EXACT_MATCHING						PARTIAL_MATCHING				
frag	id	src	score	✓	=	✓	frag	id	src	score
Barrichello	bn:01877435n	BFY	1,0000	✓	y	✓	Barrichello	bn:01877435n	BFY	1,0000
Suzuka	bn:02646813n	BFY	0,9403	✗	n	✓	Suzuka	bn:03620336n	BFY	0,8105
trio	bn:00078313n	MCS	0,0000	✗	y	✗	trio	bn:00078313n	MCS	0,0000
sul-americano	bn:00110954a	BFY	1,0000	✓	y	✓	sul-americano	bn:00110954a	BFY	1,0000
					n	✓	Montoya	bn:00891246n	BFY	0,9385
					n	✓	Da Matta	bn:03492040n	BFY	0,6433
					n	✓	Matta	bn:03492040n	BFY	0,6083
Três	bn:00000100n	BFY	0,8552	✓	y	✓	Três	bn:00000100n	MCS	0,0000
pilotos	bn:00002306n	BFY	0,5347	✗	y	✗	pilotos	bn:00002306n	MCS	0,0000
garantiram	bn:00082877v	MCS	0,0000	✗	y	✗	garantiram	bn:00082877v	MCS	0,0000
primeiras	bn:00103006a	BFY	0,8596	✓	n	✗	primeiras	bn:00103232a	MCS	0,0000
colocações	bn:00063704n	BFY	0,6667	✓	n	✗	colocações	bn:00030615n	MCS	0,0000
					n	✗	grid	bn:00239516n	BFY	0,5909
grid de largada	bn:06796270n	BFY	1,0000	✓	y	✓	grid de largada	bn:06796270n	BFY	1,0000
largada	bn:00069828n	BFY	1,0000	✓	y	✓	largada	bn:00069828n	MCS	0,0000
GP	bn:00037664n	BFY	1,0000	✗	n					
GP do Japão	bn:02771405n	BFY	1,0000	✓	y	✓	GP do Japão	bn:02771405n	BFY	1,0000
Japão	bn:00047985n	BFY	0,9999	✓	y	✓	Japão	bn:00047985n	BFY	0,6521
última	bn:00105773a	MCS	0,0000	✗	y	✗	última	bn:00105773a	MCS	0,0000
etapa	bn:00050548n	MCS	0,0000	✓	y	✓	etapa	bn:00050548n	MCS	0,0000
Campeonato	bn:00017615n	BFY	0,7778	✓	n					
Campeonato Mundial	bn:00469456n	BFY	1,0000	✓	n					
Mundial	bn:06849600n	BFY	1,0000	✗	n	✗	Mundial	bn:00070139n	BFY	0,5580
Fórmula	bn:00032395n	BFY	0,9167	✗	y	✗	Fórmula	bn:00032395n	MCS	0,0000
Fórmula Um	bn:00204507n	BFY	1,0000	✓	y	✓	Fórmula Um	bn:00204507n	BFY	0,9804
garoa	bn:00028809n	BFY	1,0000	✓	y	✓	garoa	bn:00028809n	MCS	0,0000
pista	bn:00065809n	MCS	0,0000	✓	y	✓	pista	bn:00065809n	BFY	0,5412
Suzuka	bn:02646813n	BFY	0,9403	✗	n	✓	Suzuka	bn:03620336n	BFY	0,8105
brasileiro	bn:17327015n	BFY	0,9000	✗	n	✓	brasileiro	bn:00012850n	MCS	0,0000
Rubens	bn:00061810n	BFY	0,9000	✗	n	✓	Rubens	bn:01877435n	BFY	0,8776
Rubens Barrichello	bn:01877435n	BFY	1,0000	✓	y	✓	Rubens Barrichello	bn:01877435n	BFY	1,0000
Barrichello	bn:01877435n	BFY	1,0000	✓	y	✓	Barrichello	bn:01877435n	BFY	1,0000
marcou	bn:00090660v	MCS	0,0000	✗	y	✗	marcou	bn:00090660v	MCS	0,0000
tempo	bn:00077270n	MCS	0,0000	✗	y	✗	tempo	bn:00077270n	MCS	0,0000
conquistou	bn:00082739v	MCS	0,0000	✗	y	✗	conquistou	bn:00082739v	BFY	0,6316
nona	bn:00057751n	MCS	0,0000	✗	y	✗	nona	bn:00057751n	MCS	0,0000
pole position	bn:03473434n	BFY	0,5833	✓	y	✓	pole position	bn:03473434n	BFY	0,8610
					n	✗	position	bn:03460714n	BFY	0,5342
carreira	bn:00025286n	MCS	0,0000	✗	y	✗	carreira	bn:00025286n	MCS	0,0000
colombiano	bn:00020698n	MCS	0,0000	✓	y	✓	colombiano	bn:00020698n	MCS	0,0000
Juan	bn:00805046n	BFY	1,0000	✗	n					
					n	✓	Juan Pablo	bn:00891246n	BFY	0,8327
Juan Pablo Montoya	bn:00891246n	BFY	1,0000	✓	y	✓	Juan Pablo Montoya	bn:00891246n	BFY	1,0000
Pablo	bn:03196952n	BFY	0,8065	✗	n	✓	Pablo	bn:00891246n	BFY	0,5660
					n	✓	Pablo Montoya	bn:00891246n	BFY	1,0000
					n	✓	Montoya	bn:00891246n	BFY	0,9385
fez	bn:00087107v	MCS	0,0000	✓	y	✓	fez	bn:00087107v	MCS	0,0000
segundo	bn:00095974a	BFY	0,9091	✓	y	✓	segundo	bn:00095974a	BFY	0,9091
tempo	bn:00077270n	MCS	0,0000	✗	y	✗	tempo	bn:00077270n	MCS	0,0000
brasileiro	bn:17327015n	BFY	0,9000	✗	n	✓	brasileiro	bn:00012850n	MCS	0,0000
Cristiano	bn:15729447n	BFY	1,0000	✗	n					
					n	✓	Cristiano da	bn:03492040n	BFY	0,9002
Cristiano da Matta	bn:03492040n	BFY	1,0000	✓	y	✓	Cristiano da Matta	bn:03492040n	BFY	1,0000
					n	✓	da Matta	bn:03492040n	BFY	0,6433
					n	✓	Matta	bn:03492040n	BFY	0,6083
conquistou	bn:00082739v	MCS	0,0000	✗	y	✗	conquistou	bn:00082739v	BFY	0,6316
melhor	bn:00098282a	BFY	0,5455	✓	y	✓	melhor	bn:00098282a	BFY	0,5455
posição	bn:00063703n	MCS	0,0000	✗	n	✗	posição	bn:03594712n	BFY	0,5495
largada	bn:00069828n	BFY	1,0000	✓	y	✓	largada	bn:00069828n	MCS	0,0000

Continua na página seguinte.

Quadro 8 – Comparação da desambiguação para o documento em português (continuação)

Continuação da página anterior.										
EXACT_MATCHING					PARTIAL_MATCHING					
frag	id	src	score	✓	=	✓	frag	id	src	score
Fórmula	bn:00032395n	BFY	0,9167	✗	y	✗	Fórmula	bn:00032395n	MCS	0,0000
Fórmula 1	bn:00204507n	BFY	1,0000	✓	n					
terceiro lugar	bn:03330674n	BFY	1,0000	✗	y	✗	terceiro lugar	bn:03330674n	BFY	1,0000
lugar	bn:00062699n	MCS	0,0000	✗	y	✗	lugar	bn:00062699n	MCS	0,0000
chuva	bn:00066032n	BFY	0,7428	✓	y	✓	chuva	bn:00066032n	MCS	0,0000
apertou	bn:00088055v	MCS	0,0000	✗	y	✗	apertou	bn:00088055v	MCS	0,0000
final	bn:00078895n	MCS	0,0000	✗	y	✗	final	bn:00078895n	MCS	0,0000
ajudou	bn:00082481v	MCS	0,0000	✓	y	✓	ajudou	bn:00082481v	MCS	0,0000
garantir	bn:00082877v	MCS	0,0000	✓	y	✓	garantir	bn:00082877v	MCS	0,0000
trio	bn:00078313n	MCS	0,0000	✗	y	✗	trio	bn:00078313n	MCS	0,0000
sul-americano	bn:00110954a	BFY	1,0000	✓	y	✓	sul-americano	bn:00110954a	BFY	1,0000
ponta	bn:00077365n	MCS	0,0000	✗	y	✗	ponta	bn:00077365n	MCS	0,0000
entrar	bn:00087677v	BFY	0,9524	✗	y	✗	entrar	bn:00087677v	BFY	0,9412
pista	bn:00065809n	MCS	0,0000	✓	y	✓	pista	bn:00065809n	BFY	0,5609
finlandês	bn:00034612n	BFY	0,9167	✗	y	✗	finlandês	bn:00034612n	MCS	0,0000
Raikkonen	bn:01660723n	BFY	1,0000	✓	y	✓	Raikkonen	bn:01660723n	BFY	1,0000
ficou	bn:00092756v	MCS	0,0000	✗	y	✗	ficou	bn:00092756v	MCS	0,0000
oitavo	bn:00096089a	BFY	1,0000	✓	y	✓	oitavo	bn:00096089a	BFY	1,0000
tempo	bn:00077270n	MCS	0,0000	✗	y	✗	tempo	bn:00077270n	MCS	0,0000
Michael	bn:01339659n	BFY	0,5682	✗	n	✓	Michael	bn:01311681n	BFY	0,5563
Michael Schumacher	bn:01311681n	BFY	1,0000	✓	y	✓	Michael Schumacher	bn:01311681n	BFY	1,0000
Schumacher	bn:01576979n	BFY	1,0000	✗	n	✓	Schumacher	bn:01311681n	BFY	0,6878
foi	bn:00083181v	BFY	0,8656	✓	y	✓	foi	bn:00083181v	BFY	0,8320
apenas	bn:00114626r	BFY	0,7353	✓	y	✓	apenas	bn:00114626r	BFY	0,7692
previsão	bn:00035806n	MCS	0,0000	✗	y	✗	previsão	bn:00035806n	MCS	0,0000
chuva	bn:00066032n	BFY	0,7428	✓	y	✓	chuva	bn:00066032n	MCS	0,0000
corrida	bn:00035695n	MCS	0,0000	✗	y	✗	corrida	bn:00035695n	MCS	0,0000
domingo	bn:00028190n	BFY	0,8591	✓	y	✓	domingo	bn:00028190n	MCS	0,0000

23 *synsets*, 18 foram corretamente identificados. Assim, verificou-se que, nesse caso, o método EXACT_MATCHING obteve resultados melhores, com um número um pouco maior de *synsets* identificados e menor taxa de erro.

Como nos testes realizados não houve grande discrepância entre os dois métodos configurados pelo parâmetro *Matching Type*, optou-se por valorizar a maior identificação de *synsets*. Assim, utilizou-se o valor EXACT_MATCHING para textos em inglês e PARTIAL_MATCHING para textos em português.

C.2 Resultados complementares do Capítulo 5 - Representações baseadas em *embeddings*

Nessa seção são apresentados resultados complementares da avaliação experimental apresentada na Seção 5.6 do Capítulo 5 (Página 131). Na Figura 47 são apresentadas as distribuições dos valores de $F1^M$ obtidos para cada representação avaliada. Os *box plots* foram gerados considerando os 104 classificadores que foram gerados para cada *dataset*, de acordo com a configuração experimental adotada.

A Tabela 12 apresenta os maiores valores de $F1^M$ obtidos por cada algoritmo de aprendizado entre todos os parâmetros testados. A penúltima linha apresenta o *ranking* médio das

Quadro 9 – Comparação da desambiguação para o documento em inglês

EXACT_MATCHING						PARTIAL_MATCHING				
frag	id	src	score	✓	=	✓	frag	id	src	score
disappointing	bn:00086883v	BFY	1,0000	✓	y	✓	disappointing	bn:00086883v	BFY	1,0000
Lots	bn:00008985n	MCS	0,0000	✓	n	✗	Lots	bn:00052061n	MCS	0,0000
services	bn:00070651n	BFY	0,7107	✓	y	✓	services	bn:00070651n	MCS	0,0000
promised	bn:00082668v	BFY	0,7143	✗	n	✓	promised	bn:00092065v	MCS	0,0000
provided	bn:00084692v	BFY	0,8400	✓	y	✓	provided	bn:00084692v	BFY	0,9231
shuttle	bn:01393975n	BFY	0,8000	✗	n	✗	shuttle	bn:00010609n	MCS	0,0000
running	bn:00087107v	BFY	0,6250	✗	n	✗	running	bn:00093170v	MCS	0,0000
restaurant	bn:00029545n	BFY	1,0000	✓	y	✓	restaurant	bn:00029545n	MCS	0,0000
closed	bn:00085166v	BFY	1,0000	✓	n	✗	closed	bn:00085170v	BFY	0,6618
closed down	bn:00085164v	BFY	1,0000	✓	y	✓	closed down	bn:00085164v	BFY	1,0000
explained	bn:00087870v	BFY	1,0000	✓	y	✓	explained	bn:00087870v	MCS	0,0000
business	bn:00014137n	BFY	0,8320	✓	n	✗	business	bn:00014136n	MCS	0,0000
low	bn:00106209a	BFY	0,6667	✗	n	✓	low	bn:00106205a	MCS	0,0000
right	bn:00114335r	BFY	0,6667	✗	n	✓	right	bn:00116860r	MCS	0,0000
right now	bn:13763778r	MCS	0,0000	✓	y	✓	right now	bn:13763778r	MCS	0,0000
holidays	bn:00044415n	BFY	1,0000	✓	n	✓	holidays	bn:00044416n	MCS	0,0000
guess	bn:00089105v	MCS	0,0000	✗	n	✗	guess	bn:00089106v	MCS	0,0000
make sure	bn:00090585v	BFY	1,0000	✓	y	✓	make sure	bn:00090585v	BFY	1,0000
sure	bn:00099479a	BFY	0,5714	✗	n	✗	sure	bn:00111505a	MCS	0,0000
's	bn:13666727v	MCS	0,0000	✗	y	✗	's	bn:13666727v	MCS	0,0000
busy	bn:00099057a	BFY	0,6154	✓	y	✓	busy	bn:00099057a	MCS	0,0000
business	bn:00014137n	BFY	0,8320	✗	n	✗	business	bn:00014136n	MCS	0,0000
center	bn:00017120n	MCS	0,0000	✗	y	✗	center	bn:00017120n	MCS	0,0000
apart	bn:00114309r	BFY	1,0000	✗	n	✗	apart	bn:00114308r	MCS	0,0000
front desk	bn:00066513n	BFY	1,0000	✓	y	✓	front desk	bn:00066513n	BFY	1,0000
desk	bn:00026554n	BFY	1,0000	✗	y	✗	desk	bn:00026554n	MCS	0,0000
staff	bn:00032685n	BFY	0,6923	✗	n	✓	staff	bn:00073801n	MCS	0,0000
know	bn:00085261v	MCS	0,0000	✓	y	✓	know	bn:00085261v	MCS	0,0000
idea	bn:00045800n	BFY	0,8333	✓	y	✓	idea	bn:00045800n	MCS	0,0000
apart	bn:00114309r	BFY	1,0000	✗	n	✗	apart	bn:00114308r	MCS	0,0000
quite	bn:00116771r	MCS	0,0000	✓	y	✓	quite	bn:00116771r	MCS	0,0000
shocked	bn:00083054v	MCS	0,0000	✓	n	✗	shocked	bn:00093642v	MCS	0,0000
find	bn:00086708v	BFY	0,5714	✓	n	✗	find	bn:00084231v	MCS	0,0000
find out	bn:00082811v	MCS	0,0000	✓	n					
tried	bn:00082844v	BFY	0,6923	✓	n	✗	tried	bn:00087731v	MCS	0,0000
let	bn:00082536v	MCS	0,0000	✓	n	✗	let	bn:00090268v	BFY	0,6154
Vending	bn:00089254v	BFY	1,0000	✓	y	✓	Vending	bn:00089254v	BFY	1,0000
Vending machines	bn:15034978n	BFY	0,9722	✓	y	✓	Vending machines	bn:15034978n	BFY	0,6412
machines	bn:00007309n	BFY	0,7716	✗	n	✗	machines	bn:00052556n	MCS	0,0000
lobby	bn:00004493n	BFY	0,7368	✓	n	✗	lobby	bn:00051711n	MCS	0,0000
Staff	bn:00032685n	BFY	0,6923	✗	n	✓	Staff	bn:00073801n	MCS	0,0000
friendly	bn:00103345a	BFY	1,0000	✗	y	✗	friendly	bn:00103345a	BFY	1,0000
bit	bn:00010731n	BFY	0,5926	✗	n	✗	bit	bn:00010729n	MCS	0,0000
embarrassed	bn:00096119a	MCS	0,0000	✓	y	✓	embarrassed	bn:00096119a	MCS	0,0000
available	bn:00097846a	BFY	0,9492	✓	y	✓	available	bn:00097846a	BFY	0,5414
room	bn:00068233n	BFY	0,9429	✓	y	✓	room	bn:00068233n	MCS	0,0000
fine	bn:00102972a	BFY	1,0000	✗	y	✗	fine	bn:00102972a	BFY	1,0000

Figura 47 – *Box plots* para $F1^M$ por representação utilizada, para cada *dataset*.

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Nota – Para cada *dataset*, os gráficos das representações são apresentados na seguinte ordem: (1) BOW; (2) LDA; (3) Word2Vec; (4) Doc2Vec; (5) NASARI+Babel2Vec; (6) Babel2Vec.

representações segundo o teste de Friedman, considerando as 728 configurações experimentais testadas (104 variações de algoritmos e parâmetros e 7 *datasets*). A última linha apresenta a posição das representações de acordo com esse *ranking*.

Tabela 12 – Maiores valores de $F1^M$ para as coleções *BBC*, *SE-ABSA15* e *BS-Top4*

	<i>NASARI+Babel2Vec</i>	<i>Babel2Vec</i>	BOW	LDA	Word2Vec	Doc2Vec
BBC	0,972876	0,976502	0,969916	0,970883	0,980586	0,970799
IMBHN ^C	0,956163	0,961546	0,957812	0,948156	0,970446	0,922173
IMBHN ^R	0,940258	0,956914	0,969916	0,970883	0,957409	0,937259
J48	0,865307	0,858082	0,861564	0,824172	0,855074	0,526506
k-NN	0,958878	0,966370	0,955323	0,957884	0,969669	0,915351
NB	0,934477	0,952239	0,929833	0,900808	0,951755	0,958609
SVM	0,972876	0,976502	0,964797	0,967654	0,980586	0,970799
SE-product	0,952991	0,955735	0,943081	0,938826	0,950558	0,939978
IMBHN ^C	0,873372	0,955735	0,943081	0,927439	0,939071	0,868570
IMBHN ^R	0,730892	0,765060	0,938537	0,925122	0,732921	0,698659
J48	0,747232	0,729347	0,810296	0,875319	0,764517	0,609952
k-NN	0,938724	0,945087	0,939842	0,938826	0,934453	0,769169
NB	0,884791	0,932946	0,855050	0,715410	0,873881	0,912486
SVM	0,952991	0,954739	0,883670	0,935326	0,950558	0,939978
SE-polarity	0,597225	0,594260	0,558758	0,539338	0,621238	0,556340
IMBHN ^C	0,539077	0,594260	0,558758	0,497508	0,621238	0,511960
IMBHN ^R	0,545238	0,550890	0,557566	0,524487	0,544862	0,506299
J48	0,478647	0,455994	0,474069	0,443998	0,483042	0,429528
k-NN	0,537139	0,525139	0,528043	0,539338	0,541276	0,486016
NB	0,595268	0,554302	0,468746	0,480393	0,545126	0,532572
SVM	0,597225	0,577768	0,558302	0,519539	0,592297	0,556340
SE-product-polarity	0,497043	0,494970	0,443560	0,421607	0,510545	0,461315
IMBHN ^C	0,472152	0,494970	0,443560	0,421607	0,457392	0,374233
IMBHN ^R	0,240281	0,245434	0,328582	0,399322	0,236923	0,246709
J48	0,301117	0,335454	0,358916	0,355277	0,312568	0,259420
k-NN	0,438379	0,436045	0,411538	0,419932	0,436834	0,347557
NB	0,438854	0,434714	0,384997	0,307768	0,390888	0,400885
SVM	0,497043	0,478375	0,425338	0,418636	0,510545	0,461315
BS-topic	0,996825	1,000000	1,000000	1,000000	1,000000	1,000000
IMBHN ^C	0,973680	0,990569	0,986319	0,988991	0,977630	0,969799
IMBHN ^R	0,973811	0,982885	0,994375	0,988482	0,988671	0,978358
J48	0,808339	0,850198	0,964721	0,901001	0,964628	0,664314
k-NN	0,996825	1,000000	0,994791	1,000000	1,000000	0,994042
NB	0,965391	0,965952	0,996131	0,987197	0,976647	0,991451
SVM	0,996825	1,000000	1,000000	0,989315	0,996131	1,000000
BS-semantic	0,658056	0,664682	0,695852	0,618711	0,656611	0,590126
IMBHN ^C	0,580900	0,566616	0,667463	0,591395	0,561850	0,513348
IMBHN ^R	0,570066	0,636406	0,695852	0,612795	0,599575	0,496712
J48	0,491536	0,399512	0,567107	0,489398	0,515755	0,321358
k-NN	0,616822	0,616316	0,660750	0,618711	0,608121	0,541812
NB	0,507814	0,510583	0,566944	0,511153	0,444052	0,525981
SVM	0,658056	0,664682	0,661353	0,603702	0,656611	0,590126
BS-topic-semantic	0,484579	0,501235	0,517545	0,479573	0,485417	0,439909
IMBHN ^C	0,418227	0,454941	0,476316	0,479573	0,464493	0,366523
IMBHN ^R	0,241580	0,236749	0,361117	0,338505	0,243478	0,203202
J48	0,265628	0,251758	0,392214	0,334849	0,326445	0,175700
k-NN	0,427565	0,451989	0,494361	0,462205	0,440707	0,397474
NB	0,449823	0,447991	0,440168	0,348542	0,426031	0,365001
SVM	0,484579	0,501235	0,517545	0,425766	0,485417	0,439909
<i>Ranking médio</i>	3,0790	2,5522	3,8826	3,8475	2,9725	4,6662
<i>Posição pelo ranking médio</i>	3º	1º	5º	4º	2º	6º

Nota – Valores de $F1^M$ maiores que o $F1^M$ obtido com o *baseline* BOW (coluna cinza) são apresentadas em negrito. Os maiores valores de cada linha estão sublinhados. As linhas em cinza correspondem aos melhores valores de $F1^M$ de cada representação para o respectivo *dataset*.

Considerando a combinação de classificadores gerados com a BOW e classificadores gerados com *NASARI+Babel2Vec*, as distribuições dos valores de $F1^M$ obtidos para cada estratégia de combinação são apresentadas nas Figuras 48, 49 e 50. Já os resultados da combinação de BOW e *Babel2Vec* são apresentados nas Figuras 51, 52 e 53. Nessas figuras, para cada *dataset* são apresentados 5 gráficos, correspondentes à variação de peso do voto do classificador gerado com a BOW (pesos de 0,1 a 0,9). Na Tabela 13 são apresentados os maiores valores de $F1^M$ obtidos pelas combinações de BOW com as duas representações semânticas propostas.

Figura 48 – Box plots para $F1^M$ de cada *dataset* para a combinação de BOW e NASARI+Babel2Vec, utilizando a estratégia SC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 49 – *Box plots* para $F1^M$ de cada *dataset* para a combinação de BOW e *NASARI+Babel2Vec*, utilizando a estratégia WSC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 50 – Box plots para $F1^M$ de cada *dataset* para a combinação de BOW e NASARI+Babel2Vec, utilizando a estratégia MC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 51 – *Box plots* para $F1^M$ de cada *dataset* para a combinação de BOW e *Babel2Vec*, utilizando a estratégia SC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 52 – Box plots para $F1^M$ de cada dataset para a combinação de BOW e Babel2Vec, utilizando a estratégia WSC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Figura 53 – *Box plots* para $F1^M$ de cada *dataset* para a combinação de BOW e *Babel2Vec*, utilizando a estratégia MC

(a) *BBC*

(b) *SE-product*

(c) *BS-topic*

(d) *SE-polarity*

(e) *SE-product-polarity*

(f) *BS-semantic*

(g) *BS-topic-semantic*

Fonte: Elaborada pelo autor.

Tabela 13 – Melhores $F1^M$ para as coleções *BBC*, *SE-ABSA15* e *BS-Top4* para combinação de classificadores

	BOW + <i>NASARI</i> + <i>Babel2Vec</i>			BOW + <i>Babel2Vec</i>		
	SC	WSC	MC	SC	WSC	MC
BBC	0,975346	0,975346	0,973141	0,979009	0,979009	0,977737
IMBHN ^C	0,974232	0,975124	0,970395	0,976068	0,976068	0,974134
IMBHN ^R	0,971788	0,972241	0,969916	0,972177	0,972177	0,969916
J48	0,872928	0,886845	0,875534	0,871361	0,880765	0,868419
k-NN	0,968133	0,968070	0,964530	0,968849	0,968849	0,966994
NB	0,946635	0,935203	0,950028	0,952241	0,952241	0,952411
SVM	0,975346	0,975346	0,973141	0,979009	0,979009	0,977737
SE-product	0,957468	0,957468	0,955873	0,957420	0,957420	0,955735
IMBHN ^C	0,957468	0,957468	0,948821	0,957420	0,957420	0,955735
IMBHN ^R	0,938537	0,938537	0,938537	0,938537	0,938537	0,938537
J48	0,810296	0,810296	0,810296	0,810296	0,810296	0,810296
k-NN	0,957307	0,953092	0,955873	0,956386	0,952833	0,952053
NB	0,888674	0,904095	0,915964	0,932946	0,942296	0,932946
SVM	0,955700	0,955700	0,955700	0,954739	0,954739	0,954739
SE-polarity	0,598141	0,598141	0,598141	0,578962	0,578962	0,598172
IMBHN ^C	0,582828	0,583770	0,593938	0,574957	0,573240	0,598172
IMBHN ^R	0,562958	0,563097	0,561971	0,565541	0,565541	0,557566
J48	0,500451	0,495411	0,490456	0,473478	0,468869	0,474069
k-NN	0,538494	0,539303	0,549585	0,537894	0,541553	0,561767
NB	0,592525	0,593965	0,595291	0,562946	0,562946	0,554302
SVM	0,598141	0,598141	0,598141	0,578962	0,578962	0,578962
SE-product-polarity	0,494619	0,494619	0,497043	0,494577	0,494577	0,496848
IMBHN ^C	0,494619	0,494619	0,484575	0,494577	0,494577	0,496848
IMBHN ^R	0,316158	0,323011	0,328582	0,318840	0,318616	0,328582
J48	0,358946	0,358387	0,358916	0,366844	0,361880	0,364201
k-NN	0,448241	0,447270	0,445586	0,450975	0,449944	0,453357
NB	0,440556	0,441098	0,439392	0,435840	0,435461	0,435907
SVM	0,493585	0,493585	0,497043	0,482992	0,482992	0,482583
BS-topic	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
IMBHN ^C	0,992083	0,992083	0,990208	0,995833	0,995833	0,990569
IMBHN ^R	0,994375	0,994375	0,994375	1,000000	1,000000	1,000000
J48	0,964721	0,964721	0,964721	0,964721	0,964721	0,964721
k-NN	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
NB	0,996825	0,996131	0,996131	0,996131	0,996131	0,996131
SVM	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
BS-semantic	0,691307	0,691307	0,695852	0,706537	0,709230	0,695852
IMBHN ^C	0,672554	0,675449	0,667463	0,695745	0,695745	0,668093
IMBHN ^R	0,691307	0,691307	0,695852	0,706537	0,709230	0,695852
J48	0,580822	0,580822	0,567107	0,570609	0,570609	0,567107
k-NN	0,670611	0,672388	0,663116	0,671643	0,676618	0,660750
NB	0,568456	0,566944	0,586614	0,576475	0,570551	0,594496
SVM	0,675172	0,676828	0,664928	0,686757	0,686757	0,689633
BS-topic-semantic	0,526006	0,526006	0,517545	0,517545	0,517545	0,517545
IMBHN ^C	0,509426	0,509426	0,490452	0,507559	0,507559	0,476316
IMBHN ^R	0,357371	0,357371	0,361117	0,338750	0,352458	0,361117
J48	0,393212	0,393212	0,392214	0,392214	0,392214	0,392214
k-NN	0,504935	0,504935	0,494361	0,495899	0,497842	0,494361
NB	0,472305	0,451946	0,478114	0,456686	0,450068	0,450184
SVM	0,526006	0,526006	0,517545	0,517545	0,517545	0,517545

Nota – Valores em negrito correspondem a $F1^M$ maiores do que $F1^M$ dos respectivos classificadores base (resultados apresentados na Tabela 12). O melhor valor de $F1^M$ de cada linha está sublinhado. As linhas em cinza correspondem aos melhores valores de $F1^M$ para o respectivo *dataset*.