UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Mineração de opiniões baseada em aspectos para revisões de produtos e serviços

Ivone Penque Matsuno Yugoshi

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: _____

Ivone Penque Matsuno Yugoshi

Mineração de opiniões baseada em aspectos para revisões de produtos e serviços

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

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

Y94m Yugoshi, Ivone Penque Matsuno
 Mineração de opiniões baseada em aspectos para
 revisões de produtos e serviços / Ivone Penque
 Matsuno Yugoshi; orientadora 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. Mineração de Opiniões. 2. Análise de
 Sentimentos. 3. Extração de Aspectos. 4. Aprendizado
 Semissupervisionado. 5. Aprendizado por
 Transferência entre Domínios. I. Rezende, Solange
 Oliveira, orient. II. Título.

Ivone Penque Matsuno Yugoshi

Aspect-based opinion mining for reviews of products and services

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**

À minha família.

AGRADECIMENTOS

A Deus pelo dom da vida e por colocar pessoas tão especiais nessa minha jornada que contribuíram tanto em aspectos da minha vida profissional quanto pessoal. Cada uma delas foi importante para o meu aprendizado e crescimento. A todos, minha gratidão e respeito.

Ao meu pai Yoshio e minha mãe Ordylette pela minha vida, pelo amor incondicional, pela educação e pelo incentivo e apoio constante para realizar meus sonhos.

Ao meu marido Juliano pelo amor, cumplicidade e incentivo em todos os momentos.

À minha filha Natália pelo amor e carinho.

À minha orientadora Profa. Dra. Solange Oliveira Rezende pela orientação nesse doutorado, pela sua capacidade de inovação e realização e, principalmente, pelo seu incentivo que me motivou a superar diversos desafios profissionais e pessoais.

Ao Prof. Dr. Ricardo Marcondes Marcacini pela coorientação, mesmo que não oficial, pois foi enriquecedora.

À Roberta e à Camila por toda a troca de conhecimentos técnicos e pela grande amizade que surgiu. Isso foi essencial para tornar esses anos mais leves e alegres.

Às minhas amigas de longa data, Ana Cristina e Jane que, mesmo distantes fisicamente, puderam estar presentes nesses anos de doutorado compartilhando diversas experiências.

Aos colegas de Laboratório de Inteligência Computacional (LABIC) pela alegre convivência. Aos alunos de iniciação científica Guilherme, Leonardo e Darlan que contribuíram de algum modo no tema dessa tese. Ao Rafael pelas discussões enriquecedoras em aprendizado transdutivo.

Ao Anandsing pelas revisões do inglês e pelas palavras oportunas em momentos importantes.

Aos professores e funcionários da UFMS, em especial, Rodrigo Kishi, Ronaldo Fiorilo e Ricardo Marcacini que estão trabalhando comigo desde a primeira turma do curso.

À UFMS pelo afastamento concedido para realização do doutorado.

A todos funcionários do ICMC-USP que contribuíram direta ou indiretamente para a realização do doutorado.

A toda minha família, sempre presente em minha vida. Às minhas irmãs e meus cunhados, Elisa e Dácio, Isabella e Renato, pela compreensão da minha ausência e pelo carinho e constante incentivo, mesmo que convivendo em cidades diferentes. Aos meus sobrinhos Beatriz, Dante e Mateus que trouxeram mais amor, alegria e graça a nossa família. Aos meus sogros Koiti e Tioka, aos meus cunhados Luciana, Marcelo, Eliane e Fábio, e a minha sobrinha Marina que está por vir.

*“Estabeleça suas intenções.
Ame sua jornada.”
(autor desconhecido)*

RESUMO

YUGOSHI, I. P. M. **Mineração de opiniões baseada em aspectos para revisões de produtos e serviços**. 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.

A Mineração de Opiniões é um processo que tem por objetivo extrair as opiniões e suas polaridades de sentimentos expressas em textos em língua natural. Essa área de pesquisa tem ganhado destaque devido ao volume de opiniões que os usuários compartilham na Internet, como revisões em *sites de e-commerce*, rede sociais e *tweets*. A Mineração de Opiniões baseada em Aspectos é uma alternativa promissora para analisar a polaridade do sentimento em um maior nível de detalhes. Os métodos tradicionais para extração de aspectos e classificação de sentimentos exigem a participação de especialistas de domínio para criar léxicos ou definir regras de extração para diferentes idiomas e domínios. Além disso, tais métodos usualmente exploram algoritmos de aprendizado supervisionado, porém exigem um grande conjunto de dados rotulados para induzir um modelo de classificação. Os desafios desta tese de doutorado estão relacionados a como diminuir a necessidade de grande esforço humano tanto para rotular dados, quanto para tratar a dependência de domínio para as tarefas de extração de aspectos e classificação de sentimentos dos aspectos para Mineração de Opiniões. Para reduzir a necessidade de grande quantidade de exemplos rotulados foi proposta uma abordagem semissupervisionada, denominada por *Aspect-based Sentiment Propagation on Heterogeneous Networks* (ASPHN) em que são propostas representações de textos nas quais os atributos linguísticos, os aspectos candidatos e os rótulos de sentimentos são modelados por meio de redes heterogêneas. Para redução dos esforços para construir recursos específicos de domínio foi proposta uma abordagem baseada em aprendizado por transferência entre domínios denominada *Cross-Domain Aspect Label Propagation through Heterogeneous Networks* (CD-ALPHN) que utiliza dados rotulados de outros domínios para suportar tarefas de aprendizado em domínios sem dados rotulados. Nessa abordagem são propostos uma representação em uma rede heterogênea e um método de propagação de rótulos. Os vértices da rede são os aspectos rotulados do domínio de origem, os atributos linguísticos e os candidatos a aspectos do domínio alvo. Além disso, foram analisados métodos de extração de aspectos e propostas algumas variações para considerar cenários não-supervisionados e independentes de domínio. As soluções propostas nesta tese de doutorado foram avaliadas e comparadas as do estado-da-arte utilizando coleções de revisões de diferentes produtos e serviços. Os resultados obtidos nas avaliações experimentais são competitivos e demonstram que as soluções propostas são promissoras.

Palavras-chave: Mineração de Opiniões, Análise de Sentimentos, Extração de Aspectos, Aprendizado Semissupervisionado, Aprendizado por Transferência entre Domínios.

ABSTRACT

YUGOSHI, I. P. M. **Aspect-based opinion mining for reviews of products and services**. 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.

Opinion Mining is a process that aims to extract opinions and their sentiment polarities expressed in natural language texts. This area of research has been in the highlight because of the volume of opinions that users share on the available visualization means on the Internet (reviews on e-commerce sites, social networks, tweets, others). Aspect-based Opinion Mining is a promising alternative for analyzing the sentiment polarity on a high level of detail. The traditional methods for aspect extraction and sentiment classification require the participation of domain experts to create lexicons or define extraction rules for different languages and domains. In addition, such methods usually exploit supervised machine learning algorithms, but require a large set of labeled data to induce a classification model. The challenges of this doctoral thesis are related on to how to reduce the need for great human effort both: (i) to label data; and (ii) to treat domain dependency for the tasks of aspect extraction and aspect sentiment classification for Opinion Mining. In order to reduce the need for a large number of labeled examples, a semi-supervised approach was proposed, called Aspect-based Sentiment Propagation on Heterogeneous Networks (ASPHN). In this approach, text representations are proposed in which linguistic attributes, candidate aspects and sentiment labels are modeled by heterogeneous networks. Also, a cross-domain learning approach called Cross-Domain Aspect Label Propagation through Heterogeneous Networks (CD-ALPHN) is proposed in order to reduce efforts to build domain-specific resources. This approach uses labeled data from other domains to support learning tasks in domains without labeled data. A representation in a heterogeneous network and a label propagation method are proposed in this cross-domain learning approach. The vertices of the network are the labeled aspects of the source domain, the linguistic attributes, and the candidate aspects of the target domain. In addition, aspect extraction methods were analyzed and some variations were proposed to consider unsupervised and domain independent scenarios. The solutions proposed in this doctoral thesis were evaluated and compared to the state-of-the-art solutions using collections of different product and service reviews. The results obtained in the experimental evaluations are competitive and demonstrate that the proposed solutions are promising.

Keywords: Opinion Mining, Sentiment Analysis, Aspect Extraction, Semi-supervised Learning, Cross-Domain Transfer Learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Processo de mineração de opiniões	39
Figura 2 – Exemplo de mensagem de opinião	40
Figura 3 – Exemplo das quintuplas de opinião	41
Figura 4 – Sumarização das opiniões	41
Figura 5 – Exemplo de representação da coleção de documentos por uma rede heterogê- nea bipartida	49
Figura 6 – Exemplo de como a utilização de exemplos não rotulados auxilia na classificação	56
Figura 7 – Esquema conceitual da representação proposta	69
Figura 8 – Exemplo de árvore de dependência sintática gerada para a sentença “ <i>The food was nothing much, but I loved the staff.</i> ”	71
Figura 9 – Exemplo de rede heterogênea gerada a partir da representação no modelo espaço-vetorial apresentada na Tabela 2.	72
Figura 10 – Exemplo de rede heterogênea bipartida extraída da representação no modelo espaço-vetorial apresentado na Tabela 3.	72
Figura 11 – Exemplo de representação de rede bipartida para aprendizado semissupervi- sionado.	73
Figura 12 – Exemplo de aprendizado semissupervisionado baseado em redes bipartidas considerando o algoritmo TCBHN com $\eta = 0.5$ para extração de aspectos.	74
Figura 13 – Exemplo de aprendizado semissupervisionado baseado em redes biparti- das considerando o algoritmo TCBHN com $\eta = 0.5$ para classificação de sentimentos dos aspectos.	75
Figura 14 – Resultados para extração de aspectos.	82
Figura 15 – Diagramas de diferença crítica para extração de aspectos.	83
Figura 16 – Diagramas de diferença crítica para classificação de sentimentos dos aspectos.	85
Figura 17 – Resultados para classificação de sentimentos de aspectos.	86
Figura 18 – Modelo gráfico do LDA	95
Figura 19 – <i>Framework</i> para avaliação de métodos de extração de aspectos para mineração de opiniões.	97
Figura 20 – Visão geral do processo de anotação de um documento revisão.	98
Figura 21 – Diagrama de diferença crítica para medida F^1 considerando correspondência exata	105
Figura 22 – Diagrama de diferença crítica para medida F^1 considerando correspondência aproximada	105

Figura 23 – Revisões com marcação Part-of Speech (POS)	106
Figura 24 – Part-of-Speech (PoS) e relações de dependência sintática extraídas da sentença “ <i>The camera is nice</i> ”.	110
Figura 25 – Esquema geral da representação da rede heterogênea proposta para unificar a representação dos conjuntos de atributos entre domínios fonte (vértices com aspectos rotulados) e alvo (vértices dos candidatos a aspecto), em que os atributos linguísticos são usados como vértices ponte.	114
Figura 26 – Visão geral da interface do sistema Websensors-SentimentAnalysis (Data Analytics) para mineração de opiniões baseada em aspectos.	122
Figura 27 – Exemplo de um mapa de calor que representa a ocorrência geográfica de um aspecto selecionado de acordo com a localização dos consumidores.	123
Figura 28 – Interface para criação do processo de anotação - configuração.	127
Figura 29 – Interface para criação do processo de anotação - anotadores.	127
Figura 30 – Interface para sumarização do processo de anotação.	128
Figura 31 – Interface de anotação com sugestões de rótulos.	129
Figura 32 – Formato de arquivo de exportação das anotações da revisão: “ <i>The price is reasonable although the service is poor</i> ”.	130
Figura 33 – Arquitetura geral do <i>framework</i> Opinon-meter	133
Figura 34 – Dependências sintáticas do <i>Aspectator</i>	139
Figura 35 – Árvore sintática da sentença <i>s</i>	140
Figura 36 – Arquivo de saída do <i>Aspectator</i>	142
Figura 37 – Resultados para medida de precisão.	142
Figura 38 – Resultados para medida de cobertura.	143
Figura 39 – Resultados para medida de F^1	143

LISTA DE ALGORITMOS

Algoritmo 1 – Gaussian Fields and Harmonic Functions (GFHF)	58
Algoritmo 2 – Learning with Local and Global Consistency (LLGC)	59
Algoritmo 3 – Label Propagation through Heterogeneous Networks (LPHN)	60
Algoritmo 4 – GNetMine (GNM)	61
Algoritmo 5 – Tag-based Model (TM)	63
Algoritmo 6 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 1)	64
Algoritmo 7 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 2)	65
Algoritmo 8 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 3)	66
Algoritmo 9 – Cross-Domain Aspect Label Propagation through Heterogeneous Network (CD-ALPHN)	115

LISTA DE TABELAS

Tabela 1 – Ilustração de uma matriz atributo-valor para representar um coleção com n exemplos (documentos) e m atributos (termos).	48
Tabela 2 – Representação espaço-vetorial gerada a partir da árvore sintática apresentada na Figura 8.	71
Tabela 3 – Representação no modelo espaço-vetorial para aspectos extraídos no passo anterior das três sentenças.	72
Tabela 4 – Matriz de contingência para os termos t_i e t_j	92
Tabela 5 – Descrição das coleções de revisões utilizadas no experimento	101
Tabela 6 – Resultados da medida F^1 para cada método em cada coleção de revisões considerando correspondência exata	104
Tabela 7 – Resultados da medida F^1 para cada método em cada coleção de revisões considerando correspondência aproximada	105
Tabela 8 – Exemplo de aspectos por método de extração	107
Tabela 9 – Descrição das coleções de revisões utilizadas nos experimentos	116
Tabela 10 – Level of inconsistency (β) considering the number of aspects of the target domain that occur in the source domain.	118
Tabela 11 – Resultados de F^1 para a extração de aspectos de cada abordagem.	119
Tabela 12 – <i>Tags</i> utilizadas para selecionar candidatos a aspecto	134
Tabela 13 – Dados das bases utilizadas	136
Tabela 14 – Precisão na extração de aspectos por base rotulada	136
Tabela 15 – Precisão por classe na classificação de sentimento para cada método e base rotulada	137
Tabela 16 – Matriz atributo-valor da sentença da Figura 35	140

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Contextualização e Motivação	25
1.2	Questões de Pesquisa e Objetivos	29
1.3	Principais Resultados	31
1.4	Organização da Tese	33
2	FUNDAMENTOS	35
2.1	Considerações Iniciais	35
2.2	Mineração de Opiniões baseada em Aspectos	35
2.2.1	<i>Trabalhos Relacionados com Extração de Aspectos</i>	41
2.2.1.1	<i>Extração baseada em frequência</i>	42
2.2.1.2	<i>Extração baseada relações</i>	42
2.2.1.3	<i>Extração usando aprendizado supervisionado</i>	43
2.2.1.4	<i>Extração baseada em Modelos de Tópicos</i>	44
2.2.2	<i>Trabalhos Relacionados com Classificação de Sentimentos dos As-</i>	
	<i>pectos</i>	45
2.2.2.1	<i>Classificadores baseados em Léxico</i>	45
2.2.2.2	<i>Classificadores baseados em Aprendizado de Máquina</i>	46
2.3	Representação de Dados Textuais	47
2.3.1	<i>Modelo Espaço-Vetorial</i>	47
2.3.2	<i>Modelo de Redes</i>	48
2.4	Aprendizado de Máquina	49
2.4.1	<i>Aprendizado Supervisionado</i>	51
2.4.1.1	<i>Naïve Bayes (NB)</i>	51
2.4.1.2	<i>k-NN (k-Nearest Neighbor)</i>	51
2.4.1.3	<i>Árvore de Decisão</i>	51
2.4.1.4	<i>SVM (Support Vector Machines)</i>	52
2.4.1.5	<i>Redes MLP (Multi-Layer Perceptron)</i>	52
2.4.2	<i>Aprendizado Semissupervisionado</i>	53
2.4.3	<i>Aprendizado Semissupervisionado Transdutivo em Redes</i>	55
2.4.3.1	<i>Gaussian Fields and Harmonic Functions (GFHF)</i>	58
2.4.3.2	<i>Learning with Local and Global Consistency (LLGC)</i>	59
2.4.3.3	<i>Label Propagation using Bipartite Heterogeneous Networks</i>	59

2.4.3.4	<i>GNetMine</i>	60
2.4.3.5	<i>Tag-based Model</i>	62
2.4.3.6	<i>Transductive Categorization based on Heterogeneous Networks</i>	63
2.5	Considerações Finais	66
3	MINERAÇÃO DE OPINIÕES BASEADA EM ASPECTOS POR MEIO DE APRENDIZADO SEMISSUPERVISIONADO	67
3.1	Considerações Iniciais	67
3.2	Visão Geral da Solução Proposta	69
3.3	<i>Aspect-Based Sentiment Propagation for Heterogeneous Networks (ASPHN)</i>	70
3.3.1	<i>Geração de Redes Heterogêneas Bipartidas</i>	70
3.3.2	<i>Aprendizado Semissupervisionado baseado em Redes Heterogêneas Bipartidas</i>	72
3.4	Avaliação Experimental	77
3.4.1	<i>Coleções de Revisões e Representação Estruturada</i>	77
3.4.2	<i>Configuração Experimental</i>	78
3.4.3	<i>Critérios de Avaliação</i>	80
3.4.4	<i>Análise e Discussão dos Resultados</i>	81
3.4.4.1	<i>Extração de Aspectos</i>	82
3.4.4.2	<i>Classificação de Sentimentos de Aspectos</i>	83
3.5	Considerações Finais	86
4	EXTRAÇÃO DE ASPECTOS NÃO SUPERVISIONADA E INDEPENDENTE DE DOMÍNIO PARA MINERAÇÃO DE OPINIÕES	89
4.1	Considerações Iniciais	89
4.2	Métodos para Extração de Aspectos Não-supervisionados e Independentes de Domínio	90
4.2.1	<i>Métodos Baseados em Frequência</i>	90
4.2.2	<i>Métodos baseados em Combinações</i>	91
4.2.3	<i>Métodos baseados em Medidas de Associação</i>	92
4.2.3.1	<i>Odds Ratio (OR)</i>	93
4.2.3.2	<i>Pointwise Mutual Information (PMI)</i>	93
4.2.3.3	<i>Chi-square (χ^2)</i>	94
4.2.4	<i>Métodos baseados em Modelos de Tópicos</i>	94
4.3	<i>Framework para avaliação de métodos de extração de aspectos para Mineração de Opiniões</i>	96
4.3.1	<i>Coleções de Revisões</i>	96
4.3.2	<i>Extração de Aspectos</i>	97
4.3.3	<i>Critérios de Avaliação</i>	99

4.3.4	<i>Avaliação Comparativa</i>	100
4.4	<i>Avaliação Experimental</i>	100
4.4.1	<i>Configuração Experimental</i>	101
4.4.2	<i>Análise e Discussão dos Resultados</i>	103
4.5	<i>Considerações Finais</i>	107
5	EXTRAÇÃO DE ASPECTOS PARA MINERAÇÃO DE OPINIÕES USANDO APRENDIZADO POR TRANSFERÊNCIA ENTRE DO- MÍNIOS	109
5.1	<i>Considerações Iniciais</i>	109
5.2	<i>Transferência de Aprendizado entre Domínios</i>	111
5.3	<i>Aprendizado Transdutivo para Extração de Aspectos entre Domínios</i>	112
5.3.1	<i>Representação de Dados entre Domínios para Aprendizado em Re- des Heterogêneas</i>	113
5.3.2	<i>Propagação de Rótulos de Aspectos por meio de Atributos Linguís- ticos</i>	113
5.4	<i>Avaliação Experimental</i>	116
5.4.1	<i>Coleções de Revisões</i>	116
5.4.2	<i>Configuração Experimental</i>	116
5.4.3	<i>CrITÉRIOS de Avaliação</i>	117
5.4.4	<i>Análise e Discussão dos Resultados</i>	119
5.5	<i>Aplicações</i>	120
5.5.1	<i>Websensors</i>	120
5.5.2	<i>Websensors-SentimentAnalysis</i>	121
5.6	<i>Considerações Finais</i>	122
6	PROJETOS PILOTOS	125
6.1	<i>Considerações Iniciais</i>	125
6.2	Sistema para Anotação de Documentos e Sugestão de Rótulos para Mineração de Opiniões baseada em Aspectos	125
6.2.1	<i>RotuLabic-AspectSentiment</i>	126
6.2.1.1	<i>RotuLabic-AS – Administração</i>	126
6.2.1.2	<i>RotuLabic-AS – Sugestão de Rótulos</i>	129
6.2.1.3	<i>RotuLabic-AS – Anotação</i>	130
6.2.2	<i>Anotação de Coleções de Revisões</i>	130
6.2.3	<i>Considerações Finais – Rotulabic-Aspect Sentiment</i>	131
6.3	<i>Framework para Mineração de Opiniões baseada em Aspectos</i> . . .	131
6.3.1	<i>Opinion-meter</i>	133
6.3.1.1	<i>Extração de Aspectos</i>	133
6.3.1.2	<i>Classificação de Sentimentos dos Aspectos</i>	134

6.3.1.3	<i>Baseado em Léxico</i>	134
6.3.1.4	<i>Baseado em PMI</i>	134
6.3.1.5	<i>Baseado em Classificação por Aprendizado de Máquina</i>	135
6.3.2	<i>Avaliação Experimental</i>	135
6.3.2.1	<i>Configuração Experimental</i>	135
6.3.2.2	<i>Análise e Discussão dos Resultados</i>	136
6.3.3	<i>Considerações Finais – Opinion-meter</i>	137
6.4	Estudo Comparativo de métodos para extração de aspectos baseados em Relações Sintáticas e Aprendizado de Máquina para Mineração de Opiniões	137
6.4.1	<i>Extração de Aspectos baseada em Relações Sintáticas</i>	138
6.4.2	<i>Extração de Aspectos Baseada em Aprendizado de Máquina Supervisionado</i>	139
6.4.2.1	<i>Representação de Textos para Classificação de Aspectos</i>	139
6.4.2.2	<i>Classificadores</i>	140
6.4.3	<i>Avaliação Experimental</i>	140
6.4.3.1	<i>Configurações dos Classificadores</i>	141
6.4.3.2	<i>Configuração do Algoritmo Aspectator</i>	141
6.4.3.3	<i>Medidas de Avaliação</i>	141
6.4.3.4	<i>Análise e Discussão dos Resultados</i>	141
6.4.4	<i>Considerações Finais – Métodos baseado em Relações e Aprendizado Máquina</i>	143
7	CONCLUSÕES	145
7.1	Contribuições Científicas	145
7.2	Síntese dos Resultados e Relação de Trabalhos Publicados	150
7.3	Limitações e Trabalhos Futuros	151
REFERÊNCIAS		153
APÊNDICE A	RESULTADOS – EXTRAÇÃO DE ASPECTOS NÃO SUPERVISIONADA E INDEPENDENTE DE DOMÍNIO PARA MINERAÇÃO DE OPINIÕES	169
APÊNDICE B	RESULTADOS – MINERAÇÃO DE OPINIÕES BASEADA EM ASPECTOS USANDO APRENDIZADO SEMISSUPERVISIONADO	187

INTRODUÇÃO

1.1 Contextualização e Motivação

As opiniões de outras pessoas ou de especialistas influenciam o processo de tomada de decisão para comprar produtos, escolher políticos, investir no mercado financeiro e em diversas outras instâncias. Consultar opiniões de outras pessoas é um hábito comum, e atualmente, com o advento da Web 2.0, em que o usuário passou a ser um gerador de conteúdo, essas opiniões são amplamente expressas pelos usuários em *blogs*, redes sociais, *sites* especializados ou de *e-commerce*. De acordo com pesquisas realizadas, a maioria dos usuários lêem comentários sobre produtos antes de comprá-los e utilizam essas informações para tomar decisões (ANNEX, 2016; PANG; LEE, 2008). Por outro lado, as empresas estão buscando compreender melhor os pontos fortes e fracos de seus produtos e serviços para aumentar suas vendas e conquistar clientes. Existem milhares de opiniões disponíveis na web, essas opiniões, em geral, são expressas por meio de textos livres em língua natural. O volume de dados de textos livres é expressivamente grande e tem aumentado rapidamente o que tem despertado o desenvolvimento de técnicas e ferramentas para extrair e compreender tais textos de modo automático (LIU, 2012; LIU; ZHANG, 2012).

A Mineração de Opiniões, também conhecida como Análise de Sentimentos, é um processo que tem por objetivo identificar as opiniões ou sentimentos dos usuários expressos em textos dos mais variados tipos (LIU; ZHANG, 2012). Este processo envolve técnicas de diversas áreas como Mineração de Textos, Aprendizado de Máquina e Processamento de Língua Natural. A Mineração de Opiniões é uma área de pesquisa relativamente nova, mas vem sendo amplamente estudada e apresenta diversos desafios (LIU, 2012). De modo simplificado pode-se definir que dado um texto de opinião (revisão) em língua natural referente a uma entidade (produtos, serviços, organizações, pessoas, problemas, eventos), o objetivo da Mineração de Opiniões é identificar os sentimentos em relação à entidade que são expressos nessas revisões (LIU, 2012).

Abordagens tradicionais para Mineração de Opiniões visam classificar a polaridade do sentimento das revisões como positiva, negativa ou neutra. Esta classificação é realizada analisando a revisão como um todo (Mineração de Opiniões em nível de documento) ou analisando o sentimento de cada sentença do documento (Mineração de Opiniões em nível de sentença). Nesses dois casos, não são exploradas informações sobre determinados aspectos de uma entidade (produto ou serviço) analisada, mesmo sendo frequente a existência de sentimentos diferentes para aspectos distintos nas revisões analisadas. Por exemplo, na sentença “*Eu gostei da imagem desta televisão, mas o controle remoto dela é horrível*”, tem-se uma opinião positiva e outra negativa. Além disso, cada opinião refere-se a aspectos diferentes (“imagem” e “controle remoto”) da entidade em questão (“televisão”). Para lidar com esse tipo de cenário é realizada a Mineração de Opiniões baseada em Aspectos, que potencializa o apoio à tomada de decisão gerando informações mais específicas sobre o sentimento dos aspectos de uma entidade. Porém, a Mineração de Opiniões baseada em Aspectos é um processo mais complexo e desafiador (LIU, 2015; JIMÉNEZ-ZAFRA *et al.*, 2016). Com base nessa necessidade, esta tese de doutorado tem por tema a Mineração de Opiniões baseada em Aspectos que apresenta uma visão mais detalhada das opiniões expressas sobre determinado assunto.

Formalmente, nesta tese, define-se que uma opinião é representada pela tripla $O = (e_i, a_{ij}, s_{ijk})$, em que e_i é uma entidade i (produto ou serviço), a_{ij} é o aspecto j (propriedade) da entidade i , e s_{ijk} é a polaridade do sentimento em relação ao aspecto a_{ij} da entidade e_i (por exemplo, positiva, negativa ou neutra). Considerando essa definição, no texto de opinião “*As sobremesas deste restaurante são deliciosas, mas o atendimento é muito ruim.*” é possível extrair as seguintes triplas de opinião: $O_1 = (\text{restaurante}, \text{sobremesas}, \text{positivo})$ e $O_2 = (\text{restaurante}, \text{atendimento}, \text{negativo})$. Dada uma coleção de textos de revisões, a Mineração de Opiniões baseada em Aspectos tem por objetivo extrair essas triplas de opiniões. Para isso várias tarefas são necessárias, como: extração e categorização das entidades, extração e categorização dos aspectos, classificação de sentimentos dos aspectos e sumarização das opiniões. Devido a amplitude da Mineração de Opiniões baseada em Aspectos, nesta tese, os objetivos estão relacionados às tarefas de: (i) identificar quais termos são aspectos e (ii) classificar os sentimentos para cada aspecto.

A maioria dos trabalhos relacionados com extração de aspectos e classificação de sentimentos é baseada em frequência, relações e modelos de tópicos. Entretanto, para que essas abordagens tenham um bom desempenho, elas utilizam recursos adicionais: (i) léxico de sentimentos, (ii) dicionários, (iii) conjunto de regras de relações, entre outros. Para definir esses recursos são necessários especialistas de domínio (esforço humano), uma vez que os recursos são dependentes de domínio. Em geral, um termo que é aspecto em um domínio não necessariamente é aspecto em outros domínios. Por exemplo, os termos “comida” e “higiene do local” são aspectos no domínio de restaurante, porém não são aspectos em outros domínios como câmera digital e laptop. Além da dependência do domínio em relação aos aspectos, há também a dependência de domínio em relação à polaridade dos sentimentos dos aspectos. Por

exemplo, o termo “grande” quando associado ao aspecto “fila de espera” no domínio de restaurante, representa um sentimento negativo, mas quando associado ao aspecto “porção”, no mesmo domínio, representa um sentimento positivo.

Uma alternativa para diminuir a necessidade de recursos adicionais é o uso de aprendizado de máquina para Mineração de Opiniões (HU; LIU, 2004; ZHUANG; JING; ZHU, 2006; KOBAYASHI; INUI; MATSUMOTO, 2007; BLAIR-GOLDENSOHN *et al.*, 2008; SOMASUNDARAN *et al.*, 2009; QIU *et al.*, 2011). Aprendizado de Máquina é uma área que envolve métodos computacionais que são capazes de “aprender” um determinado comportamento a partir de observações de dados e experiências acumuladas (MITCHELL, 1997). Dentre as estratégias existentes de Aprendizado de Máquina, o aprendizado supervisionado é muito utilizado. Nesse processo é fornecido um conjunto de treinamento formado pelos exemplos e seus respectivos rótulos. Nesse caso, o objetivo do aprendizado é utilizar o conjunto de treinamento para aprender uma função de mapeamento entre os exemplos e rótulos (FLACH, 2012). É esperado que tal função seja capaz de mapear os exemplos não vistos (conjunto de teste) aos prováveis rótulos. Quando os rótulos são definidos como valores discretos sem relação de ordem, o método de aprendizado é chamado de classificação. Desse modo, para usar aprendizado de máquina supervisionado em Mineração de Opiniões baseada em Aspectos, um conjunto de treinamento é construído por meio da rotulação de aspectos pelos usuários para, posteriormente, aprender um classificador que identifique aspectos em textos de opinião não vistos. De forma análoga, os usuários também rotulam exemplos positivos, negativos e neutros que envolvam os aspectos, possibilitando aprender um classificador que defina o sentimento de aspectos ainda não vistos. Embora trabalhos envolvendo aprendizado supervisionado sejam o estado da arte na área, existe uma limitação prática devido a necessidade de rotular um grande conjunto de dados para a tarefa de aprendizado, e isso também exige um grande esforço humano.

Assim, apesar de diversas propostas para tratar os desafios da Mineração de Opiniões baseada em Aspectos, ainda existem desafios que, em alguns casos, limitam a aplicação prática. Os desafios de pesquisa desta tese de doutorado estão relacionados a como diminuir a necessidade de grande esforço humano tanto para rotular dados, quanto para tratar a dependência de domínio para as tarefas de extração de aspectos e classificação de sentimentos dos aspectos.

Uma alternativa promissora para reduzir a necessidade de dados rotulados é a investigação de métodos de aprendizado semissupervisionado para Mineração de Opiniões baseada em Aspectos. Nesse caso, considera-se que a quantidade de exemplos rotulados é muito menor (exigindo menos esforço humano), porém, necessita-se de métodos específicos para lidar com este cenário (CHAPELLE; SCHÖLKOPF; ZIEN, 2006). Entre os métodos existentes, vale destacar o aprendizado semissupervisionado por meio de redes, no qual a informação do problema é modelada em um grafo e a topologia da rede é explorada para propagar a informação dos exemplos rotulados para dados não rotulados. Em algumas situações, inclusive, o uso de aprendizado semissupervisionado em redes possibilita obter desempenho similar ou até superior aos

métodos de aprendizado supervisionado (LIU; HU; CHENG, 2005; JINDAL; LIU, 2006; JIN; HO; SRIHARI, 2009; JAKOB; GUREVYCH, 2010; CHOI; CARDIE, 2010; YANG; CARDIE, 2013).

Para diminuir a dependência de domínio para extração de aspectos, uma alternativa é aproveitar os aspectos rotulados em um domínio para suportar a extração de aspectos em outro domínio, em que não há aspectos rotulados (CHEN; MUKHERJEE; LIU, 2014; ZHANG *et al.*, 2016; RANA; CHEAH, 2017). Essa estratégia é denominada por Aprendizado por Transferência entre Domínios (*Cross-Domain Transfer Learning*) e foi investigada como uma solução promissora para problemas em que a rotulação de dados é impossível ou custosa (PAN; YANG, 2010; SCHOUTEN; FRASINCAR, 2016). O uso de métodos de aprendizado semissupervisionado por meio de redes e aprendizado por transferência entre domínios para Mineração de Opiniões baseada em Aspectos apresentam muitos desafios. A seguir são apresentados os desafios relacionados a esta tese.

1. **Representação de textos das revisões em redes para a extração de aspectos e classificação de sentimentos:** existem diversas possibilidades de representações de textos das revisões para classificação do documento como um todo, porém a representação em redes para a extração de aspectos e respectiva classificação de sentimento por meio de aprendizado semissupervisionado ainda é um desafio relevante e pouco explorado nessa área.
2. **Representações de textos das revisões em redes para diferentes domínios de aplicação:** algumas abordagens do estado da arte utilizam recursos adicionais como dicionários de domínio ou ontologias, para melhorar o desempenho da classificação. Desse modo, essas representações são dependentes de domínio e não podem ser diretamente utilizadas em outras aplicações. Sendo assim, o desenvolvimento de redes para a extração de aspectos e classificação de sentimentos que sejam independentes de domínio é um desafio relevante dessa área, tanto para o aprendizado semissupervisionado, quanto para o aprendizado por transferência entre domínios.
3. **Algoritmos de aprendizado de máquina eficientes para Mineração de Opiniões baseada em Aspectos sem o uso de grande quantidade de dados rotulados:** em geral, algoritmos baseados em aprendizado de máquina para extração de aspectos e classificação de sentimentos que apresentam resultados superiores utilizam uma grande quantidade de dados rotulados. Porém, essa área apresenta poucos exemplos rotulados devido a diferentes tipos de domínios de aplicações. Além disso, os algoritmos de aprendizado de máquina que utilizam poucos exemplos rotulados obtêm resultados inferiores. Portanto, é um desafio relevante a adaptação ou o desenvolvimento de algoritmos de aprendizado de máquina para extração de aspectos e classificação de sentimentos que sejam eficientes e que apresentem resultados superiores ao estado da arte.

4. **Algoritmos de aprendizado de máquina eficientes que façam um melhor uso das representações em redes e que sejam independentes de domínio para Mineração de Opiniões baseada em Aspectos:** em geral, algoritmos baseados em aprendizado de máquina para extração de aspectos e classificação de sentimentos que apresentam resultados superiores utilizam recursos adicionais e dependentes de domínio. Por outro lado, os algoritmos que não utilizam esses recursos obtêm resultados inferiores. Portanto, é um desafio relevante o desenvolvimento de algoritmos de classificação baseados em redes para extração de aspectos e classificação de sentimentos que sejam independentes de domínio e apresentem resultados melhores que o estado da arte.

1.2 Questões de Pesquisa e Objetivos

Motivados pelos desafios discutidos na Seção 1.1 e pela necessidade de Mineração de Opiniões baseada em Aspectos de modo automático, o objetivo geral desta tese de doutorado é contribuir com novas técnicas para as tarefas de extração de aspectos e classificação de sentimentos de revisões de produtos e serviços que minimizam a necessidade de interferência humana tanto para tratar a dependência de domínio, quanto para rotular dados. Assim, nesta tese de doutorado os objetivos gerais são divididos em dois grupos de acordo com as respectivas motivações: (I) minimizar necessidade de dados rotulados e (II) minimizar a necessidade de recursos adicionais dependentes de domínio. A seguir são apresentadas as questões de pesquisa organizadas nesses grupos.

(I) A maioria dos métodos para Mineração de Opiniões baseada em Aspectos que utilizam aprendizado de máquina supervisionado necessita de grande quantidade de dados rotulados para treinamento. Com o intuito de utilizar menos dados rotulados, nesta tese é investigado o uso de métodos de aprendizado semissupervisionado por meio de redes. As seguintes questões de pesquisa norteiam o desenvolvimento deste trabalho:

- **QP-I.1:** *Como as representações de textos das revisões de produtos e serviços podem impactar no processo de extração de aspectos por meio de aprendizado semissupervisionado baseado em redes?*
- **QP-I.2:** *Como as representações de textos das revisões de produtos e serviços podem impactar no processo de classificação de sentimentos dos aspectos por meio de aprendizado semissupervisionado baseado em redes?*
- **QP-I.3:** *Qual a influência da quantidade de exemplos rotulados no processo de extração de aspectos e classificação dos sentimentos?*
- **QP-I.4:** *O uso de métodos de aprendizado semissupervisionados baseado em redes pode aperfeiçoar o processo de extração de aspectos?*

- **QP-I.5:** *O uso de métodos de aprendizado semissupervisionados baseado em redes pode aperfeiçoar o processo de classificação de sentimentos dos aspectos?*

Para responder essas questões de pesquisa, os objetivos específicos são:

- Propor e avaliar uma representação de textos das revisões de produtos e serviços para extração de aspectos para uso de aprendizado semissupervisionado baseado em redes.
- Propor e avaliar uma representação de textos das revisões de produtos para classificação de sentimentos de aspectos para uso de aprendizado semissupervisionado baseado em redes.
- Avaliar o desempenho e a influência da quantidade de exemplos rotulados no processo de extração de aspectos e classificação dos sentimentos dos aspectos.
- Avaliar se o uso de métodos de aprendizado semissupervisionado baseado em redes podem aperfeiçoar o processo de extração de aspectos e no processo de classificação de sentimentos dos aspectos.

(II) Os métodos de extração de aspectos para Mineração de Opiniões baseada em Aspectos são dependentes dos exemplos do conjunto de treinamento. Se existem apenas exemplos rotulados de um determinado domínio para treinamento do modelo de aprendizado, somente serão aprendidas informação desse domínio. Com o propósito geral de minimizar a necessidade de esforço humano para definir recursos adicionais dependentes de domínio, nesta tese são investigados métodos independentes de domínio e de aprendizado por transferência entre domínios. Assim, as seguintes questões de pesquisa norteiam o desenvolvimento deste trabalho:

- **QP-II.1:** *Como a ausência de informações de domínio pode impactar os métodos de extração de aspectos para Mineração de Opiniões?*
- **QP-II.2:** *Como as informações de outros domínios podem ser utilizadas para extrair aspectos de um domínio não conhecido?*
- **QP-II.3:** *Como representações de textos de revisões podem impactar no aprendizado por transferência para extração de aspectos?*
- **QP-II.4:** *Qual o impacto das informações de outros domínios no processo de extração de aspectos de um domínio não conhecido?*

Nesse sentido, para responder essas questões de pesquisa, os objetivos específicos para diminuir a dependência de domínio dos métodos de extração de aspectos são:

- Avaliar métodos não-supervisionados para extração de aspectos que não utilizem informações de domínio.

- Propor e avaliar uma representação de textos das revisões de produtos para extração de aspectos para uso de aprendizado por transferência entre domínios.
- Propor e avaliar um método para extração de aspectos por meio de aprendizado por transferência entre domínios.
- Avaliar como as informações de outro domínio utilizadas no aprendizado por transferência entre domínios impactam no processo de extração de aspectos.

1.3 Principais Resultados

Baseado nos desenvolvimentos das propostas para atingir os objetivos definidos nesta tese de doutorado, foram obtidas as seguintes contribuições científicas:

- Proposta de uma nova representação de textos para coleções de revisões para produtos e serviços para extração de aspectos;
- Proposta de uma nova representação de textos para coleções de revisões para produtos e serviços para classificação de sentimentos de aspectos;
- Proposta de um novo uso de aprendizado semissupervisionado para Mineração de Opiniões baseada em Aspectos;
- Proposta de novas estratégias para métodos extração de aspectos independentes de domínio;
- Proposta de uma nova Representação de Textos para coleções de revisões para produtos e serviços para aprendizado por transferência entre domínios;
- Proposta de um novo método para propagação de rótulos para aprendizado por transferência entre domínios;
- Padronização, pré-processamento e disponibilização de um conjunto de coleções de revisões de produtos e serviços;
- Proposta e desenvolvimento de novo sistema de rotulação de aspectos e respectivos sentimentos com algoritmos de sugestão de rótulos (*Rotulabic-Aspect Sentiment*);
- Divulgação dos resultados científicos diretos desta tese por meio de artigos:

- PUBLICADOS:

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 - JIDM*, 2016.

MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Análise de Sentimentos baseada em Aspectos usando Aprendizado Semissupervisionado em Redes Heterogêneas. In: Symposium on Knowledge Discovery, Mining and Learning, 2015, Petrópolis. Proceeding of the 3rd Symposium on Knowledge Discovery, Mining and Learning, 2015. p. 90-93.

FARIAS, D. S.; MATSUNO, I. P.; MARCACINI, R. M.; REZENDE, S. O. Opinion-meter: um Framework para Análise de Sentimentos Baseada em Aspectos. In: Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2016, Teresina-PI.

FARIAS, D. S.; MATSUNO, I. P.; REZENDE, S. O. Annotation of Corpus for Aspect-Based Sentiment Analysis. In: Simpósio Internacional de Iniciação Científica e Tecnológica da USP – SIICUSP, 2016.

– SUBMETIDOS ou EM DESENVOLVIMENTO:

MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Towards domain-independent aspect extraction methods for sentiment analysis. Expert Systems with Applications, 2018. [submetido]

MARCACINI, R. M.; MATSUNO, I. P.; ROSSI, R. G.; REZENDE, S. O. Cross-domain aspect extraction for sentiment analysis: a transductive learning approach. Decision Support Systems, 2018. [em revisão para publicação]

MATSUNO, I. P.; FARIAS, D. S.; MARCACINI, R. M.; REZENDE, S. O. Rotulabic-Aspect Sentiment: Sistema de Anotação com Sugestão de Rótulos para Mineração de Opiniões baseada em Aspectos.[em desenvolvimento]

ANZAI, R.; MATSUNO, I. P.; Dos SANTOS, A. C.; MARCACINI, R. M.; REZENDE, S. O. Estudo Comparativo de Métodos de Extração de Aspectos para Mineração de Opiniões. [em desenvolvimento].

– Artigos científicos publicados e submetidos em outras áreas relacionadas:

MARCACINI, R. M.; SINOARA, R. A. ; MATSUNO, I. P.; REZENDE, S. O. Aprendizado não supervisionado de websensors a partir de dados textuais. In: Symposium on Knowledge Discovery, Mining and Learning (KDMile).[S.l.:s.n.], 2013.

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 (IGDP), v. 13, p. 57-66, 2015.

MARQUES, C. A. N. ; MATSUNO, I. P.; SINOARA, R. A. ; REZENDE, S. O.; ROZENFELD, H. An exploratory study to evaluate the practical application of PSS methods and tools based on text mining. In: International Conference on Engine-

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

BARBOSA, J. R.; MATSUNO, I. P.; GUIMARAES, E. R.; REZENDE, S. O.; VINCENZI, A. M. R.; DELAMARO, M. E. Mineração de Textos para Apoiar a Predição de Severidade de Relatórios de Incidentes: um Estudo de Viabilidade. In: Simpósio Brasileiro de Qualidade de Software, 2017, Rio de Janeiro. Simpósio Brasileiro de Qualidade de Software, 2017. p. 89-103.

BARBOSA, J. R.; MATSUNO, I. P.; MARCACINI, R. M.; REZENDE, S. O.; VINCENZI, A. M. R.; DELAMARO, M. E. Predicting the Severity of Bug Reports with Semi-Supervised Learning: A Case Study, The Journal of Systems and Software, 2018 [submetido]

1.4 Organização da Tese

O restante desta tese está organizado do seguinte modo:

Capítulo 2 – Fundamentos e Trabalhos Relacionados. Neste capítulo são apresentados os conceitos e terminologias utilizados nesta tese sobre Mineração de Opiniões baseada em Aspectos e Aprendizado de Máquina Semissupervisionado.

Capítulo 3 – Método Semissupervisionado para Mineração de Opiniões baseada em Aspectos. Neste capítulo é apresentada uma abordagem denominada ASPHN (*Aspect-based Sentiment Propagation on Heterogeneous Networks*) proposta nesta tese. Na ASPHN são propostas: (i) representação de textos das coleções de revisões para a extração de aspectos; (ii) representação de textos das coleções de revisões para classificação de sentimentos; e (iii) uso de métodos semissupervisionados transdutivos. Por fim, uma avaliação experimental para avaliar a proposta ASPHN é apresentada.

Capítulo 4 – Métodos Não-supervisionados Independentes de Domínio para Extração de Aspectos para Mineração de Opiniões. Neste capítulo são apresentados métodos para extração de aspectos com base em critérios estatísticos e estratégias propostas para eliminar o uso de léxicos predefinidos e de construção de regras de sintaxe supervisionada. Na sequência é proposto um método para avaliação dos aspectos extraídos.

Capítulo 5 – Método *Cross-Domain* para Extração de Aspectos para Mineração de Opiniões. Neste capítulo é apresentada uma abordagem para extração de aspectos entre domínios baseada na aprendizagem transdutiva denominada por CD-ALPHN (*Cross-Domain Aspect Label Propagation through Heterogeneous Networks*) proposta nesta tese. Na CD-ALPHN são propostas: (i) nova representação de textos das coleções de domínio conhecido (origem), e (ii) novo algoritmo de propagação de rótulos entre os domínios

(origem e destino). Além disso, é apresentada uma avaliação prática experimental em sistemas de suporte à decisão.

Capítulo 6 – Projetos Pilotos. Neste capítulo são apresentados três projetos pilotos: (i) *Rotulabic Aspect-Sentiment*; (ii) *Opinion-Meter*; e (iii) Comparação entre método baseado em relações de regras sintáticas e de aprendizado de máquina supervisionado.

Capítulo 7 – Conclusões. Neste capítulo são apresentadas as conclusões desta tese. Os objetivos e as respectivas questões de pesquisa propostos no Capítulo de Introdução são retomados e discutidos. Por fim, são descritas as limitações deste trabalho e direções para trabalhos futuros.

FUNDAMENTOS

2.1 Considerações Iniciais

Neste capítulo são apresentados os conceitos e trabalhos relacionados a esta tese. Na Seção 2.2, são apresentadas uma revisão da literatura sobre Mineração de Opiniões baseada em Aspectos com ênfase nas tarefas de extração de aspectos e classificação de sentimentos. Na Seção 2.3 são apresentadas as representações de coleção de dados textuais utilizadas para o aprendizado de máquina. Os conceitos e os métodos de aprendizado de máquina relacionados ao trabalho desenvolvido na tese são apresentados na Seção 2.4. Por fim, na Seção 2.5 são apresentadas as considerações finais do conteúdo desse capítulo.

2.2 Mineração de Opiniões baseada em Aspectos

A Mineração de Opiniões, também chamada de Análise de Sentimentos, é uma área de pesquisa que estuda as opiniões, sentimentos, avaliações e emoções das pessoas, em relação às entidades e seus respectivos aspectos (atributos), como produtos, serviços, organizações, pessoas, tópicos, problemas, eventos (LIU, 2012). O início e o rápido crescimento dessa área de pesquisa coincidem com a rápida emergência da comunicação social na Internet por diversos meios, por exemplo, comentários, fóruns de discussão, *blogs*, *microblogs*, *Twitter* e redes sociais, pois, pela primeira vez na história humana, tem-se um volume grande de dados opinativos registrados em formulários digitais (GANTZ; REINSEL, 2010). Devido à importância das opiniões para as empresas e a sociedade como um todo, nos últimos anos, há um crescimento de pesquisas acadêmicas (LIU, 2012), e desenvolvimento de sistemas industriais que usam Mineração de Opiniões para avaliar seus produtos e serviços (WANG W. E ZHOU, 2009). Sistemas de Mineração de Opiniões podem ser aplicadas a praticamente todos os negócios e domínio social. Desde o início de 2000, essa área de pesquisa trabalha com grandes desafios, pois os dados são dependentes de idioma, de domínios de aplicação e com problemas de

representação (gírias, símbolos, etc.). Essa área cresceu e tem sido tema de pesquisa nas áreas de Processamento de Linguagem Natural, Mineração de Dados, Mineração de Web, Mineração de Textos e Recuperação da Informação (ARCHACK; IPEIROTIS, 2007; DAS; CHEN, 2007; DELLAROCAS; AWAD, 2007; GHOSE; SUNDARARAJAN, 2007; HU; PAVLOU; ZHANG, 2006; PARK; LEE; HAN, 2007)

A Mineração de Opiniões pode ser feita considerando as seguintes granularidade: de documentos, de sentenças e de aspectos. No nível de documentos visa classificar o sentimento do documento como um todo em “positivo”, “negativo” ou “neutro”. Neste tipo de análise considera que o documento contém informações referentes a um único assunto (tópico/tema/entidade/produto/serviço) (PANG; VAITHYANATHAN, 2002; TURNEY, 2002). Já em nível de sentença visa classificar sentimento expresso em cada frase de um documento. Neste nível de análise, o primeiro passo é identificar se a sentença é objetiva (ausência de sentimentos/opiniões) ou subjetiva (contém sentimentos/opiniões) (WIEBE; BRUCE; O’HARA, 1999). Se a sentença é subjetiva, será determinado se a sentença inteira expressa uma opinião ‘positiva’, “negativa” ou “neutra”.

Nesses dois primeiros tipos de Mineração de Opiniões não são apresentados informações sobre características da entidade analisada. Para obter a classificação dessas características utiliza-se a Mineração de Opiniões em nível de aspectos (características, atributos) das entidades. Em uma sentença, podem ser expressos sentimentos diferentes e de aspectos distintos, por exemplo, a sentença “A câmera e o design desse celular são excelentes, mas a qualidade de voz e a bateria é muito ruim”. Nessa sentença é observado opiniões positivas e negativas. Além disso, cada opinião refere-se aos aspectos diferentes (“câmera”, “design”, “qualidade de voz”, “bateria”) da entidade em questão (“celular”). Este projeto de doutorado tem como foco a Mineração de Opiniões em nível de aspecto, também chamada de Mineração de Opiniões baseada em Aspectos.

Os conceitos fundamentais e terminologias da Mineração de Opiniões baseada em Aspectos utilizados nesta tese são baseadas nas definições apresentadas em Liu (2012). Alguns termos foram adaptados para utilização na proposta deste trabalho. Essas definições relacionadas à Mineração de Opinião baseada em aspecto são baseadas em Liu (2015).

Opinião: de acordo com o dicionário¹ são apresentadas as seguintes definições para opinião.

*1 Maneira de opinar; modo de ver pessoal; parecer, voto emitido ou manifestado sobre certo assunto. 2 Asserção sem fundamento; presunção. 3 Conceito, reputação. 4 Juízo ou sentimento que se manifesta em assunto sujeito a deliberação. 5 Capricho, teimosia. Opinião pública: juízo coletivo adotado e exteriorizado por um grupo ou, em sociedades diferenciadas e estratificadas, por diversos grupos ou camadas sociais.*² Outras definições

¹ Michaelis. Opinião. <http://michaelis.uol.com.br/moderno/portugues/index.php>

² Michaelis. Opinião. <http://michaelis.uol.com.br/moderno/portugues/index.php?lingua=portugues-portugues&palavra=opinio> [Acessado em: 20/05/2015]

são que *opinião* é a ideia confusa acerca da realidade e que se opõe ao conhecimento verdadeiro³; e/ou *opinião* é um julgamento, ponto de vista, ou declaração sobre assuntos considerados comumente como subjetivos.⁴

Sentimento: são apresentadas as seguintes definições para sentimentos. *Sentimento* é 1 Ação ou efeito de sentir. 2 Faculdade ou capacidade de sentir, de receber impressões mentais. 3 Sensação psíquica, tal como as paixões, o pesar, a mágoa, o desgosto, etc. 4 Disposição para ser facilmente comovido ou impressionado. 5 Emoção terna ou elevada, tal como o amor, a amizade, o patriotismo. 6 Atitude mental a respeito de alguém ou de alguma coisa: Sentimento de estima, de respeito, de ódio, do dever. 7 Exibição ou manifestação de sensibilidade ou sentimentalismo, ou propensão às emoções ternas em literatura, arte ou música. 8 Pressentimento, suspeita. **9 Opinião.** 10 Faculdade intuitiva de perceber ou apreciar as qualidades ou méritos de uma coisa. 11 Filos Conjunto de emoções.⁵ Como observado nas definições *opinião* e *sentimento* podem ser considerados sinônimos em algumas circunstâncias.

Documento (d): é um texto em língua natural que relata sobre um determinado assunto, tema, problema, produto, organização, etc. Este texto pode representar uma mensagem, uma revisão, uma análise ou uma notícia sobre assunto específico.

Conjunto de Documentos (D): é uma coleção de documentos d sobre um ou mais assuntos específicos.

Entidade (e): para $0 < i \leq n$, uma entidade e_i é um produto, serviço, tópico, pessoa, organização, acontecimento ou evento do qual está sendo referenciada nos documentos. Uma entidade é descrita por um conjunto de componentes e seus aspectos (HU; LIU, 2004; LIU, 2006). Exemplos de entidades são: produto, serviço, tópico, pessoa, organização, acontecimento ou evento. Na literatura, entidades podem ser denominadas por *objetos*.

Aspectos (a): um aspecto a_{ij} é uma propriedade, componente, características de entidade. Uma entidade pode ser descrita por um conjunto de componentes suas características. Exemplos de aspectos são: *tamanho* do produto, *preço* do produto, *qualidade* de atendimento de um serviço. Na literatura, aspectos podem ser denominados por *características* ou *atributos*.

O sentimento é um termo usado na linguística que se refere à direção em que um conceito ou opinião é interpretado. Neste trabalho, o termo **opinião** é usado para denotar um texto em linguagem natural descrevendo a opinião, avaliação, atitude e emoção em relação a um assunto específico. Essa opinião expressa um sentimento (bom, neutro ou ruim) em relação ao objeto

³ <[pt.wikipedia.org/wiki/Opini~ao](http://pt.wikipedia.org/wiki/Opini%C3%A7%C3%A3o)>

⁴ <http://en.wikipedia.org/wiki/Opinion>

⁵ <http://michaelis.uol.com.br/moderno/portugues/index.php?lingua=portugues-portugues-palavra=sentimento>

alvo opinado. Formalmente, uma opinião O será definida pela quintupla, $O = (e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, em que:

- e_i é o nome de uma entidade;
- a_{ij} é um aspecto (característica, propriedade) da entidade e_i ;
- h_k é o titular da opinião;
- t_l é o momento em a opinião O foi expressa por h_k ; e,
- s_{ijkl} é a polaridade do sentimento em relação ao aspecto a_{ij} da entidade e_i expressa por h_k no tempo t_l .

Nesta tese será considerado apenas a tripla, $O = (e_i, a_{ij}, s_{ijkl})$, e para cada coleção de revisões de opiniões referencia uma única entidade.

A polaridade do sentimento s_{ijkl} ‘positiva’, ‘negativa’ ou ‘neutra’. Essa polaridade também pode ser expressa de outros modos diferentes, tais como categorias ou diferentes níveis de força ou de intensidade. Um exemplo para expressão do sentimento é o uso de uma a cinco estrelas usado por sistemas de avaliação e de recomendações na *web*.

A Mineração de Opiniões tem por objetivo extrair automaticamente o sentimento expressos nas opiniões descritas em diversos documentos tendo como foco um objeto alvo. Considerando as definições da seção anterior, pode-se dizer que a Mineração de Opiniões tem por objetivo descobrir todas as quintuplas de opiniões O expressas em um documento de opinião d , ou em uma coleção de documentos D em relação a um objeto alvo. Neste caso, o objeto alvo será definido pela entidade e_i e seus aspectos a_{ij} . Assim, uma entidade e_i é um objeto do mundo real, como um produto, pessoa, acontecimento, organização ou tópico. Um aspecto a_{ij} é definido por uma característica de e_i . Se e_i é um produto, então características do produto como tamanho (a_{i1}), cor (a_{i2}), preço (a_{i3}) e sabor (a_{i4}) são aspectos de e_i . Por exemplo, na frase “A *televisão Sony* tem excelente qualidade de imagem.”, a entidade é “*televisão Sony*” e “*excelente*” é um sentimento para o aspecto “*qualidade de imagem*”. Já na frase “A *televisão Sony* é muito boa!.” a opinião expressa é sobre o objeto alvo como um todo, nesses casos, é considerado que o aspecto da entidade é “*GERAL*” (a_{i0}).

Todo o processo de Mineração de Opiniões Baseado em Aspectos é bastante complexo e envolve técnicas de diversas áreas de pesquisa. De forma geral, este processo pode ser dividido nas seguintes tarefas principais: (1) extração e categorização das entidades; (2) extração e categorização dos aspectos; (3) extração e categorização do titular da opinião; (4) extração e padronização do tempo; (5) classificação do sentimento do aspecto a da entidade e ; (6) geração das quintuplas de opinião; e, (7) sumarização e visualização das opiniões. Na Figura 1 é apresentado um esquema que resume o Processo de Análise de Sentimentos, na sequência, uma explicação breve de cada tarefa é apresentada. Embora, as tarefas sejam apresentadas em uma sequência na Figura 1, elas podem ser executadas em paralelo e/ou integradas.

Figura 1 – Processo de mineração de opiniões

Fonte: Elaborada pelo autor.

1. **Extração e categorização da entidade:** identificar todas as expressões de entidade em d , e categorizar as expressões de entidades sinônimas em categorias de entidade. Cada expressão de entidade indica uma única entidade única e_i .
2. **Extração e categorização do aspecto:** identificar todas as expressões de aspectos das entidades, e categorizar essas expressões de aspectos em categorias. Cada categoria de aspecto da entidade e_i representa um aspecto único a_{ij} .
3. **Extração e categorização do titular da opinião:** identificar quem emitiu a opinião ou dos dados emissor de opinião. Esses dados devem ser estruturados e categorizados.
4. **Extração e padronização do tempo:** identificar quando as opiniões foram publicadas (emitidas) e padronizar os formatos de tempo diferentes.
5. **Classificação do sentimento do aspecto:** determinar a polaridade do sentimento sobre um aspecto a_{ij} , ou seja, classificar o sentimento em positivo, negativo ou neutro. A classificação pode ser numérica (1, 2, 3,...), de categorias (regular, ruim, bom, muito bom, ótimo) ou outras formas de classificação.
6. **Geração das quintuplas de opinião:** produzir todas as quintuplas de opinião $O = (e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ expressas no(s) documento(s) d com base nos resultados das tarefas anteriores.
7. **Sumarização:** um documento de opinião d não é suficiente para a tomada de decisão, é necessário realizar todas as tarefas anteriores para um conjunto de documentos D e então computar e sumarizar as opiniões. A sumarização de texto consiste em reduzir um documento ou múltiplos documentos com textos longos em um conjunto de palavras ou de

sentenças que possibilite maior compreensão do significado do texto. Portanto, as opiniões são ordenadas, categorizadas e sumarizadas de forma que os resultados sobre o objeto alvo (entidade e seus aspectos e respectivos sentimentos) sejam apresentados de forma que possibilite melhor interpretação dos dados.

Para exemplificar as etapas de Mineração será considerada a mensagem apresentada na Figura 2 extraída de um *blog*. Na sequência será apresentado qual o resultado esperado na execução de cada tarefa do Processo de Mineração de Opiniões baseada em Aspectos.

Figura 2 – Exemplo de mensagem de opinião

Posted by: bigJohn Date: 15/set/2011

"(1)Eu comprei uma câmera Samsung e meu amigo pegou uma câmera Canon ontem. (2)Na semana passada, nós usamos muito as câmeras. (3)As fotos da minha Samy não são tão grandes, e a duração da bateria é curta demais. (4)O meu amigo estava muito feliz com a câmera dele e amou a qualidade de suas imagens. (5)Eu quero uma câmera que possa tirar boas fotos. (6)Eu vou devolvê-la amanhã."

Fonte: Adaptada de Liu (2012).

Os resultados esperados para cada tarefa são:

1. **Extração e categorização das entidades:** nesta etapa são identificadas por exemplo as expressões “*Samsung*”, “*Samy*” e “*Canon*”. As expressões “*Samsung*”, “*Samy*” representam a mesma entidade “*Câmera Samsung*” .
2. **Extração e categorização dos aspectos:** identifica as expressões tais como: “*imagem*”, “*foto*”, e “*duração da bateria*”. As duas primeiras expressões são sinônimos para o aspecto “*imagem*” das câmeras. Além disso, relacionar o aspecto à entidade relacionada.
3. **Extração do(s) titular(es) da opinião:** na sentença (3), pode ser *bigJohn* (o autor do *blog*) e na sentença (4) pode ser o *amigo de bigJohn*.
4. **Extração e padronização do tempo:** a mensagem foi postada no blog em 15 de setembro de 2011. Deve-se colocar no formato padrão a ser utilizado, por exemplo, *2001-09-15*
5. **Classificação do sentimento do aspecto:** a sentença (3) emite uma opinião negativa da qualidade da imagem da câmera *Samsung* e da duração da bateria. A sentença (4) emite uma opinião positiva para a câmera *Canon* como um todo (aspecto=*GERAL*) e também da sua qualidade de imagem. A sentença (5) aparentemente expressa uma opinião positiva, mas não é. Para gerar as quintuplas de opinião contidas na sentença (4), é necessário conhecer a qual câmera referem-se as expressões: “*a câmera dele*” e “*suas*”.

6. **Geração das quintuplas de opinião**: após a execução dessas etapas as quintuplas geradas são apresentadas na Figura 3.
7. **Sumarização**: a sumarização pode ser apresentada de várias formas. Neste exemplo, optou-se por apresentar a quantidade de opiniões positivas e negativas organizadas por cada aspecto da entidade. Na Figura 4 um exemplo de resumo é apresentado.

Figura 3 – Exemplo das quintuplas de opinião

$O_1 = (\text{Samsung, qualidade_imagem, negativo, bigJohn, 15/set/2011})$
$O_2 = (\text{Samsung, tempo_bateria, negativo, bigJohn, 15/set/2011})$
$O_3 = (\text{Canon, geral, positivo, Amigo de bigJohn, 15/set/2011})$
$O_4 = (\text{Canon, qualidade_imagem, positivo, Amigo de bigJohn, 15/set/2011})$

Fonte: Adaptada de Liu (2012).

Figura 4 – Sumarização das opiniões

Câmera Digital: Canon
Aspecto: GERAL
Positivo: 105 < sentenças individuais >
Negativo: 12 < sentenças individuais >
Aspecto: Qualidade da Imagem
Positivo: 95 < sentenças individuais >
Negativo: 10 < sentenças individuais >
Aspecto: Duração da Bateria
Positivo: 50 < sentenças individuais >
Negativo: 9 < sentenças individuais >

Fonte: Adaptada de Liu (2012).

Como pôde ser observado, a Mineração de Opiniões Baseada em Aspectos, na maioria dos casos, é bastante subjetiva e apresenta uma série de desafios para realização desse processo de forma automática. Nesta tese os desafios estão relacionados às tarefas de extração de aspectos e classificação de sentimentos associados aos aspectos. Os trabalhos relacionados com essas tarefas são apresentadas nas Seções 2.2.1 e 2.2.2.

2.2.1 Trabalhos Relacionados com Extração de Aspectos

Em Liu (2012) foram propostas quatro categorias de abordagens para a extração de aspectos de acordo com as técnicas que estas utilizam. Embora as abordagens possam utilizar mais que uma técnica (abordagens híbridas), as principais abordagens para extração de aspectos

são baseadas em: (i) frequência; (ii) relações; (iii) aprendizado supervisionado; e, (iv) modelos de tópicos.

2.2.1.1 Extração baseada em frequência

Um aspecto pode ser expresso por um substantivo, adjetivo, verbo ou advérbio. No entanto, uma pesquisa mostra que 60-70% dos aspectos explícitos são substantivos (LIU, 2006). Devido ao fato que em comentários de opinião, as pessoas são mais propensas a falar sobre os aspectos relevantes, isso sugere que, em geral, os aspectos devem ser substantivos frequentes. No entanto, nem todos os aspectos que são substantivos, e nem todos os aspectos são frequentes. Por isso, diferentes técnicas de seleção são aplicadas em substantivos frequentes para distinguir quais dos substantivos podem ser aspectos ou não. Métodos que baseiam na frequência normalmente aplicam um conjunto de restrições em sintagmas nominais de alta frequência para identificar aspectos.

Hu e Liu (2004), Popescu e Etzioni (2005), Long, Zhang e Zhut (2010) apresentam métodos para extração de aspectos baseado na frequência de substantivos presentes nos textos analisados. Em geral, esses métodos geram um conjunto de candidatos a aspectos e propõem um critério de seleção que podem ser baseados em coocorrência, medida da *Pointwise Mutual Information* (PMI), padrão sintático, entre outros. Outras abordagens são aquelas que baseiam-se em um léxico de sentimentos e nas relações das palavras candidatas a aspectos às palavras de sentimentos (HU; LIU, 2004; ZHUANG; JING; ZHU, 2006). Os métodos baseados em frequência dos termos nas revisões são independentes de domínio e de idioma, porém utilizam para tratar são consideradas e são dependentes domínio.

2.2.1.2 Extração baseada relações

Devido ao fato das opiniões terem um alvo (entidades ou aspectos), é natural que existam muitas relações entre as expressões de sentimento e os alvos de opinião. Devido ao fato das palavras e frases de sentimentos são muitas vezes conhecidas é possível explorar essas relações. Essas relações podem ser de (i) dependências sintáticas que descrevem as relações de opinião e alvo de opinião; (ii) codificação de padrões léxico-sintática; e (iii) as relações de parte/atributo.

Se a sentença não tem um aspecto frequente, mas tem algumas palavras de sentimento, o substantivo mais próximo a uma palavra de sentimento é extraído como um aspecto (HU; LIU, 2004; BLAIR-GOLDENSOHN *et al.*, 2008). Por exemplo, na frase “Esta câmera tira fotos ótimas.” o termo que indica sentimento é “ótimas” e o alvo é opinião “fotos” e é possível uma relação entre esses dois termos palavras(substantivo-adjetivo/objeto-complemento).

Em Zhuang, Jing e Zhu (2006) é proposto o uso de um analisador de dependência sintática para identificar as relações de dependência das palavras de sentimento e seus alvos de opinião para extrair pares aspecto-sentimento. Depois da análise, as palavras de uma frase são ligadas por relações de dependência. Com base nos dados de treinamento, o algoritmo identifica

os primeiros modelos de relação de dependência confiáveis, depois usa essas relações para identificar pares aspecto-sentimento válidos nos dados de teste. Abordagens semelhantes a esta foram propostas por [Somasundaran et al. \(2009\)](#) e [Kobayashi, Inui e Matsumoto \(2007\)](#).

Em [Qiu et al. \(2009\)](#) e [Qiu et al. \(2011\)](#), a ideia de dependência foi generalizada, por meio do método de Dupla Propagação (DP) para a extração simultânea de ambas as palavras de sentimento e alvo de opinião. Este foi originalmente projetado para idioma inglês, mas também tem sido aplicado com sucesso em análise de *revisões* no idioma chinês ([ZHAI et al., 2011](#)). Este método pode também ser reduzido para encontrar apenas os aspectos utilizando um grande léxico sentimento e sem qualquer propagação.

Em [Wang e Wang \(2008\)](#) é utilizado um algoritmo de *bootstrapping* para identificar os aspectos do produto e as palavras de sentimento com base em uma lista inicial (semente) de palavras de sentimentos. A medida de Informação Mútua foi empregada para avaliar a associação entre candidatos a aspectos e palavras de sentimento. Além disso, regras linguísticas foram extraídas para identificar aspectos e palavras de sentimento que não são muito frequentes.

[Hai, Chang e Cong \(2012\)](#) apresenta uma proposta similar, porém, em vez de empregar um conjunto de palavras de sentimento como sementes, este método inicia com um pequeno conjunto de aspectos como sementes. O algoritmo aumenta de forma iterativa pela mineração de pares do tipo aspecto-opinião, aspecto-aspecto, e relações de dependência opinião-opinião. Nessa proposta, dois modelos de associação, ou seja, testes de razão de verossimilhança e análise semântica latente, foram utilizados para calcular as associações de pares entre essas expressões (aspectos e/ou opiniões).

2.2.1.3 Extração usando aprendizado supervisionado

Muitos algoritmos clássicos de aprendizado de máquina supervisionado são utilizados para a extração de informação e podem ser empregados nas tarefas de extração de aspectos ([MOONEY; BUNESCU, 2005](#); [SARAWAGI, 2008](#); [HOBBS; RILOFF, 2010](#)). Em geral, os métodos para extração de aspectos são baseados em aprendizagem sequencial (*sequential labeling*). Os métodos atuais de aprendizagem sequencial do estado-da-arte são *Conditional Random Field* (CRF) ([LAFFERTY; MCCALLUM; PEREIRA, 2001](#)) e *Hidden Markov Model* (HMM) ([RABINER, 1989](#)).

O trabalho de [Jakob e Gurevych \(2010\)](#) utiliza CRF para extrair aspectos de sentenças que contém opinião. O treinamento é feito com sentenças de diferentes domínios e utiliza recursos (atributos) adicionais, que também são independentes de domínio, como anotações de PoS, de dependência sintática, de distância entre as palavras e de expressões de sentimento. [Li et al. \(2010\)](#) apresenta uma proposta semelhante para modelar a dependência de longa distância entre as sentenças. Esse modelo também considera o uso das conjunções entre as sentenças: "e" (mantém a polaridade) e "mas" (inverte a polaridade). [Yang e Cardie \(2013\)](#) integram duas variações de CRF: (i) Skip-chain CRF e (ii) Tree-CRF para extrair opiniões e aspectos. [Choi](#)

e Cardie (2010), Qiu *et al.* (2009) e Qiu *et al.* (2011) também propõem abordagens utilizando CRF.

Jin, Ho e Srihari (2009) utilizam um HMM integrado a recursos linguísticos, como *part-of-speech* (POS) e padrões lexicais para identificar aspectos e suas polaridades. As classes consideradas são: aspecto; explícito positivo; explícito negativo; implícito positivo; implícito negativo; e *palavras conhecidas*. Um estado observável para o HMM nesta abordagem é representada por um par formado por uma palavra e sua respectiva anotação POS.

Os trabalhos de Liu, Hu e Cheng (2005) e Jindal e Liu (2006) utilizam ambos HMM e CRF. Nestes trabalhos são utilizados regras sequenciais, que são obtidas com base na mineração supervisionada do padrão sequencial de classes rotuladas.

Outros métodos de aprendizado também são utilizados, em Ghani *et al.* (2006) utilizam métodos tradicionais de aprendizado supervisionados e semissupervisionados para extrair os aspectos. Kobayashi, Inui e Matsumoto (2007) utiliza um método de classificação em árvore estruturada. Yu *et al.* (2011) utiliza método semissupervisionado *One-class SVM* (MANEVITZ; YOUSEF, 2002). Kovelamudi *et al.* (2011) utiliza método supervisionado, mas também utiliza algumas informações relevantes da *Wikipedia*. Klinger e Cimiano (2013) utilizam um grafo fator (KSCHISCHANG; FREY; LOELIGER, 2001) (grafo bipartido que representa uma fatoração de função) para extrair os alvos de opinião e expressões de opinião.

2.2.1.4 Extração baseada em Modelos de Tópicos

Algumas propostas utilizam modelos probabilísticos para extração dos aspectos e dos sentimentos. A modelagem de tópicos é um método de aprendizado não supervisionado que assume que cada documento é composto por um conjunto de tópicos e cada tópico tem uma distribuição de probabilidade sobre as palavras. Existem dois principais modelos de tópicos que são utilizados para extração de aspectos: *Latent Dirichlet Allocation*-LDA (BLEI; NG; JORDAN, 2003) e *Probabilistic Latent Semantic Indexing*-PLSI (HOFMANN, 1999).

Em Mei *et al.* (2007) é proposto um método baseado no PLSI para identificar aspectos e sentimentos utilizando palavras conhecidas de sentimentos na distribuição. Guo *et al.* (2009) propõe um método baseado no Latent Semantic Association (LaSA) para agrupar candidatos a aspectos. Nessa abordagem assume-se que as palavras com o mesmo contexto tem associação semântica similar.

O método proposto em Titov e McDonald (2008a) (expandido em Titov e McDonald (2008b)) é baseado no LDA e extrai dois tipos de tópicos: global e local. Os tópicos globais correspondem aos tipos dos produtos, como a marca ou modelo dos produtos, em geral, estes são conhecidos e fixos. Os tópicos locais correspondem aos aspectos dos tipos de produtos em questão. Com base na conclusão desse trabalho, em Lu, Zhai e Sundaresan (2009) é proposto um método que assume que cada um comentário pode ser analisado como uma frase de opinião

representada por um par $p = (\text{termo}, \text{modificador})$. Na sequência, todos comentários que possuem opinião são representados dessa forma. Um modelo probabilístico baseado em pLSI é utilizado para identificar os principais aspectos de um produto, agrupando os termos.

Em Zhao *et al.* (2010) é proposto um método híbrido que combina Entropia Máxima e LDA para identificar palavras que são aspectos e palavras de sentimentos para aspectos específicos. Esta abordagem também pode usar atributos linguísticos para separar palavras que indicam aspectos das palavras indicam de sentimentos.

Embora o modelo de tópicos seja uma abordagem baseada em princípios de inferência com base probabilística e que pode ser estendido para modelar muitos tipos de informação, ele tem algumas limitações que dificultam seu uso em aplicações de Mineração de Opiniões na vida real. A principal desvantagem é a necessidade de um grande volume de dados e uma quantidade significativa de ajustes para que se possa

2.2.2 *Trabalhos Relacionados com Classificação de Sentimentos dos Aspectos*

Para a tarefa de classificação de sentimentos dos aspectos existem duas principais abordagens baseadas em Liu (2012): (i) classificadores léxico e (ii) classificadores baseados em aprendizado supervisionado. Na sequência são apresentados os trabalhos relacionados com essas duas abordagens.

2.2.2.1 *Classificadores baseados em Léxico*

Os classificadores baseados em léxico são tipicamente não supervisionados. Esses usam um léxico de sentimentos (lista de palavras, frases e expressões idiomáticas que representam sentimento), expressões compostas, regras de opiniões e (possivelmente) a árvore sintática da sentença para determinar a orientação do sentimento de cada aspecto em uma frase. Esses classificadores também consideram “*modificadores de sentimento*”, “*cláusulas-mas*” e muitas outras construções gramaticais que possam afetar sentimentos.

Em geral, uma abordagem baseada em léxico para classificação de sentimentos em relação aos aspectos usa os seguintes recursos (DING; LIU; YU, 2008; LIU, 2010):

- um léxico de expressões de sentimento incluindo palavras de sentimento, frases, expressões idiomáticas e regras de composição;
- um conjunto de regras para lidar com diferentes construções de linguagem (por exemplo, modificadores de sentimento e cláusulas-mas) e tipos de frases; e,
- uma função de agregação sentimento ou um conjunto de sentimento e de alvo das relações derivadas da árvore sintática para determinar a orientação sentimento em cada destino em uma frase.

Em Blair-Goldensohn *et al.* (2008) é proposto um método baseado em um léxico integrado com aprendizado supervisionado. Em Kessler e Nicolov (2009) é utilizado quatro estratégias diferentes para determinar o sentimento de cada aspecto e também mostrou várias estatísticas sobre por que é tão difícil de relacionar as palavras sentimento aos seus alvos. Embora esses métodos não foram desenvolvidos para tratar frases comparativas, não é difícil adaptá-los para tratar esses casos como apresentado em Ganapathibhotla e Liu (2008) e Ding, Liu e Yu (2008). Existem muitos trabalhos baseados nessa abordagem (HU; LIU, 2004; KIM; HOVY, 2004; WAN, 2008; ZHU *et al.*, 2009).

As abordagens baseadas em léxico podem evitar alguns dos problemas da classificação baseada em aprendizado supervisionado e tem apresentado um bom desempenho em um grande número de domínios. Uma das limitações das abordagens baseada em léxico é a necessidade de um grande esforço e tempo de especialistas para construir a base de conhecimento inicial do léxico, padrões e regras. Já existe um legado acumulado ao longo dos anos que pode ajudar na construção inicial de sistemas de análise de sentimentos. E embora a abordagem baseada em léxico apresente ser independente do domínio, é necessário algum trabalho adicional é para cuidar das peculiaridades de cada domínio.

2.2.2.2 Classificadores baseados em Aprendizado de Máquina

Para classificação de sentimentos relacionadas aos aspectos, os algoritmos de aprendizado de máquina tradicionais, como SVM e Naive Bayes, que são utilizados para a Classificação de Sentimentos em nível de sentença e documento, são insuficientes (LIU, 2012). O principal motivo é que esses algoritmos não consideram um alvo de opinião (entidade e/ou aspecto) e, portanto, são incapazes de determinar a que se refere o sentimento classificado. Para resolver esse problema, é necessário adequar os algoritmos e/ou as representações dos dados para que sejam capazes de considerar um alvo opinião no aprendizado de um modelo.

Em Wei e Gulla (2010) é proposto um modelo de classificação hierárquica. No entanto, a questão-chave é a forma de determinar o escopo de cada expressão de sentimento, ou seja, se ela está relacionada ao aspecto de interesse na sentença. A principal abordagem atual é usar um parser sintático para determinar a dependência sintática e as demais informações pertinentes. Por exemplo, em Jiang *et al.* (2011), um analisador de dependência foi utilizado para gerar um conjunto de características de aspectos dependentes para a classificação. Uma abordagem semelhante também foi usada em Boiy e Moens (2009), em que os pesos de cada aspecto é baseado na posição da palavra em relação ao aspecto alvo na árvore sintática.

A principal vantagem da abordagem baseada no aprendizado de máquina é a possibilidade de construir um modelo automaticamente a partir de todos os tipos de atributos e recursos úteis para a classificação de sentimentos em relação aos aspectos. A maioria desses atributos são difíceis de serem utilizadas por um método baseado em léxico, porque os usuários não sabem como aplicá-las sem um algoritmo para descobrir essas regras.

A desvantagem dessa técnica é que o aprendizado supervisionado é dependente de dados para treinamento. Além disso, um modelo ou classificador treinado a partir de dados rotulados em um domínio, muitas vezes não tem bom desempenho em outro domínio. Embora o aprendizado por transferência (*transfer learning*) tem sido investigado em diversos trabalhos (AUE; GAMON, 2005; BLITZER; DREDZE; PEREIRA, 2007; TAN *et al.*, 2007) ainda não apresenta bons resultados para classificação de aspectos.

Os métodos atuais são usados principalmente para a classificação sentimento em nível do documento. Como documentos são longos e contêm muitos atributos para a classificação de sentenças individuais, a aprendizagem supervisionada apresenta dificuldade para dimensionar um grande número de domínios de aplicação.

2.3 Representação de Dados Textuais

Para aprender um modelo por meio de aprendizado de máquina é necessária uma representação estruturada da coleção de dados. O modelo espaço-vetorial (SALTON, 1989) é muito utilizado para representar coleções de textos por ser um modelo de representação simples e intuitivo (AGGARWAL; ZHAI, 2012). Nesse modelo, cada documento textual é denominado por um “exemplo”, fazendo uma analogia de que exemplos passados são uma experiência sobre o domínio do problema em que seja deseja “aprender”. A representação de dados textuais é a base para a aplicação de algoritmos de aprendizado de máquina. Para isso é necessário definir as representações e realizar o pré-processamento da coleção de documentos. No pré-processamento, um conjunto de técnicas são empregadas com o objetivo de selecionar um subconjunto conciso e representativo de termos.

Na etapa de pré-processamento são realizadas as seguintes tarefas: preparação dos documentos, extração de termos e seleção de atributos e geração de uma representação estruturada da coleção de documentos apropriada para os métodos de extração de padrões utilizados. Exemplos de técnicas empregadas na preparação de documentos, foram realizadas padronização, remoção das *stopwords*, radicalização dos termos (*stemming*) e seleção de atributos. Em relação à representação da coleção de dados textuais são utilizadas duas representações: (i) modelo espaço-vetorial, *bag-of-words* (*bow*) (TAN; STEINBACH; KUMAR, 2005), comumente utilizada e (ii) modelo de redes heterogêneas bipartidas.

2.3.1 Modelo Espaço-Vetorial

O modelo espaço-vetorial é uma matriz atributo-valor, em que linhas representam os exemplos (documentos, frases, termos, outros) e as colunas representam os atributos (palavras, termos, frases, conceitos, outros) da coleção de textos. Os atributos são obtidos após uma etapa de pré-processamento dos textos, que tem o objetivo de selecionar um subconjunto conciso e representativo de termos. Esse modelo tem como vantagem permitir a utilização de algoritmos

de classificação tradicionais que lidam com vetores numéricos. Na Tabela 1 está representado uma coleção contendo n documentos (exemplos), m atributos (termos) e uma coluna de “classe”, sendo essa utilizada na tarefa de classificação para representar a classe ou o rótulo do exemplo caso o mesmo seja rotulado.

Tabela 1 – Ilustração de uma matriz atributo-valor para representar um coleção com n exemplos (documentos) e m atributos (termos).

	t_1	t_2	$\dots$	t_{m-1}	t_m	Classe
d_1	w_{d_1,t_1}	w_{d_1,t_2}	$\dots$	$w_{d_1,t_{m-1}}$	w_{d_1,t_m}	C_{d_1}
d_2	w_{d_2,t_1}	w_{d_2,t_2}	$\dots$	$w_{d_2,t_{m-1}}$	w_{d_2,t_m}	C_{d_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
d_{n-1}	w_{d_{n-1},t_1}	w_{d_{n-1},t_2}	$\dots$	$w_{d_{n-1},t_{m-1}}$	w_{d_{n-1},t_m}	$C_{d_{n-1}}$
d_N	w_{d_N,t_1}	w_{d_N,t_2}	$\dots$	$w_{d_N,t_{m-1}}$	w_{d_N,t_m}	C_{d_N}

2.3.2 Modelo de Redes

Uma alternativa para representação dos dados textuais é o uso de redes, que vem ganhando atenção nas últimas décadas (SUN; HAN, 2012; SILVA; ZHAO, 2012; NEWMAN, 2010; AGGARWAL; WANG, 2010). A maior vantagem dessa representação é tornar explícito os diferentes tipos de relações entre os diferentes tipos de entidades dos textos (ROSSI *et al.*, 2014; ROSSI; LOPES; REZENDE, 2016; AGGARWAL; ZHAO, 2013; FERREIRA *et al.*, 2013; MISHRA *et al.*, 2012; JI *et al.*, 2010; SOLÉ *et al.*, 2010; GEE; COOK, 2005; HENSMAN, 2004; LU; GETOOR, 2003). Estudos recentes indicam que a classificação de textos por meio de redes alcança resultados promissores quando comparados com classificação de textos utilizando o modelo espaço-vetorial (ROSSI; REZENDE; LOPES, 2015; ROSSI; LOPES; REZENDE, 2016; BREVE *et al.*, 2012; BELKIN; MATVEEVA; NIYOGI, 2004; XUE *et al.*, 2004). Como dito anteriormente, este projeto tem como foco o aprendizado transdutivo que utiliza uma rede para representar os dados.

Formalmente, uma rede é definida pela tripla $\mathcal{N} = (\mathcal{O}, \mathcal{E}, \mathcal{W})$, em que $\mathcal{O}$ representa o conjunto de objetos, $\mathcal{E}$ o conjunto de conexões (também denominadas arestas ou relações) entre objetos, e $\mathcal{W}$ representa o conjunto de pesos das conexões. Se $\mathcal{O}$ contém um único tipo de objeto, esta rede é denominada rede homogênea. Quando $\mathcal{O}$ contém diferentes tipos de objetos, a rede é denominada de rede heterogênea (JI *et al.*, 2010). E quando $\mathcal{O}$ é composto por apenas dois tipos de objetos e $\mathcal{E}$ contém conexões que ocorrem somente entre objetos de tipos diferentes esta rede é denominada de rede heterogênea bipartida.

Em relação às redes heterogêneas, a topologia da rede é dependente do domínio e das informações disponíveis a respeito dos objetos.

Uma forma simples de representar uma coleção de textos é utilizar uma rede heterogênea bipartida. Para construir essa rede, não é necessário calcular semelhanças entre documentos, coocorrência de termos, nem informações explícitas sobre as relações entre os objetos. Nesta

rede, $\mathcal{O} = \mathcal{D} \cup \mathcal{T}$, em que $\mathcal{D} = d_1, d_2, \dots, d_m$ representa os objetos do tipo *documento* e $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ representa os objetos do tipo *termo*. $\mathcal{D}$ consiste de documentos rotulados ($\mathcal{D}_L$) e não rotulados ($\mathcal{D}_U$), ou seja, $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$. A conexão ocorre somente a partir de um objeto $d_i \in \mathcal{D}$ a um objeto $t_i \in \mathcal{T}$ e w_{d_i, t_j} é o peso entre o documento d_i e o termo t_j . O peso pode ser a frequência ou a ocorrência do termo t_i no documento d_i . $C = \{c1, c2, \dots, cl\}$ é o conjunto de classes da coleção de documentos.

Para cada objeto da rede o_i é atribuído um vetor $\mathbf{f}_i = \{f_1, f_2, \dots, f_{|C|}\}$, em que são armazenados os pesos de cada objeto o_i para a cada classe da coleção. O vetor de pesos define a relevância do objeto de discriminar cada classe. Portanto, esse vetor também é referenciado por vetor de informações da classe. A matriz $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{|\mathcal{O}|}\}^T$ armazena todos os vetores de pesos dos objetos de rede. O vetor $\mathbf{y}_i = \{y_1, y_2, \dots, y_{|C|}\}$ armazena as informações da classe real de um documento rotulado d_i , em que y tem um valor na posição correspondente à classe de documento i e 0 nas outras posições. A matriz $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{|\mathcal{D}_L|}\}$ representa informações de classe de todos os documentos rotulados. Na Figura 5 é apresentada a representação gráfica dessa rede heterogênea bipartida. É importante observar que a grande contribuição das redes heterogêneas é tornar explícitas as relações entre os objetos- algo dependente do domínio da aplicação. Assim, para obter resultados promissores é necessário um estudo sobre os tipos de dados, bem como o objetivo da classificação, para justificar o uso de uma determinada topologia.

Figura 5 – Exemplo de representação da coleção de documentos por uma rede heterogênea bipartida

Fonte: Adaptada de Rossi (2016).

2.4 Aprendizado de Máquina

O Aprendizado de Máquina (AM) tem por objetivo o desenvolvimento de técnicas computacionais para construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado pode ser definido como um programa de computador que toma decisões com base em experiências acumuladas (exemplos) por meio de solução bem sucedida de problemas anteriores. O modelo pode ser preditivo para fazer previsões futuras, ou descritiva para extrair conhecimento dos dados, ou ambos (MITCHELL, 1997; ALPAYDIN, 2010).

Segundo [Mitchell \(1997\)](#), aprendizado de máquina é definido como “um algoritmo que aprende a partir de experiências E com a relação a alguma classe de tarefas T e medida de desempenho P, se o seu desempenho em tarefas em T, medida pelo P, melhora com a experiência E”. Em outras palavras, algoritmos de aprendizado de máquina têm como objetivo aprender, generalizar ou encontrar padrões a partir de uma amostra representativa de exemplos (dados) sobre o problema.

Os métodos de aprendizado de máquina são muito utilizados nas tarefas de Mineração de Opiniões em nível de documento e em nível de sentença. A maioria das abordagens baseada em aprendizado de máquina já propostas utilizam técnicas supervisionadas ([LIU, 2012](#)). Porém, para uso desses métodos de Mineração de Opiniões baseada em Aspectos são necessárias abordagens que considerem os alvos de opinião (entidades ou aspectos) que variam em um único documento ou sentença.

Os métodos de aprendizado de máquina são organizados em três tipos: (i) aprendizado supervisionado; (ii) aprendizado não supervisionado; e (iii) semissupervisionado. Uma das principais diferenças entre esses três tipos é o conjunto de exemplos para treinamento. No aprendizado supervisionado, o conjunto de exemplos é significativamente grande e cada exemplo de treinamento apresenta a classe (rótulo) que este pertence. Novas ocorrências são classificadas com base no que foi aprendido no conjunto de treinamento. Já no aprendizado não supervisionado não existe rótulo pré-definido para nenhum dos exemplos do conjunto de treinamento. A vantagem desse tipo de aprendizado é não depender de informações rotuladas. No aprendizado supervisionado a acurácia, em geral, é maior, entretanto em um cenário real é difícil e custoso ter uma quantidade suficiente de exemplos rotulados para gerar um bom modelo.

O aprendizado semissupervisionado também considera um conjunto de exemplos rotulados, porém a quantidade de exemplos em que os rótulos são conhecidos é bem inferior que no aprendizado supervisionado e, neste tipo de aprendizado são necessários métodos específicos para tratar este cenário. O objetivo do aprendizado semissupervisionado é fazer uso dos exemplos não rotulados para melhorar o desempenho da classificação. O modo como os exemplos não rotulados são tratados no aprendizado semissupervisionado pode resultar em classificação com eficácia/performance superior a do aprendizado supervisionado considerando a mesma quantidade de exemplos rotulados, ou resultar em performance equivalente considerando uma menor quantidade de exemplos rotulados ([ZHU *et al.*, 2009](#)). Esse tipo de aprendizado de máquina têm despertado o interesse para uso na classificação automática de textos pois existe um grande volume de textos produzidos e disponibilizados atualmente, e que não são rotulados ([XU; KING, 2015](#); [SUBRAMANYA; TALUKDAR, 2014](#); [SØGAARD, 2013](#); [BREVE *et al.*, 2012](#); [CULP; MICHAILIDIS, 2008](#)). Devido ao fato que a existem poucos dados rotulados para Mineração de Opiniões baseada em aspectos. Nas seções próximas subseções são apresentados os principais conceitos e algoritmos de aprendizado de máquina utilizados neste trabalho.

2.4.1 *Aprendizado Supervisionado*

Os algoritmos de aprendizado supervisionado (TAN; STEINBACH; KUMAR, 2005) tem o objetivo de aprender uma função de mapeamento entre os exemplos e os rótulos. Esses algoritmos utilizam somente exemplos rotulados como conjunto de treinamento. O aprendizado supervisionado, dependendo do tipo de atributo alvo, pode ser do tipo classificação com valores discretos ou do tipo regressão, em que os valores possuem relação de ordem e são contínuos. Uma descrição resumida dos algoritmos de classificação utilizados no desenvolvimento desse projeto de doutorado são descritos a seguir. O conteúdo desta seção são baseados em Witten e Frank (2005).

2.4.1.1 *Naïve Bayes (NB)*

Este algoritmo é um classificador probabilístico simples que calcula um conjunto de probabilidades a partir do cálculo de frequência e combinações dado um conjunto de dados. O algoritmo utiliza o teorema de Bayes e assume que todos atributos são independentes dado o valor da classe. Dado um conjunto de dados rotulados, a etapa de treinamento consiste em calcular a probabilidade de ocorrência cada atributo em relação às classes. Na etapa de teste, o algoritmo calcula a probabilidade de cada atributo de um exemplo não visto ocorrer com base nas probabilidades obtida na etapa de treino. O exemplo não visto é classificado com a classe que possui maior probabilidade de ocorrer. Esse algoritmo é considerado ingênuo, uma vez que ele considera que todos os atributos são independentes durante o cálculo dessas probabilidades. Para dados textuais, geralmente se utiliza uma versão mais apropriada do algoritmo, denominado Multinomial Naïve Bayes. A principal diferença está na forma em que são computadas as probabilidades, que neste caso considera que os atributos seguem uma distribuição multinomial, que é o caso das palavras, nos textos.

2.4.1.2 *k-NN (k-Nearest Neighbor)*

Este algoritmo é baseado em distâncias (semelhanças) entre objetos, a classificação de um novo objeto é feita considerando os exemplos do conjunto de treinamento mais próximos a ele. Esse algoritmo classifica um novo exemplo por meio de um cálculo de similaridade com todos os exemplos previamente rotulados do conjunto de treinamento. Após essa comparação são selecionados os k vizinhos mais semelhantes ao exemplo não visto. Uma vez tendo selecionado os k vizinhos, o exemplo não visto receberá o rótulo com maior frequência dentre seus k vizinhos mais semelhantes — uma estratégia conhecida como voto majoritário. Também há a votação ponderada que considera a distância dos vizinhos.

2.4.1.3 *Árvore de Decisão*

Este algoritmo faz a divisão hierárquica dos atributos dos exemplos, em que os possíveis rótulos são as folhas e os atributos estão distribuídos em grau de importância nos outros nós

da árvore. Desse modo, o atributo com maior relevância para discriminar a classe está mais perto do topo. Isto é, durante a etapa de treinamento, o atributo no topo da árvore é o que possui menor entropia (maior ganho de informação) computada. Assim, conforme a noção de entropia do conjunto, um atributo que melhor organiza os exemplos em cada classe é identificado na etapa de treinamento. O conjunto de eventos é dividido em subconjuntos uma vez que o atributo é identificado e enquanto não ocorre um critério de parada o processo de identificar o melhor atributo para cada subconjunto prossegue.

O algoritmo C.45 (QUINLAN, 1993), este método é baseado em árvores de decisão que usa a estratégia de divisão e conquista para segmentar recursivamente o espaço de busca em subespaços, e cada subespaço é ajustado usando diferentes modelos. As árvores de decisões são simples de entender e interpretar. No entanto, podem criar árvores tendenciosas se algumas classes dominarem outras. O algoritmo J48 proposto por Quinlan (1993) é considerado o que apresenta o melhor resultado na construção de árvores de decisão, a partir de um conjunto de dados de treinamento, pois o mesmo se mostra adequado para os procedimentos, envolvendo as variáveis (dados) qualitativas contínuas e discretas presentes nas bases de dados.

2.4.1.4 SVM (*Support Vector Machines*)

Este algoritmo é baseado em aprendizado estatísticos desenvolvido por Vapnik (1995) que estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa capacidade de generalização. É um classificador que tem como objetivo prever corretamente a classe de novos dados construindo um hiperplano que permite separar diferentes classes de dados com uma margem máxima. Em função da maioria dos problemas práticos de classificação serem não lineares, o SVM utiliza uma técnica denominada *kernel trick* que gera a partir do espaço original dos dados um espaço novo de maior dimensionalidade, desse modo, possibilitando uma maior chance de que os exemplos sejam linearmente separáveis.

2.4.1.5 Redes MLP (*Multi-Layer Perceptron*)

Este algoritmo utiliza o perceptron, neurônio artificial, composto por um vetor de entrada, um bias, um vetor de pesos e uma função de ativação. Três camadas compõe o MLP. A primeira, a camada de entrada, recebe os atributos e valores dos exemplos que são estímulos do ambiente. Em seguida, os neurônios realizam a combinação linear dos dados da entrada nas camadas intermediárias. Dependendo da ativação de neurônio, o estímulo é propagado para as próximas camadas se houverem ou para a camada de saída que computa a resposta da rede. A identificação dos pesos dos neurônios para minimizar a diferença entre a saída da rede e a saída desejada é denominado aprendizado da rede. O algoritmo mais utilizado para essa tarefa é chamado de *Backpropagation* (RUMELHART; HINTON; WILLIAMS, 1986). Nesse algoritmo, que é repetido até uma condição de parada ou a rede convergir, a propagação do erro da camada de saída até a camada de entrada faz com que as camadas intermediárias atualizem os pesos das

conexões dos neurônios.

2.4.2 Aprendizado Semissupervisionado

Em relação ao aprendizado semissupervisionado, nesse paradigma são utilizados tanto exemplos não rotulados quanto exemplos rotulados durante a tarefa de aprendizado (ZHU *et al.*, 2009). O objetivo é aumentar o conjunto de exemplos disponíveis para treinamento, visando melhor discriminar os padrões (CHAPELLE; SCHÖLKOPF; ZIEN, 2006). Em geral, esse tipo de aprendizado é utilizado quando existem poucos exemplos rotulados, sendo útil em diversas aplicações práticas; uma vez que normalmente é fácil coletar exemplos, mas é custoso rotulá-los (devido à necessidade de um especialista do domínio e o tempo gasto para rotulação).

Existem dois tipos de aprendizado semissupervisionado: (i) transdutivo e (ii) indutivo. No aprendizado transdutivo, dada uma coleção de exemplos que se deseja classificar e poucos exemplos rotulados, o objetivo é classificar diretamente os exemplos não rotulados sem a necessidade de induzir um modelo de classificação (GAMMERMAN; VOVK; VAPNIK, 1998). Já no aprendizado indutivo semissupervisionado, os exemplos previamente rotulados e não rotulados são utilizados para induzir um modelo de classificação, que é utilizado para classificar novos exemplos (BLUM; MITCHELL, 1998). O processo de indução do modelo de classificação é dividido em duas partes. A primeira parte consiste em utilizar o aprendizado transdutivo para rotular os exemplos. A segunda parte consiste em utilizar os exemplos previamente rotulados e os rotulados através do aprendizado transdutivo para construir o modelo de classificação. Os algoritmos tradicionais para classificação semissupervisionada são o Self-Training, o Co-Training, o *Expectation Maximization* (EM) e o *Transductive Support Vector Machine* (TSVM), descritos a seguir

Self-Training (CULP; MICHAILIDIS, 2008; HAFFARI; SARKAR, 2007; YAROWSKY, 1995): esse algoritmo faz uso de um conjunto de exemplos rotulados (D_L) como conjunto de treinamento e um algoritmo supervisionado para classificar os eventos não rotulados EU. O classificador é construído com base nos exemplos do conjunto com exemplos rotulados (D_L) e usado para classificar o conjunto exemplos não rotulados. Após rotular todos os exemplos do conjunto EU, o algoritmo seleciona os eventos do conjunto EU que foram rotulados com maior taxa de confiança (dado um limiar) e adiciona esses eventos ao conjunto EL, construindo-se o classificador indutivo novamente. O processo é repetido até que todos os eventos do conjunto EU tenham sido adicionados no conjunto EL ou outra condição de parada.

Co-Training (BLUM; MITCHELL, 1998; ZHU; GOLDBERG, 2009): tem como base a ideia de que o mesmo conjunto de exemplos é representado por duas visões V1 e V2, ou seja, dois diferentes tipos de atributos. São utilizados dois algoritmos supervisionados, um para cada visão, para construir os modelos de classificação a partir do conjunto de dados

rotulados. Após a etapa inicial treinamento, os eventos classificados com maior confiança na visão V1 são adicionados no conjunto de treino na visão V2; e vice-versa. Assim, um classificador complementa a visão de outro classificador. Esse processo se repete até que não se tenha mais dados rotulados ou outra condição de parada .

Expectation Maximization (EM) (DEMPSTER; LAIRD; RUBIN, 1977): este algoritmo permite atribuir informação de classe aos documentos não rotulados iterativamente, sem a necessidade de induzir modelos de classificação. No caso da classificação de textos, o objetivo é definir os componentes de misturas que geram os textos da coleção. Cada componente está associado a uma classe e o objetivo é definir as probabilidades dos termos para cada componente utilizando tanto documentos rotulados quanto não rotulados. No contexto desta tese, os componentes correspondem às informações de classe dos termos para as classes. Em Nigam e Ghani (2000) é apresentada uma abordagem que combina EM com a técnica Multinomial Naïve Bayes para realizar a classificação transdutiva de textos. Nesta abordagem, a probabilidade de ocorrer uma classe dado um documento, ou a informação de classe do documento no contexto desta tese, é dada de acordo com teorema de Bayes. O cálculo da informação de classe dos documentos não rotulados (E-step) e o cálculo da informação de classe dos termos (M-step). E-step e M-step são executados iterativamente até a convergência, isto é, até que não haja alterações significativas nas informações de classe dos documentos não rotulados ou até um determinado número fixo de iterações. O algoritmo EM executa uma busca do tipo *Hill-Climbing* no espaço de hipóteses, garantindo que os valores das informações de classe dos termos convirjam.

Transductive Support Vector Machine (TSVM) (VAPNIK; VASHIST, 2009; JOACHIMS, 1999): é uma extensão de algoritmo SVM que permite considerar os exemplos não rotulados durante a etapa de treinamento. O objetivo é que os exemplos não rotulados auxiliem na identificação de regiões de alta densidade nos dados. Essas regiões de alta de densidade, junto com as informações de exemplos rotulados, permitem obter hiperplanos que identifiquem uma melhor margem de separação dessas regiões. A principal estratégia para esse treinamento consiste em duas etapas. Na primeira etapa (etapa indutiva), são utilizados apenas os exemplos rotulados para obtenção do hiperplano. Em seguida, os exemplos não rotulados são classificados considerando o hiperplano obtido. Na segunda etapa (etapa transdutiva), todos os exemplos (rotulados originalmente e rotulados pelo classificador) são utilizados para obter um novo hiperplano de separação. Caso algum evento tenha sido classificado erroneamente (considerando os rótulos da etapa anterior), então o exemplo é alocado para a outra classe. Esse processo continua até a convergência ou outra condição de parada.

Além disso, independente do tipo de aprendizado a ser realizado, deve ser definido a forma de representação da coleção de textos para que esta seja compreendida pelos algoritmos

de aprendizado de máquina. Existem duas representações que estão sendo utilizadas: (i) modelo espaço-vetorial (amplamente utilizada) e (ii) redes.

Em um modelo espaço-vetorial cada documento é representado por um vetor e cada posição desse vetor corresponde a uma dimensão (atributo ou termo) da coleção de documentos (SALTON, 1989). Essa representação tem como principais características a alta dimensionalidade, na qual algumas centenas de documentos podem gerar milhares de dimensões, e a alta esparsidade, na qual cada documento contém poucos termos frente a todos os termos da coleção, e portanto, a maioria dos valores das dimensões é 0.

As representações em redes permitem modelar relações entre entidades de um problema e em um formalismo matematicamente tratável (BLANCO; LIOMA, 2012). Dessa forma, esta representação facilita a extração de padrões de classes que dificilmente são capturados por algoritmos baseados no modelo espaço-vetorial. Além disso, melhoram o desempenho de classificação dos algoritmos de aprendizado de máquina (BREVE *et al.*, 2012). Esse tipo de representação tem sido uma alternativa para representar conjuntos de dados e textos e que atualmente vem sendo muito empregado na área de Aprendizado de Máquina (SUN; HAN, 2012; HAN, 2012; ANGELOVA; KASNECI; WEIKUM, 2012; NEWMAN, 2010; AGGARWAL; WANG, 2010).

Neste projeto de doutorado o foco está nos métodos de aprendizado semissupervisionado transdutivo baseados em redes. Nas seções seguintes são apresentadas as terminologias utilizadas para aprendizado semissupervisionado transdutivo e os principais algoritmos do estado-da-arte.

2.4.3 *Aprendizado Semissupervisionado Transdutivo em Redes*

O aprendizado transdutivo apresenta a vantagem de utilizar exemplos não rotulados para melhorar o processo de aprendizado. Esta característica é muito útil no escopo deste trabalho, já que há uma grande quantidade de textos, como notícias e boletins informativos, publicados diariamente na web. Para exemplificar, em algumas situações os algoritmos de aprendizado transdutivo utilizando somente 10% de exemplos rotulados conseguem ter uma performance de classificação comparável ao algoritmos de aprendizado indutivo supervisionado (ROSSI *et al.*, 2014). Na Figura 6 existem três textos, mas apenas o texto “Ensino e educação” possui o rótulo “Educação”. Nessa situação, o texto “Matemática para crianças” não pode ser rotulado, pois, não possui termos em comum com o documento rotulado. No entanto, ao utilizar textos não rotulados é possível encontrar termos comuns entre os três textos e pode-se atribuir o rótulo “Educação” a todos eles. Assim, a utilização de exemplos não rotulados permite descobrir novos padrões que não seriam descobertos considerando exemplos rotulados.

Figura 6 – Exemplo de como a utilização de exemplos não rotulados auxilia na classificação

Fonte: Rossi (2016).

Inicialmente, os algoritmos de aprendizado transdutivo foram propostos usando o modelo espaço-vetorial. Entretanto, estudos anteriores reportaram que não foram obtidos resultados satisfatórios com esse tipo de representação (ZHU *et al.*, 2009; COZMAN; COHEN; CIRELO, 2003; ELWORTHY, 1994). Por outro lado, estudos recentes têm apresentado resultados promissores de classificação transdutiva por meio de redes (ROSSI; LOPES; REZENDE, 2016; ROSSI; LOPES; REZENDE, 2014). Tal resultado é um aspecto motivador para investigar algoritmos de classificação transdutiva em redes na área de Mineração de Opiniões baseada em Aspectos.

Ao utilizar a modelagem de rede na tarefa de classificação, cada exemplo do conjunto de objetos $\mathcal{O}$ de uma rede possui um vetor de informação de classe $\mathbf{f}$. Nesse vetor existe uma posição para cada possível rótulo que armazena o nível de aderência do exemplo ao rótulo. Caso seja um exemplo rotulado, ele possuirá um vetor $\mathbf{y}$ de mesmas dimensões de $\mathbf{f}$, mas com o objetivo de armazenar o(s) rótulo(s) original(is) do exemplo.

A classificação transdutiva utilizando redes visa classificar objetos não rotulados considerando informações provindas dos objetos vizinhos, como os atributos ou informações de classe. Diversos trabalhos da literatura demonstraram que a utilização das informações de classes dos objetos vizinhos são suficientes para obter uma acurácia de classificação tão boa quanto utilização dos atributos em conjunto com as informações de classe (JENSEN; NEVILLE; GALLAGHER, 2004; OH; MYAENG; LEE, 2000; CHAKRABARTI; DOM; INDYK, 1998), além de diminuir o custo computacional da classificação.

Formalmente, o objetivo da classificação transdutiva é gerar uma função $F : \mathcal{O}^{L+U} \rightarrow \mathcal{Y}^{L+U}$ tal que F seja um bom preditor sobre os dados não rotulados. Tanto os exemplos rotulados quanto não rotulados são considerados para gerar a função F .

A rede que representa o conjunto de dados é composta por objetos rotulados e não rotulados. No caso do tipo de conjunto de dados tratado neste projeto de doutorado são as coleções de documentos (textos), que usualmente possui alguns documentos como objetos rotulados. Os demais termos da rede, outros documentos e termos não possuem rótulos. Os

objetos que se deseja classificar são denominados como “*alvos*” e os demais tipos de objetos como “*pontes*”.

Os algoritmos semissupervisionados transdutivos em redes tem como base um *framework* de regularização, em que o objetivo é minimizar uma função de custo satisfazendo duas premissas: (i) as informações de classe de objetos vizinhos (conectados) devem ser semelhantes e (ii) as informações de classe dos objetos rotulados atribuídas durante o processo de classificação devem ser semelhantes às informações de classe reais (ZHU, 2005). O *framework* está expresso de forma genérica na seguinte função (DELALLEAU; BENGIO; ROUX, 2005):

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{o_i, o_j \in \mathcal{O}} w_{o_i, o_j} \Omega(\mathbf{f}_{o_i}, \mathbf{f}_{o_j}) + \mu \sum_{o_i \in \mathcal{O}^L} \Omega'(\mathbf{f}_{o_i}, \mathbf{y}_{o_i}) \quad (2.1)$$

na qual μ é o parâmetro de regularização que define a importância que dada a cada uma das premissas. Os algoritmos baseados em regularização se diferenciam nas instanciações da função de similaridade e dos valores de μ . Independente da função de similaridade adotada, a ideia é que se w_{o_i, o_j} é grande, as informações de classe f_{o_i}, f_{o_j} são próximas.

O primeiro termo $\Omega(\cdot)$ da função de regularização é responsável por computar a proximidade entre os vetores de informação de classe entre cada par de objetos relacionados na rede. Tal proximidade pode ser derivada de funções de distâncias ou de dissimilaridades. Já o segundo termo $\Omega(\cdot)'$ é responsável por computar a proximidade entre a informação de classe de objetos rotulados e as respectivas informações reais de classe. Ainda nesta equação, w_{o_i, o_j} indica o peso da relação entre os objetos e o parâmetro μ indica a importância da informação de classe durante a propagação dos rótulos.

A Equação 2.1 pode ser resolvida através de soluções iterativas, denominadas “propagação de rótulos” (WANG; ZHANG, 2008). Nesse caso, gradualmente os objetos propagam sua informação de classe para os objetos vizinhos sendo essa propagação ponderada pelo peso da sua relação com outros objetos.

Diversos métodos com diferentes características vêm sendo proposto ao longo dos anos para a classificação transdutiva. A seguir são apresentados os principais algoritmos para classificação transdutiva usando redes homogêneas e heterogêneas que são utilizados neste projeto de doutorado. Os principais algoritmos de classificação homogêneas são: (i) *Gaussian Fields and Harmonic Functions* (GFHF) e (ii) *Learning with Local and Global Consistency* (LLGC). respectivamente. Os principais algoritmos de de classificação transdutiva em redes heterogêneas são: (i) *Tag-based Model*; (ii) *GNetMine*; (iii) *Label Propagation using Bipartite Heterogeneous Networks* (LPHN) e (iv) TCHN . Esses algoritmos são descritos nas próximas subseções.

2.4.3.1 Gaussian Fields and Harmonic Functions (GFHF)

Um dos principais algoritmos para classificação transdutiva em redes homogêneas é o GFHF (*Gaussian Fields and Harmonic Functions*) (ZHU; GHAHRAMANI; LAFFERTY, 2003). Sua principal característica é computar a informação de classe de um objeto não rotulado com base na média de informação de classe dos objetos vizinhos ponderada pelos pesos das respectivas conexões. Assim, dado um determinado objeto o_i , seu vetor de informação de classe é computado pela Equação:

$$f(o_i) = \frac{\sum_{o_j \in \mathcal{O}} w_{o_i, o_j} \cdot f_{o_j}}{\sum_{o_j \in \mathcal{O}} w_{o_i, o_j}}. \quad (2.2)$$

Outra característica importante do GFHF é que a informação de classe de objetos rotulados não deve ser alterada durante a propagação. Para satisfazer essa restrição, o valor de μ no segundo termo da função de regularização é definido como $\lim_{\mu \rightarrow \infty} \mu$. Na Equação 2.3 é apresentada a função de regularização completa do algoritmo GFHF.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{o_i, o_j \in \mathcal{O}} w_{o_i, o_j} (\mathbf{f}_{o_i} - \mathbf{f}_{o_j})^2 + \lim_{\mu \rightarrow \infty} \mu \sum_{o_i \in \mathcal{O}^L} (\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2 \quad (2.3)$$

No Algoritmo 1 é apresentado o pseudocódigo do GFHF. A entrada do algoritmo é o conjunto de objetos da rede $\mathcal{O}$, as informações de classe $\mathbf{Y}$, os pesos de conexões entre objetos $\mathbf{W}$ e o grau dos objetos $\mathbf{D}$. Na linha 2 é definida uma matriz $\mathbf{P}$ de dimensão $|\mathcal{O}| \times |\mathcal{O}|$. Uma célula $\mathbf{P}_{o_i, o_j}$ possui a probabilidade de ocorrer uma conexão entre o objeto o_i com o objeto o_j . Na linha 4 é feita a propagação de rótulos propriamente dita. Na linha 5, a informação de rótulo dos objetos previamente rotulados $\mathcal{O}^L$ são sobrescritos pelos valores originais armazenados em $\mathbf{Y}$. O algoritmo apresenta como saída a informação de classe dos objetos não rotulados $\mathbf{F}(\mathcal{O}^U)$.

Algoritmo 1 – Gaussian Fields and Harmonic Functions (GFHF)

Entrada: $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

1 início
2    $\mathbf{P} \leftarrow \mathbf{D}^{-1} \cdot \mathbf{W}$ 
3   repita
4      $\mathbf{F}(\mathcal{O}) \leftarrow \mathbf{P} \cdot \mathbf{F}(\mathcal{O})$ 
5      $\mathbf{F}(\mathcal{O}^L) \leftarrow \mathbf{Y}(\mathcal{O}^L)$ 
6   até convergência ou número fixo de iterações
7 fim
8 retorna  $\mathbf{F}(\mathcal{O}^U)$ 

```

2.4.3.2 Learning with Local and Global Consistency (LLGC)

Uma adaptação do algoritmo GFHF foi proposta por [Zhou et al. \(2004\)](#), denominado LLGC (*Learning With Local and Global Consistency*). No LLGC, é assumido que alguns exemplos podem ser erroneamente rotulados pelo usuário, causando com isso uma deterioração da performance de classificação. Dessa forma o algoritmo permite que os exemplos rotulados possam ter seus rótulos alterados durante a classificação. Na Equação 2.4 está representada a função de regularização do LLGC. Observe que no segundo termo da função de regularização, o parâmetro μ permite definir o grau de importância de objetos inicialmente rotulados. Outra característica importante do LLGC é a redução da influência de objetos com alto grau durante a propagação de rótulos. Essa característica é expressa na normalização realizada no primeiro termo do regularizador.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{\mathcal{O}_i, \mathcal{O}_j \in \mathcal{O}} w_{\mathcal{O}_i, \mathcal{O}_j} \left\| \frac{\mathbf{f}_{\mathcal{O}_i}}{\sqrt{\sum_{\mathcal{O}_k \in \mathcal{O}} w_{\mathcal{O}_i, \mathcal{O}_k}}} - \frac{\mathbf{f}_{\mathcal{O}_j}}{\sqrt{\sum_{\mathcal{O}_k \in \mathcal{O}} w_{\mathcal{O}_j, \mathcal{O}_k}}} \right\|^2 + \mu \sum_{\mathcal{O}_i \in \mathcal{O}^L} \|\mathbf{f}_{\mathcal{O}_i} - \mathbf{y}_{\mathcal{O}_i}\|^2 \quad (2.4)$$

No Algoritmo 2 é apresentado o pseudocódigo do LLGC. O algoritmo recebe como entrada o conjunto de objetos $\mathcal{O}$, a matriz de pesos $\mathbf{W}$, a informação de classe dos objetos rotulados $\mathbf{Y}$, uma matriz $\mathbf{D}$ com o grau dos objetos e o valor de μ que deve estar no intervalo $0 < \mu \leq 1$. A matriz $\mathbf{S}$ computada na Linha 10 representa uma função análoga à matriz $\mathbf{P}$ do GFHF. Porém, no seu cálculo é reduzido a influência da matriz de grau $\mathbf{D}$ na associação entre dois objetos da rede. A propagação de rótulos propriamente dita é realizada na linha 12, com o diferencial de que o peso de informação de classe de objetos previamente rotulados pelo usuário pode ser ajustado conforme o valor de μ .

Algoritmo 2 – Learning with Local and Global Consistency (LLGC)

Entrada: $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$, μ

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

9  início
10   $\mathbf{S} \leftarrow \mathbf{D}^{-(1/2)} \cdot \mathbf{W} \cdot \mathbf{D}^{-(1/2)}$ 
11  repita
12  |  $\mathbf{F} \leftarrow \mu \cdot \mathbf{S} \cdot \mathbf{F} + (1 - \mu) \cdot \mathbf{Y}$ 
13  até convergência ou número fixo de iterações
14  fim
15  retorna  $\mathbf{F}(\mathcal{O}^U)$ 

```

2.4.3.3 Label Propagation using Bipartite Heterogeneous Networks

O algoritmo LPHN (Label Propagation through Heterogeneous Networks (LPHN)) proposto por [Rossi, Lopes e Rezende \(2014\)](#) é uma extensão do algoritmo GFHF para redes heterogêneas. Assim, é mantida a propriedade de que objetos rotulados pelos usuários não devem

ter seus rótulos alterados. A função de regularização do LPHN é descrita na Equação 3.1 que usa uma regularização análoga a do GFHF (Equação 2.3), com a restrição de que a propagação considera diferentes tipos de relação no primeiro termo do regularizador. Assim, o LPHN seleciona dois conjuntos de objetos $\mathcal{O}_j$ e $\mathcal{O}_l$ e propaga os rótulos do conjunto $\mathcal{O}_j$ para o conjunto $\mathcal{O}_l$ utilizando a mesma estratégia do GFHF.

$$Q(\mathbf{F}) = \sum_{\mathcal{O}_j, \mathcal{O}_l \subset \mathcal{O}} \frac{1}{2} \sum_{o_i \in \mathcal{O}_j} \sum_{o_k \in \mathcal{O}_l} w_{o_i, o_k} (\mathbf{f}_{o_i} - \mathbf{f}_{o_k})^2 + \lim_{\mu \rightarrow \infty} \mu \sum_{o_i \in \mathcal{O}^L} (\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2 \quad (2.5)$$

No Algoritmo 3 é apresentado o pseudocódigo do LPHN. A entrada do algoritmo é o conjunto de objetos da rede $\mathcal{O}$, as informações de classe $\mathbf{Y}$, os pesos de conexões entre objetos $\mathbf{W}$ e o grau dos objetos $\mathbf{D}$. Na linha 2 é definida uma matriz $\mathbf{P}$ que possui dimensão $|\mathcal{O}| \times |\mathcal{O}|$. Uma célula $\mathbf{P}_{o_i, o_j}$ possui a probabilidade de ocorrer uma conexão entre o objeto o_i com o objeto o_j . Na linha 5 é feita a propagação de rótulos entre os objetos de $\mathcal{O}_i$ e $\mathcal{O}_j$. A informação de rótulo dos objetos previamente rotulados é sempre mantida (linhas 5 – 6). O algoritmo retorna a informação de classe dos objetos não rotulados $\mathbf{F}(\mathcal{O}^U)$.

Algoritmo 3 – Label Propagation through Heterogeneous Networks (LPHN)

Entrada: $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

1  início
2       $\mathbf{P} = \mathbf{D}^{-1} \cdot \mathbf{W}$ 
3      repita
4          para cada  $\mathcal{O}_j, \mathcal{O}_l \subset \mathcal{O}$  (subconjuntos de objetos  $i$  e  $j$ ) faça
5               $\mathbf{F}(\mathcal{O}_i, \mathcal{O}_j) \leftarrow \mathbf{P}(\mathcal{O}_i, \mathcal{O}_j) \cdot \mathbf{F}(\mathcal{O}_i, \mathcal{O}_j)$ 
6               $\mathbf{F}(\mathcal{O}_i^L) \leftarrow \mathbf{Y}(\mathcal{O}_i^L)$ 
7               $\mathbf{F}(\mathcal{O}_j^L) \leftarrow \mathbf{Y}(\mathcal{O}_j^L)$ 
8          fim
9      até convergência ou número fixo de iterações
10 fim
11 retorna  $\mathbf{F}(\mathcal{O}^U)$ 

```

2.4.3.4 GNetMine

O GNetMine (GNM) proposto por Ji *et al.* (2010) é um algoritmo para propagação de rótulos em redes heterogêneas que considera a importância das relações, pois considera diferentes semânticas entre os objetos de diferentes tipos. Além disso, permiti reduzir a importância de objetos inicialmente rotulados. A função de regularização do GNM é descrita na Equação 3.2. Nesta função, $\lambda_{\mathcal{O}_i, \mathcal{O}_j}$ é utilizado para definir a importância da relação entre um objeto do tipo $\mathcal{O}_i$ e um objeto do tipo $\mathcal{O}_j$. α_{o_j} é utilizado para definir a importância dada para as informações de

classe real, ou seja, de um objeto inicialmente rotulado por o_j . Ambos os parâmetros possuem valores no intervalo de entre $[0; 1]$, $0 \leq \lambda_{\mathcal{O}_i, \mathcal{O}_j} \leq 1$ e $0 \leq \alpha_{o_j} \leq 1$.

$$Q(\mathbf{F}) = \sum_{\mathcal{O}_i, \mathcal{O}_j \in \mathcal{O}} \lambda_{\mathcal{O}_i, \mathcal{O}_j} \sum_{o_k \in \mathcal{O}_i} \sum_{o_l \in \mathcal{O}_j} w_{o_k, o_l} \left\| \frac{\mathbf{f}_{o_k}(\mathcal{O}_i)}{\sqrt{\sum_{o_m \in \mathcal{O}_j} w_{o_k, o_m}}} - \frac{\mathbf{f}_{o_l}(\mathcal{O}_j)}{\sqrt{\sum_{o_m \in \mathcal{O}_i} w_{o_l, o_m}}} \right\|^2 + \sum_{o_j \in \mathcal{O}^L} \alpha_{o_j} (\mathbf{f}_{o_j} - \mathbf{y}_{o_j}) \quad (2.6)$$

Algoritmo 4 – GNetMine (GNM)

Entrada: $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$, λ , α

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

1  início
2   $\mathbf{F}^{(0)} \leftarrow \mathbf{Y}$ 
3  para cada  $\mathcal{O}_i, \mathcal{O}_j \subset \mathcal{O}$  faça
4       $S_{\mathcal{O}_i, \mathcal{O}_j} = \mathbf{D}_{\mathcal{O}_i, \mathcal{O}_j}^{-1/2} \cdot \mathbf{W}_{\mathcal{O}_i, \mathcal{O}_j} \cdot \mathbf{D}_{\mathcal{O}_j, \mathcal{O}_i}^{-1/2}$ 
5  fim
6  repita
7      para cada  $\mathcal{O}_i \subset \mathcal{O}$  faça
8          para cada  $o_j \in \mathcal{O}_i$  faça
9               $\mathbf{f}_{o_j} \leftarrow \sum_{\mathcal{O}_k \subset \mathcal{O}, o_j \notin \mathcal{O}_k} \lambda_{\mathcal{O}_i, \mathcal{O}_k} \cdot \sum_{o_m \in \mathcal{O}_k} S(\mathcal{O}_i, \mathcal{O}_k)_{o_j, o_m} \cdot \mathbf{f}_{o_m}$ 
10              $\mathbf{f}_{o_j} \leftarrow \mathbf{f}_{o_j} + (2 \cdot \lambda_{\mathcal{O}_i, \mathcal{O}_i} \cdot \sum_{o_m \in \mathcal{O}_i} S(\mathcal{O}_i, \mathcal{O}_i)_{o_j, o_m} \cdot \mathbf{f}_{o_m}) + \alpha_{o_j} \cdot \mathbf{y}_{o_j}$ 
11              $\mathbf{f}_{o_j} \leftarrow \frac{\mathbf{f}_{o_j}}{\sum_{\mathcal{O}_k \subset \mathcal{O}, o_j \notin \mathcal{O}_k} \lambda_{\mathcal{O}_i, \mathcal{O}_k} + 2 \cdot \lambda_{\mathcal{O}_i, \mathcal{O}_i} + \alpha_{o_j}}$ 
12         fim
13     fim
14  até convergência ou número fixo de iterações
15 fim
```

No Algoritmo 4 é apresentado o pseudocódigo do GNM. A entrada do algoritmo é o conjunto de objetos da rede $\mathcal{O}$, as informações de classe $\mathbf{Y}$, os pesos de conexões entre objetos $\mathbf{W}$, grau dos objetos $\mathbf{D}$, a importância α das informações previamente rotuladas e um vetor λ com a importância de cada tipo de relação. Na linha 4 é definido um valor de confiança da conexão entre dois objetos na rede, dada dois tipos de relações. A propagação de rótulos é realizada nas linhas 9–11, em que é realizada a atualização de informação de classe de um objeto $o_j \in \mathcal{O}_i$ conforme os objetos vizinhos, a importância da relação dos objetos vizinhos e a importância

da informação rotulada. Na prática, o GNM pode ser visto como uma extensão do LLGC para redes heterogêneas, uma vez que o primeiro termo do regularizador também ameniza o efeito de objetos com um alto valor de grau. Apesar da possibilidade de definir a importância das relações entre os objetos, esse valor de importância deve ser definido previamente pelo usuário ou por algum processo de calibragem.

2.4.3.5 Tag-based Model

O *Tag-based Model* (TM) proposto por Ji *et al.* (2010) é um algoritmo para propagação de rótulos em redes heterogêneas que permite classificar diferentes tipos de objetos conectados todos a um outro tipo de objeto. Além disso esse algoritmo permite adicionar objetos de outros domínios que possam auxiliar na classificação de objetos. Portanto, neste algoritmo, além dos tipos de objetos considerados anteriormente ($\mathcal{O}_i$ e $\mathcal{O}_j$), também é utilizado um conjunto de objetos do tipo auxiliar ($\mathcal{A}$).

O algoritmo considera (i) a informação de classe inferidas inicialmente e as informações de classe atribuídas aos objetos auxiliares; (ii) as informações de classe reais e as informações atribuídas aos documentos rotulados; (iii) as informações de classe inferidas inicialmente e as informações de classe atribuídas aos objetos não rotulados; e, (iv) as informações de classe dos objetos em $\mathcal{O}_i$ e $\mathcal{A}$ com as informações de classe dos objetos em $\mathcal{O}_j$. A função de regularização do TM é descrita na Equação 3.3. Nesta função, α define a importância das informações de classe real dos objetos auxiliares, β dos objetos inicialmente rotulados o_i e γ dos objetos inicialmente sem rótulos.

$$Q(\mathbf{F}) = \alpha \sum_{a_i \in \mathcal{A}} \|\mathbf{f}_{a_i} - \mathbf{y}_{a_i}\|^2 + \left(\beta \sum_{o_i \in \mathcal{O}^L} \|\mathbf{f}_{o_i} - \mathbf{y}_{o_i}\|^2 + \gamma \sum_{o_i \in \mathcal{O}^U} \|\mathbf{f}_{o_i} - \mathbf{y}_{o_i}\|^2 \right) \quad (2.7)$$

$$+ \sum_{o_i \in \mathcal{O}_i \cup \mathcal{A}} \sum_{o_j \in \mathcal{O}_j} w_{o_i, o_j} \|\mathbf{f}_{o_i} - \mathbf{f}_{o_j}\|^2$$

No Algoritmo 5 é apresentado o pseudocódigo do *Tag-based-Model*. A entrada do algoritmo é o conjunto de objetos da rede $\mathcal{O}$, as informações de classe $\mathbf{Y}$, os pesos de conexões entre objetos $\mathbf{W}$, grau dos objetos $\mathbf{D}$, o peso da importância das informações dos objetos auxiliares α , o peso dos objetos previamente rotulados β e o peso dos objetos previamente sem rótulos γ , um conjunto objetos do tipo auxiliar $\mathcal{A}$ e conjunto objetos conectam aos demais tipos de objetos $\mathcal{P}$. Vale ressaltar que o uso de objetos auxiliares é opcional. Na linha 5 é definido um valor influência em relação aos objetos auxiliares. Na linha 8 é definido um valor influência em relação aos objetos já rotulados. A propagação de rótulos é realizada nas linhas 11–14, em que é realizada a atualização de informação de classe de um objeto $o_j \in \mathcal{O}_j$ conforme os objetos vizinhos, a importância da relação dos objetos vizinhos e a importância da informação rotulada e da informação auxiliar.

Algoritmo 5 – Tag-based Model (TM)**Entrada:** $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$, $\mathcal{A}$ ($\mathcal{A} \subset \mathcal{O}$) $\mathcal{P}$ ($\mathcal{P} \subset \mathcal{O}$), α , β , γ **Saída:** $\mathbf{F}(\mathcal{O}^U)$

```

1 início
2   repita
3     para cada  $a_i \in \mathcal{A}$  faça
4        $\mathbf{f}_{a_i}(\mathcal{A}) \leftarrow \alpha \cdot \mathbf{f}_{a_i} / (\alpha + \sum_{t_j \in \mathcal{T}} w_{a_i, t_j}) + \sum_{t_j \in \mathcal{T}} w_{a_i, t_j} \mathbf{f}_j(\mathcal{T}) / (\alpha + \sum_{t_j \in \mathcal{T}} w_{a_i, t_j})$ 
5     fim
6     para cada  $d_i \in \mathcal{D}^L$  faça
7        $\mathbf{f}_{d_i}(\mathcal{D}^L) \leftarrow \beta \mathbf{f}_{d_i} / (\beta + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j}) + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j} \mathbf{f}_j(\mathcal{T}) / (\beta + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j})$ 
8     fim
9     para cada  $d_i \in \mathcal{D}^U$  faça
10       $\mathbf{f}_{d_i}(\mathcal{D}^U) \leftarrow \gamma \mathbf{f}_{d_i} / (\gamma + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j}) + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j} \mathbf{f}_{t_j}(\mathcal{T}) / (\gamma + \sum_{t_j \in \mathcal{T}} w_{d_i, t_j})$ 
11    fim
12    para cada  $t_i \in \mathcal{T}$  faça
13       $\mathbf{f}_i(\mathcal{T}) \leftarrow \sum_{o_k \in \mathcal{A} \cup \mathcal{D}} w_{o_k, t_i} \mathbf{f}_{o_k} / \sum_{o_j \in \mathcal{A} \cup \mathcal{D}} w_{o_j, t_i}$ 
14    fim
15  até convergência ou número fixo de iteracoes
16 fim

```

2.4.3.6 Transductive Categorization based on Heterogeneous Networks

O *Transductive Categorization based on Heterogeneous Networks* (TCHN) proposto por [Rossi, Lopes e Rezende \(2016\)](#) é um algoritmo para propagação de rótulos em redes heterogêneas que tem por premissa que (i) as informações de classe dos objetos em $\mathcal{O}^L$ e em $\mathcal{O}^U$ são úteis para definir as informações de classe dos "termos" e (ii) as informações de classe inferidas para os "termos" são úteis para inferir as informações de classe dos "documentos" não rotulados.

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2} \left(\sum_{c_k \in \mathcal{C}} \left(\sum_{t_i \in \mathcal{T}^U} f_{t_i, c_k} - \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} \cdot f_{b_j, c_k} \right) \right)^2 + \frac{1}{2} \left(\sum_{c_k \in \mathcal{C}} \left(\sum_{t_i \in \mathcal{T}^L} y_{t_i, c_k} - \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} \cdot f_{b_j, c_k} \right) \right)^2 \quad (2.8)$$

O objetivo do TCHN é minimizar: (i) as diferenças entre as **informações de classe dos termos** e as informações de classe dos **objetos rotulados**; (ii) as diferenças entre as **informações de classe dos diferentes objetos** e as informações de classe atribuídas aos **objetos não rotulados**; (iii) a diferença entre as informações de classe entre os **objetos-t relacionados**; e (iv) as diferenças entre as informações de classe dos **objetos relacionados**. É importante ressaltar que

quanto maior o peso das relações, menor a diferença entre as informações de classe dos objetos relacionados. Portanto o algoritmo TCHN minimiza a função apresentada na Equação 2.8.

Para definir as informações de classe dos documentos não rotulados, o TCHN realiza dois processos: um para otimização e outro para propagação de rótulos. No Algoritmo 6 é apresentado o pseudocódigo do TCHN. Nas linhas 20—30 são induzidas as informações de classe iniciais para os termos considerando apenas documentos rotulados. Na sequência são realizadas as seguintes iterações locais:

- (i) otimização das informações de classe dos termos considerando as informações de classe dos documentos não rotulados (linhas 17–27);
- (ii) minimização das informações de classe entre termos relacionados (linhas 28–37);
- (iii) propagação das informações de classe dos termos para definir as informações de classe dos documentos não rotulados (38–41);
- (iv) otimização das informações de classe dos documentos não rotulados considerando as relações entre os documentos (linhas 42–50); e,
- (v) otimização das informações de classe dos termos considerando as informações de classe dos documentos rotulados (linhas 55–65).

Algoritmo 6 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 1)

Entrada: $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$, η

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

1  início
2      /* First Step                                     */
3      repita
4          para cada  $d_i \in \mathcal{D}^L$  faça
5              InfClasses  $\leftarrow$  Classificar_documento( $d_i, \mathbf{F}(\mathcal{T})$ )
6              para cada  $c_j \in \mathcal{C}$  faça
7                  error =  $y_{d_i, c_j} - \text{InfClasses}[c_j]$ 
8                  para cada  $t_k \in T$  faça
9                       $f_{t_k, c_j}(\mathcal{T}) \leftarrow f_{t_k, c_j}(\mathcal{T}) + \eta \cdot w_{d_i, t_k} \cdot \text{error}$ 
10                 fim
11             fim
12         até convergência ou número fixo de iterações
13         /* Second Step                                 */
14         para cada  $d_i \in \mathcal{D}^U$  faça
15              $f_{d_i} \leftarrow$  Classificar_documento( $d_i, \mathbf{F}(\mathcal{T})$ )
16              $f_{d_i} \leftarrow$  Padronizar( $f_{d_i}$ )
17         fim
18 fim

```

Algoritmo 7 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 2)Entrada: $\mathcal{O}, \mathbf{Y}, \mathbf{W}, \mathbf{D}, \eta$ Saída: $\mathbf{F}(\mathcal{O}^U)$

```

18 início
19   /* Third Step - Iterações Globais */
20   repita
21     /* Iterações Locais 1 */
22     repita
23       para cada  $d_i \in \mathcal{D}^U$  faça
24         InfClasses  $\leftarrow$  Classificar_documento( $d_i, F(\mathcal{T})$ )
25         para cada  $c_j \in \mathcal{C}$  faça
26           error  $\leftarrow f_{d_i, c_j} - \text{InfClasses}[c_j]$ 
27           para cada  $t_k \in T$  faça
28              $f_{t_k, c_j} \leftarrow f_{t_k, c_j} + \eta \cdot w_{d_i, t_k} \cdot \text{error}$ 
29           fim
30         fim
31       até até que um critério de parada seja atingido
32     /* Iterações Locais 2 */
33     repita
34       para cada  $t_i \in \mathcal{T}$  faça
35         para cada  $c_j \in \mathcal{C}$  faça
36           para cada  $t_k \in T$  faça
37             erro  $\leftarrow f_{t_k, c_j} - f_{t_i, c_j}$ 
38              $f_{t_i, c_j} \leftarrow f_{t_i, c_j} + \eta \cdot w_{t_i, t_k} \cdot \text{erro}$ 
39           fim
40         fim
41       até até que um critério de parada seja atingido
42     /* Iterações Locais 3 */
43     para cada  $d_i \in \mathcal{D}^U$  faça
44        $f_{d_i} \leftarrow$  Classificar_documento( $d_i, F(\mathcal{T})$ )
45        $f_{d_i} \leftarrow$  Padronizar( $f_{d_i}$ )
46     fim
47     /* Iterações Locais 4 */
48     repita
49       para cada  $d_i \in \mathcal{D}^U$  faça
50         para cada  $c_j \in \mathcal{C}$  faça
51           para cada  $d_k \in \mathcal{D}$  faça
52             erro  $\leftarrow f_{d_k, c_j} - f_{d_i, c_j}$ 
53              $f_{d_i, c_j} \leftarrow f_{d_i, c_j} + \eta \cdot w_{d_i, d_k} \cdot \text{erro}$ 
54           fim
55         fim
56       até até que um critério de parada seja atingido
57     /* Iterações Locais 5 */
58     para cada  $d_i \in \mathcal{D}^U$  faça
59        $f_{d_i} \leftarrow$  Padronizar( $f_{d_i}$ )
60     fim
61   até até que um critério de parada seja atingido
62 fim

```

Algoritmo 8 – Transductive Categorization based on Heterogeneous Networks (TCHN - parte 3)**Entrada:** $\mathcal{O}$, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{D}$, η **Saída:** $\mathbf{F}(\mathcal{O}^U)$

```

60 início
    /* Repita Global ainda está aberto-colocado novamente para manter
       mesma ident */
61 repita
    /* Iterações Locais 6 */
62     repita
63         para cada  $d_i \in \mathcal{D}^L$  faça
64             InfClasses  $\leftarrow$  Classificar_documento( $d_i, F(\mathcal{T})$ )
65             para cada  $c_j \in \mathcal{C}$  faça
66                 error  $\leftarrow f_{d_i, c_j} - \text{InfClasses}[c_j]$ 
67                 para cada  $t_k \in T$  faça
68                      $f_{t_k, c_j} \leftarrow f_{t_k, c_j} + \eta \cdot w_{d_i, t_k} \cdot \text{error}$ 
69                 fim
70             fim
71         fim
72     até até que um critério de parada seja atingido
73 até até que um critério de parada seja atingido
74 fim

```

2.5 Considerações Finais

Neste capítulo foram apresentados conceitos básicos sobre representação de dados textuais, aprendizado de máquina e mineração de opiniões. Foram apresentados com maior ênfase os trabalhos relacionados à Mineração de Opiniões baseadas em Aspectos que possibilita extrair informações mais específicas das revisões de usuários e no Aprendizado Semissupervisionado Transdutivo em Redes que foi base para as soluções propostas nesta tese.

MINERAÇÃO DE OPINIÕES BASEADA EM ASPECTOS POR MEIO DE APRENDIZADO SEMISSUPERVISIONADO

3.1 Considerações Iniciais

As abordagens tradicionais para Mineração de Opiniões visam classificar a polaridade do sentimento de documentos textuais como “positivo”, “negativo” ou “neutro”. Normalmente, essa classificação é realizada considerando os documentos como uma única unidade (mineração de opiniões no nível de documento) ou considerando cada sentença de um documento (mineração de opiniões em nível de sentença) (LIU, 2015). Em ambos os casos, informações sobre aspectos, ou seja, características ou propriedades de produtos ou serviços, não são exploradas. No entanto, a polaridade do sentimento pode ser diferente para aspectos distintos do mesmo produto ou serviço (CHEN; MUKHERJEE; LIU, 2014). Por exemplo, na seguinte frase “*Eu gostei da resolução da imagem da TV, mas seu controle remoto é terrível.*”, existem duas opiniões: (i) positiva sobre a “resolução da imagem” e (i) negativa sobre o “controle remoto”. Para lidar com esse tipo de cenário, é possível utilizar a Mineração de Opiniões baseada em Aspectos. Desse modo é possível aprimorar a tomada de decisões por meio de informações específicas sobre a polaridade do sentimento de cada aspecto de uma entidade (produto, serviço, empresa, instituição, etc.).

A Mineração de Opiniões em nível de aspectos é mais desafiadora e complexa do que em nível de documento ou nível de sentença (LIU, 2015; JIMÉNEZ-ZAFRA *et al.*, 2016). Apesar das diversas etapas necessárias realizar Mineração de Opiniões em nível de aspectos, é possível dividi-la em duas etapas principais: (i) extração de termos ou palavras relacionados aos aspectos, que trazem a informação sobre características e propriedades da entidade e (ii) classificação de sentimentos para cada aspecto identificado. Pesquisas promissoras sobre mineração de opiniões

baseada em aspectos combinam técnicas de Processamento de Língua Natural (*Natural Language Processing* (NLP)) e Aprendizado de Máquina para extração de aspectos e classificação de sentimentos (LIU, 2015; GANESHBHAI; SHAH, 2015).

O aprendizado não supervisionado é aplicado em Mineração de Opiniões, considerando técnicas de Modelos de Tópicos, como *Probabilistic Latent Semantic Analysis* (PLSA) (HOFMANN, 1999) e *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003), para gerar um modelo de tópicos de sentimentos (JIMÉNEZ-ZAFRA *et al.*, 2016; MOGHADDAM; ESTER, 2012; ZHAO *et al.*, 2010; TITOV; MCDONALD, 2008a; TITOV; MCDONALD, 2008b). Nessa abordagem, cada tópico é composto por um conjunto de termos. Na sequência é utilizada uma lista predefinida de aspectos e outra lista predefinida de termos (palavras e expressões) positivos e negativos para definir o aspecto e a polaridade do sentimento. A desvantagem dessa abordagem é definir o número correto de tópicos e a necessidade de um léxico predefinido. Além disso, definir um recurso léxico é muitas vezes problemático em cenários reais. Por exemplo, a palavra “rapidamente” tem um sentimento positivo para o aspecto “tempo de inicialização do sistema operacional” e um sentimento negativo para aspecto “descarregamento de bateria”. Assim, a polaridade da palavra pode mudar para diferentes contextos de domínios (ZHANG; LIU, 2011; BOLLEGALA; WEIR; CARROLL, 2011).

Os métodos baseados em Aprendizado de Máquina são alternativas potencialmente úteis para superar essas limitações. Neste caso, o objetivo é aprender um modelo que extraia automaticamente os aspectos dos textos, identificando a importância do léxico para o domínio de contexto, atributos linguísticos e relações sintáticas desses atributos (LIU, 2012; CHEN; MUKHERJEE; LIU, 2014). Embora as técnicas de aprendizado de máquina supervisionado forneçam melhores resultados que as baseadas em aprendizado não supervisionado (LIU, 2015; GANESHBHAI; SHAH, 2015), tais abordagens requerem um enorme esforço humano para rotular uma grande quantidade de aspectos e seus respectivos sentimentos para aprender um modelo (CHEN; MUKHERJEE; LIU, 2014; PONTIKI *et al.*, 2014).

Neste capítulo é apresentada a representação das revisões propostas nesta tese de doutorado para diminuir a necessidade de uso de dados rotulados. Além dessas considerações iniciais, o restante deste capítulo está organizado em mais quatro seções. Na Seção 3.2 é apresentada o esquema conceitual da representação proposta. Na Seção 3.3 são apresentados os detalhes da proposta para realizar extração de aspectos e classificação de sentimentos dos aspectos por meio de aprendizado semissupervisionado baseado em redes heterogêneas bipartidas. Na Seção 3.4 são apresentados a avaliação experimental realizada para comparar os resultados obtidos com a nossa proposta e com os resultados. Finalmente, na Seção 3.5 são apresentadas as considerações finais deste capítulo.

3.2 Visão Geral da Solução Proposta

Nesta tese de doutorado é proposta uma abordagem semissupervisionada para realizar Mineração de Opiniões baseada em Aspectos, que permite aprender um modelo usando poucos exemplos rotulados, o que a torna mais aplicável aos cenários do mundo real. Essa abordagem proposta, denominada (*Aspect-based Sentiment Propagation on Heterogeneous Networks* (ASPHN)), utiliza técnicas de aprendizado semissupervisionado baseado em modelos de grafos, em que os atributos linguísticos, os aspectos candidatos e os rótulos de sentimentos são modelados por meio de redes heterogêneas. Especificamente, foram utilizadas redes heterogêneas bipartidas para modelar os dados. Esse tipo de rede não tem parâmetros, é rapidamente gerada e apresentou resultados competitivos ou melhores do que os outros tipos de redes.

Na Figura 7 é apresentado o esquema conceitual da proposta ASPHN, a qual utiliza uma rede heterogênea composta por (i) aspectos $\mathcal{A} = \{a_1, \dots, a_r\}$; (ii) atributos gramaticais $\mathcal{G} = \{g_1, \dots, g_q\}$ conectados aos candidatos à aspectos; e (iii) termos $\mathcal{T} = \{t_1, \dots, t_s\}$ conectados aos aspectos classificados. As arestas indicam ausência ou presença das relações, especificamente, (i) quando o atributo gramatical está relacionado ao aspecto e (ii) quando um termo ocorre na mesma sentença de um aspecto. Na modelagem aqui proposta, os candidatos à aspectos são termos identificados na coleção textual, compostos por substantivos, verbos, adjetivos e advérbios. Já os atributos gramaticais são extraídos das sentenças, bem como da respectiva rede de dependência sintática.

Na implementação desse esquema conceitual, a ASPHN, são geradas duas redes bipartidas. Uma para extrair os aspectos, composta por dois tipos de objetos - atributos linguísticos e candidatos a aspecto. E outra, para analisar o sentimento dos aspectos, composta pelos aspectos identificados no passo anterior e termos que serão processados para identificar o sentimento (“positivo”, “negativo” ou “neutro”) do aspecto. Em ambas as redes são rotulados alguns objetos de acordo com a tarefa e os rótulos são propagados por meio da rede para classificar os objetos restantes. Para validar os resultados dessa abordagem proposta foi realizada uma avaliação experimental comparativa com os métodos tradicionais do estado-da-arte.

Figura 7 – Esquema conceitual da representação proposta

Fonte: Elaborada pelo autor.

3.3 Aspect-Based Sentiment Propagation for Heterogeneous Networks (ASPHN)

Dado um conjunto de documentos escritos em língua natural, representando mensagens, revisões, análises, ou notícias sobre um assunto específico, o problema da Mineração de Opiniões baseada em Aspectos pode ser definido como extrair opiniões representadas pela tripla $O = (e_i, a_{ij}, s_{ijk})$, em que e_i é uma entidade i (por exemplo: produto ou serviço), a_{ij} é o aspecto j (propriedade) da entidade i , e s_{ijk} é a polaridade do sentimento em relação ao aspecto a_{ij} da entidade e_i (por exemplo: “positivo”, “negativo” ou “neutro”). Nesta tese de doutorado, a entidade é uma informação prévia do problema, de forma que o objetivo do aprendizado é a extração de aspectos e a classificação dos respectivos sentimentos desses aspectos referentes à entidade definida.

Nesta proposta, denominada *Aspect-Based Sentiment Propagation for Heterogeneous Networks* (ASPHN), faz uso de redes heterogêneas para realizar a Mineração de Opiniões baseada em Aspectos. Esta opção de modelar este problema como uma rede heterogênea e realizar a propagação de rótulos, foi escolhida, pois esta combinação de técnicas foi utilizada com sucesso no aprendizado semissupervisionado (ROSSI; LOPES; REZENDE, 2016; JI *et al.*, 2010). Nas próximas seções é apresentado essa representação da abordagem proposta para Mineração de Opiniões baseada em Aspectos por meio de aprendizado semissupervisionado. Além disso, é discutido como essa abordagem é bem apropriada neste contexto.

A ASPHN utiliza duas redes heterogêneas bipartidas, uma para cada etapa da Mineração de Opiniões baseada em Aspectos: (i) extração de aspectos e (ii) classificação de sentimentos dos aspectos. Para isso, é necessário que: (i) gerar as redes heterogêneas bipartidas de forma a permitir a identificação/extração do aspecto e a classificação do sentimento do aspecto e (ii) executar a propagação de rótulos na rede para classificar os objetos como aspectos e não aspectos (primeiro passo) e como positivo, negativo e neutro (segundo passo). Na próxima seção, é detalhado como são geradas as redes heterogêneas bipartidas e como é realizado o aprendizado semissupervisionado por meio da propagação de rótulos em tais redes.

3.3.1 Geração de Redes Heterogêneas Bipartidas

Formalmente, uma rede heterogênea bipartida pode ser definida como $N = \langle \mathcal{O}, \mathcal{E}, \mathcal{W} \rangle$, em que $\mathcal{O}$ representa dois conjuntos de rede de objetos (também chamados de vértices ou nós), $\mathcal{O} = \mathcal{O}_\infty \cup \mathcal{O}_\epsilon$, $\mathcal{E}$ representa o conjunto de conexões (também chamadas de relações ou links), que ocorre apenas de objetos de um conjunto $\mathcal{O}_\infty$ para os de outro conjunto $\mathcal{O}_\epsilon$, e $\mathcal{W}$ representa os pesos das conexões.

A primeira rede heterogênea bipartida é composta por candidatos a aspecto $\mathcal{C} = \{c_1, \dots, c_r\}$ e por atributos linguísticos $\mathcal{L} = \{l_1, \dots, l_q\}$. Um candidato a aspecto é conectado ao atributo linguístico se houver uma relação entre eles. Neste trabalho, os aspectos candidatos podem

ser substantivos, verbos¹, adjetivos e advérbios (LIU, 2015). Os atributos linguísticos são extraídos das árvores de dependência sintática das sentenças. Por exemplo, na seguinte sentença “*The food was nothing much, but I loved the staff.*”, pode-se obter a árvore de dependência sintática apresentada na Figura 8. Neste projeto, essa árvore foi gerada usando o *Stanford CoreNLP toolkit* (MANNING *et al.*, 2014). A partir da árvore sintática, são extraídos os atributos linguísticos considerando cada classe gramatical ou cada relação de dependência sintática. Como passo seguinte, são removidas as *stopwords* (preposições, artigos, conjunções e pronomes) das sentenças. O peso do atributo é igual a 1 se o aspecto candidato entrar em uma relação sintática e 0 caso contrário. Na Tabela 2 é apresentada uma representação no modelo espaço-vetorial extraída da árvore sintática apresentada na Figura 8. Para melhor ilustrar, são diferenciados os atributos da classe gramatical dos atributos da relação de dependência sintática por meio de letras maiúsculas e minúsculas, respectivamente.

Figura 8 – Exemplo de árvore de dependência sintática gerada para a sentença “*The food was nothing much, but I loved the staff.*”

Fonte: Elaborada pelo autor.

A partir da representação no modelo espaço-vetorial, é extraída a rede heterogênea bipartida, que consiste em um mapeamento direto da representação no modelo espaço-vetorial para o grafo. Nesse caso, a rede possui dois tipos de objetos: candidatos a aspecto e atributos linguísticos. Um aspecto candidato está conectado a um atributo linguístico se o peso na representação no modelo espaço-vetorial for igual ao valor 1. Na Figura 9 é apresentado um exemplo de uma rede heterogênea bipartida extraída da representação no modelo espaço-vetorial apresentada na Tabela 2.

Tabela 2 – Representação espaço-vetorial gerada a partir da árvore sintática apresentada na Figura 8.

	NN	RB	JJ	PR	DT	VB	CC	det	nsubj	cop	advmod	cc	conj	dobj
food	1	0	0	0	0	0	0	1	1	0	0	0	0	0
nothing	1	0	0	0	0	0	0	0	0	1	1	1	1	0
much	0	1	0	0	0	0	0	0	0	0	1	0	0	0
staff	1	0	0	0	0	0	0	1	0	0	0	0	0	1

A segunda rede heterogênea bipartida é composta por (i) aspectos $\mathcal{A} = \{a_1, \dots, a_s\}$ (identificados na etapa anterior), e (ii) termos $\mathcal{V} = \{v_1, \dots, v_n\}$. Os aspectos são conectados com os termos apresentados na mesma sentença. Destaca-se que as *stopwords* das sentenças são removidas. Para exemplificar essa representação proposta, considere estas três sentenças: (i)

¹ Alguns verbos de estado são considerados *stopwords*, como formas do verbo “to be”.

Figura 9 – Exemplo de rede heterogênea gerada a partir da representação no modelo espaço-vetorial apresentada na Tabela 2.

Fonte: Elaborada pelo autor.

“The french fries were excellent. The drinks were terrible”, (ii) “In general, the snacks and french fries were good.”, e (iii) “In general, drinks were terrible, mainly the wine.”. A representação no modelo espaço-vetorial resultante dessa proposta é apresentada na tabela Tabela 3 e a partir desta representação, foi construída a rede heterogênea bipartida apresentada na Figura 10.

Tabela 3 – Representação no modelo espaço-vetorial para aspectos extraídos no passo anterior das três sentenças.

	excellent	good	general	mainly	terrible
french-fries	1	1	0	0	0
snacks	0	1	1	0	0
wine	0	0	1	1	1
drinks	0	0	0	1	1

Figura 10 – Exemplo de rede heterogênea bipartida extraída da representação no modelo espaço-vetorial apresentado na Tabela 3.

Fonte: Elaborada pelo autor.

3.3.2 Aprendizado Semissupervisionado baseado em Redes Heterogêneas Bipartidas

O aprendizado semissupervisionado baseado em redes heterogêneas bipartidas apresentou resultados satisfatórios em comparação com outros tipos de redes (JI *et al.*, 2010; YIN *et al.*, 2009), principalmente quando se considera dados textuais (ROSSI; LOPES; REZENDE, 2016; ROSSI; LOPES; REZENDE, 2014). Além do fato de que a rede bipartida é um mapa direto de uma representação espaço-vetorial, o uso dessa rede permite fazer uso da regularização de grafo para realizar o aprendizado. Foi demonstrado que a regularização do grafo é o estado-da-arte para realizar o aprendizado semissupervisionado (ROSSI; LOPES; REZENDE, 2016; ZHU *et al.*, 2009; CHAPELLE; SCHÖLKOPF; ZIEN, 2006). A ideia do aprendizado semissupervisionado

baseado em redes bipartidas é classificar um tipo de objetos, denominados de objetos alvo, usando outro tipo de objeto, objetos de ponte, para propagar informações de classe entre os objetos alvo.

Para isso, tem-se que $\mathcal{C} = \{c_1, c_2, \dots, c_l\}$ representa o conjunto de rótulos de classe, $\mathcal{B} = \{b_1, b_2, \dots, b_m\}$ representa o conjunto de objetos de ponte, e $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ representa o conjunto de objetos alvo. No cenário de aprendizado semissupervisionado, $\mathcal{T} = \mathcal{T}^L \cup \mathcal{T}^U$, em que $\mathcal{T}^L$ representa o conjunto de objetos alvos rotulados e $\mathcal{T}^U$ representa o conjunto de objetos alvos não rotulados. Além disso, $\mathbf{f}_{o_i} = \{f_{c_1}, f_{c_2}, \dots, f_{c_{|\mathcal{C}|}}\}$ representa o i -ésimo vetor de peso de um objeto de rede o_i que armazena os pesos de um objeto o_i para todas as classes em $\mathcal{C}$ atribuídos durante o aprendizado semissupervisionado. Por isso, também é referido como vetor de informação de classe. Seja $\mathbf{F}(\mathcal{O}) = \{\mathbf{f}_{o_1}, \mathbf{f}_{o_2}, \dots, \mathbf{f}_{o_{|\mathcal{O}|}}\}^T$ uma a matriz que armazena todos os vetores de pesos dos objetos, os rótulos predefinidos para um alvo $t_i \in \mathcal{T}^L$ são armazenados como um vetor de pesos $\mathbf{y}_{t_i} = \{t_1, t_2, \dots, t_{|\mathcal{C}|}\}$, que tem valor 1 na posição correspondente da classe do objeto alvo t_i e valor 0 nas outras posições. Os pesos das relações entre os objetos da rede são armazenados em uma matriz $\mathbf{W}$. Os pesos das relações entre o objeto o_i e os outros objetos da rede são representados pelo vetor $\mathbf{w}_{o_i} = \{w_{o_1}, w_{o_2}, \dots, w_{o_{|\mathcal{O}|}}\}$. Este vetor também é usado para representar um documento d_i no modelo espaço-vetorial e armazenar a frequência, ou outra medida de frequência, dos termos no documento d_i . Na Figura 11 é ilustrada uma rede bipartida que usa essa estrutura descrita.

Figura 11 – Exemplo de representação de rede bipartida para aprendizado semissupervisionado.

Fonte: Elaborada pelo autor.

Nas Figura 12 e Figura 13 são apresentados exemplos de aprendizado semissupervisionado baseado em redes bipartidas considerando o algoritmo TCBHN (descrito no Capítulo 2) com $\eta = 0.5$. Neste primeiro exemplo, foi realizado para a rede heterogênea bipartida apresentada na Figura 9 para realizar a extração de aspecto, em que $\mathcal{C} = \{\text{ASPECT}, \text{NOT-ASPECT}\}$, $\mathcal{B} = \{NN, det, nsubj, dobj, cop, advmod, cc, conj, RB\}$, $\mathcal{T}^L = \{\text{food}, \text{much}\}$ e $\mathcal{T}^U = \{\text{staff}, \text{nothing}\}$. No segundo exemplo, foi considerada a rede heterogênea bipartida apresentada na Figure 10 para realizar a classificação do sentimento do aspecto, em que

$\mathcal{C} = \{\text{POSITIVE}, \text{NEGATIVE}\}$, $\mathcal{B} = \{\text{excellent}, \text{good}, \text{general}, \text{mainly}, \text{terrible}\}$, $\mathcal{T}^L = \{\text{"french-fries"}, \text{"drinks"}\}$ e $\mathcal{T}^U = \{\text{"snacks"}, \text{"wine"}\}$. Para facilitar a ilustração da propagação dos rótulos, as frequências dos termos são iguais 1 e $\eta = 0.5$.

Figura 12 – Exemplo de aprendizado semissupervisionado baseado em redes bipartidas considerando o algoritmo TCBHN com $\eta = 0.5$ para extração de aspectos.

(a) Estado inicial da estrutura da rede bipartida para extração de aspectos.

(b) Estado intermediário da estrutura da rede bipartida.

(c) Estado intermediário da estrutura da rede bipartida.

(d) Estado final da estrutura da rede bipartida com pesos atribuídos a todos os objetos de rede.

ASPECTO NÃO ASPECTO Não Classificado

Fonte: Elaborada pelo autor.

Figura 13 – Exemplo de aprendizado semissupervisionado baseado em redes bipartidas considerando o algoritmo TCBHN com $\eta = 0.5$ para classificação de sentimentos dos aspectos.

(a) Estado inicial da estrutura da rede bipartida para classificação de sentimentos dos aspectos.

(b) Estado intermediário da estrutura da rede bipartida.

(c) Estado intermediário da estrutura da rede bipartida.

(d) Estado final da estrutura da rede bipartida com pesos atribuídos a todos os objetos de rede.

POSITIVO
 NEGATIVO
 Não Classificado

Fonte: Elaborada pelo autor.

A regularização do grafo deve satisfazer duas premissas: (i) dois objetos conectados no grafo tendem a ser classificados com o mesmo rótulo e (ii) os rótulos dos objetos devem estar próximos da informação de classe real (conjunto de treinamento). Na sequência são apresentados os algoritmos para realizar a regularização dos grafos em redes bipartidas utilizados nesta tese de doutorado, com suas respectivas funções de regularização de grafos. Em cada função de regularização, o primeiro termo está relacionado com a primeira premissa e, analogamente, o segundo termo descreve a segunda premissa.

- **Label Propagation based on Bipartite Heterogeneous Network (LPBHN)** (ROSSI; LOPES; REZENDE, 2014): este é um algoritmo sem parâmetros para realizar o aprendizado semissupervisionado em redes bipartidas. Este algoritmo é uma extensão do algoritmo *Gaussian Fields e Harmonic Function* (GFHF) (ZHU; GHAMRANI; LAFFERTY, 2003) para redes heterogêneas bipartidas. A função de regularização a ser minimizada pelo LPBHN é dada pela Equação 3.1

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{t_i \in \mathcal{T}} \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} (\mathbf{f}_{t_i} - \mathbf{f}_{b_j})^2 + \lim_{\mu \rightarrow \infty} \mu \sum_{t_i \in \mathcal{T}^L} (\mathbf{f}_{t_i} - \mathbf{y}_{t_i})^2 \quad (3.1)$$

em que há uma restrição que $\mathbf{f}_{t_i} = \mathbf{y}_{t_i}$, então o segundo termo da Equação 3.1 tem valor que tende ao infinito.

- **GNetMine (GNM)** (JI *et al.*, 2010): esta é uma extensão do algoritmo *Learning with Local and Global Consistency* (LLGC) (ZHOU *et al.*, 2004). A função de regularização a ser minimizada pelo GNetMine é dada pela Equação 3.2

$$Q(\mathbf{F}) = \sum_{t_i \in \mathcal{T}} \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} \left\| \frac{\mathbf{f}_{t_i}}{\sqrt{\sum_{b_k \in \mathcal{B}} w_{t_i, b_k}}} - \frac{\mathbf{f}_{b_j}}{\sqrt{\sum_{t_k \in \mathcal{T}} w_{t_k, b_j}}} \right\|^2 + \sum_{t_i \in \mathcal{T}^L} \alpha_{t_i} (\mathbf{f}_{t_i} - \mathbf{y}_{t_i}), \quad (3.2)$$

em que $0 < \alpha < 1$ define a importância de cada termo da Equação 3.2.

- **Tag-based Model (TM)** (YIN *et al.*, 2009): este algoritmo foi inicialmente proposto para classificar objetos conectados a *tag* sociais. No contexto deste trabalho, a função de regularização a ser minimizada por TM é dada pela Equação 3.3.

$$Q(\mathbf{F}) = \left(\beta \sum_{b_i \in \mathcal{B}^L} \|\mathbf{f}_{b_i} - \mathbf{y}_{b_i}\|^2 + \gamma \sum_{B_i \in \mathcal{B}^U} \|\mathbf{f}_{b_i} - \mathbf{y}_{b_i}\|^2 \right) + \left(\sum_{b_i \in \mathcal{B}} \sum_{t_j \in \mathcal{T}} w_{b_i, t_j} \|\mathbf{f}_{b_i} - \mathbf{f}_{t_j}\|^2 \right), \quad (3.3)$$

em que os parâmetros β e γ controlam a importância dada ao termo da Equação 3.3.

- **Transductive Classification based on Bipartite Heterogeneous Network (TCBHN)** (ROSSI; LOPES; REZENDE, 2016): este é um algoritmo que realiza a otimização e propagação dos rótulos minimizando a Equação 3.4.

$$Q(\mathbf{F}) = \frac{1}{2} \left(\sum_{c_k \in \mathcal{C}} \left(\sum_{t_i \in \mathcal{T}^U} f_{t_i, c_k} - \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} \cdot f_{b_j, c_k} \right) \right)^2 + \frac{1}{2} \left(\sum_{c_k \in \mathcal{C}} \left(\sum_{t_i \in \mathcal{T}^L} y_{t_i, c_k} - \sum_{b_j \in \mathcal{B}} w_{t_i, b_j} \cdot f_{b_j, c_k} \right) \right)^2 \quad (3.4)$$

Na abordagem proposta, a ASPHN, a primeira rede bipartida foi utilizada para a classificação de aspectos nos quais os vértices são os atributos linguísticos e poucos aspectos são rotulados como “ASPECTO” ou “NÃO-ASPECTO”. Após a propagação de rótulos desta etapa, os aspectos classificados como “ASPECTO” são usados na segunda rede bipartida, nos quais os rótulos dos aspectos são “POSITIVO”, “NEGATIVO” ou “NEUTRO”; e os vértices são os termos. A ideia desta segunda rede bipartida é que, se um termo estiver conectado com um aspecto rotulado, essas informações de rótulos são propagadas para aspectos não marcados relacionados ao mesmo termo. Após este estágio, a rede heterogênea contém os aspectos classificados, bem como a polaridade do aspecto.

3.4 Avaliação Experimental

Para realizar a avaliação experimental foi analisada a viabilidade da ASPHN aqui proposta para realizar a extração de aspectos e classificação de sentimentos dos aspectos. Mais detalhes sobre as coleções de revisões, a configuração experimental, os critérios de avaliação, e os resultados são apresentados nas próximas seções.

3.4.1 Coleções de Revisões e Representação Estruturada

Para a realização dos experimentos, como conjunto de dados foram utilizadas duas coleções de revisões que contém anotações de aspectos e suas respectivas polaridades de sentimentos. Essas coleções possuem revisões no idioma inglês. A primeira coleção de revisões é composta por 3.044 revisões sobre restaurantes e a segunda é composta por 3.048 sobre laptops. Ambas coleções estão disponíveis em Pontiki *et al.* (2014). Cada revisão pode conter um ou mais de um aspecto. A polaridade do aspecto pode ser positiva, negativa ou neutra. Os aspectos com polaridade com conflito não foram considerados neste experimento. A coleção de *revisões* de restaurantes contém 3608 aspectos com os seguintes sentimentos de polaridade: 2164 são positivos, 637 são neutros, 807 são negativos e 91 apresentaram polaridades conflitantes. A coleção de *revisões* de laptop contém 2328 aspectos com os seguintes sentimentos de polaridade: 994 são positivos, 464 são neutros, 870 são negativos e 45 apresentaram polaridades conflitantes.

Além disso, foi calculado o *S-Index* (ROSSI; MARCACINI; REZENDE, 2013) para as coleções de *revisões* de restaurante e laptop. O *S-Index* indica a sobreposição entre as classes. Quando *S-Index* é próximo ao valor 1 significa que as classes estão bem separadas, isto é, não há sobreposição entre as classes. Caso contrário, o valor é próximo de 0. Portanto, o *S-Index* pode ser usado para entender a complexidade da classificação para cada conjunto de dados em cenários controlados. Os valores do *S-Index* para a extração de aspecto e a classificação do sentimento na coleção de *revisões* de laptop são 0,82 e 0,75, respectivamente. Da mesma forma, os valores do *S-Index* são 0,89 e 0,77 para a coleção de *revisões* de restaurante. Assim, observa-se (*apriori*) que a coleção de *revisões* de restaurante é um problema de classificação

mais complexo do que coleção de laptop. Devido a razões de reprodução, todos os códigos-fonte e coleções de *revisões* usados em nesta avaliação experimental estão disponíveis em <http://sites.labc.icmc.usp.br/ipm/tese/cap3>.

Para representar a coleção de *revisões* para os algoritmos de aprendizado de máquina para realizarem a extração de aspectos, foram extraídos os atributos linguísticos, conforme apresentado na Seção 3.3.1. Na sequência, foi gerada uma representação do modelo espaço-vetorial, conforme apresentado na Tabela 2 e, em seguida, foram geradas as redes bipartidas conforme apresentado na Figura 9. Foram gerados 75 atributos linguísticos compostos por classes gramaticais e dependências sintáticas. As representações do modelo de espaço vetorial e as representações de rede bipartida foram geradas de acordo com as etapas apresentadas na Seção 3.3.1, conforme apresentado na Tabela 3 e Figura 10 respectivamente.

3.4.2 Configuração Experimental

Na avaliação experimental, a proposta ASPHN foi comparada com outros algoritmos de aprendizado semissupervisionado com base no modelo de espaço-vetorial. Também, foi comparada com algoritmos de aprendizado supervisionado, que tradicionalmente são utilizados na Mineração de Opiniões. Essas comparações foram realizadas por dois motivos: (i) verificar se o uso de aprendizado semissupervisionado baseado em redes heterogêneas bipartidas é superior ao desempenho do aprendizado semissupervisionado com base no modelo de espaço-vetorial e (ii) para analisar se a ASPHN proposta é superior ao desempenho obtido pelo aprendizado de máquina tradicional, ou seja, o aprendizado supervisionado baseado no modelo de espaço-vetorial e que usa uma quantidade grande de dados rotulados. Além disso, foi avaliado se quantidade de exemplos não rotulados são úteis para melhorar a extração de aspectos e a classificação de sentimentos de aspectos.

Os algoritmos de aprendizado semissupervisionados baseados em redes bipartidas apresentadas na Seção 3.3.2 foram utilizados em nossa avaliação experimental. Os algoritmos e seus respectivos parâmetros foram baseados no trabalho de Rossi, Lopes e Rezende (2016) que realizaram uma pesquisa para estimar os valores dos parâmetros e quantidades de iterações necessárias para uma avaliação completa. Esse algoritmos são apresentados a seguir.

- **Tag-based Model (TM)**: foi utilizado $\beta = \{0.1, 1, 10, 100, 1000\}$ e $\gamma = \{0.1, 1, 10, 100, 1000\}$.
- **GNetMine**: foi utilizado $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$.
- **Transductive Categorization based on Bipartite Heterogeneous Networks (TCBHN)**: foi considerada uma solução iterativa do TCBHN apresentada em (ROSSI; LOPES; REZENDE, 2016). Esta solução iterativa tem dois parâmetros: η (taxa de correção de erro) e ε (erro quadrático mínimo). Foram utilizados $\eta = \{0.01, 0.05, 0.1, 0.5\}$, $\varepsilon = 0.01$,

10 como número máximo de iterações globais e 100 como número máximo de iterações locais, o que dá um total de 1000 iterações.

- **Label Propagation using Bipartite Heterogeneous Networks (LPBHN)**: este é um algoritmo sem parâmetros.

Nesta avaliação experimental, também foram considerados os algoritmos de aprendizado semissupervisionado tradicionais com base no modelo de espaço-vetorial (ZHU *et al.*, 2009). Os algoritmos e seus parâmetros são listados a seguir.

- **Self-Training**: em cada etapa da abordagem de *Self-Training*, os X exemplos classificados como mais confiáveis, considerando um modelo de classificação anterior (induzido por meio de aprendizado supervisionado), são adicionados ao conjunto de exemplos rotulados. Este processo é repetido até que todos os documentos não rotulados sejam adicionados ao conjunto de documentos rotulados. Foi considerado $X = \{5, 10, 15, 20\}$. Como o algoritmo de aprendizado indutivo para *Self-Training* foi utilizado o *Multinomial Naïve Bayes* (MNB), pois apresenta o melhor *trade-off* entre o desempenho de classificação e o tempo para dados textuais.
- **Expectation Maximization (EM)**: considerou-se o algoritmo EM para a classificação de dados textuais apresentado em (NIGAM *et al.*, 2000). Neste algoritmo é necessário definir os parâmetros referente ao peso dos exemplos não rotulados durante o aprendizado semissupervisionado (λ) e o número de componentes para cada classe (*comp*). Foram utilizados $\lambda = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ e $comp = \{1, 2, 5, 10\}$.

Também foram executados algoritmos de aprendizado supervisionados indutivos do estado-da-arte para análise e comparação com a ASPHN proposta para aprendizado semissupervisionado. Isso também possibilitou analisar se o uso de exemplos não rotulados realmente melhora o desempenho da classificação. Os algoritmos, parâmetros e considerações dos algoritmos de aprendizado supervisionado indutivo são:

- **Multinomial Naïve Bayes (MNB)**: foi utilizado o MNB, pois é o algoritmo de aprendizado usado no *Self-Train*, *Expectation Maximization* (EM). Isso possibilitou medir a diferença no desempenho e no tempo de classificação para passar de aprendizado supervisionado indutivo para o aprendizado transdutivo para algoritmos baseados no modelo de espaço-vetorial. O algoritmo MNB não possui parâmetros.
- **Support Vector Machine (SVM)**: foram considerados *kernel* (i) Linear, (ii) Polinomial (expoente = 2) e (iii) RBF (*Radial Basis Function*). O *kernel* linear é o mais aconselhável para representações com alta dimensionalidade (CARUANA; NICULESCU-MIZIL, 2006). Como o parâmetro C é real, positivo e sem limite, utilizou-se esses valores como 10^Y . Para

cada tipo de *kernel* foi considerado $Y = \{-5; -4; -3; -2; -1; 0; 1; 2; 3; 4; 5\}$. Foi utilizada a implementação do SVM disponível na ferramenta Weka².

- ***k*-Nearest Neighbors *k*-NN**: foi considerado $k = \{7; 17; 37; 57\}$. Também foi considerado o algoritmo *k*-NN com e sem votação ponderada, que dá para cada um dos vizinhos mais próximos um voto em peso igual a $(1 - s)$, em que s é uma medida de similaridade entre os vizinhos. A medida de similaridade adotada foi *co-seno*.

Foram utilizadas soluções iterativas para todos os algoritmos que possuem soluções iterativas (EM, LPBHN, GNetMine, TCBHN e TM). O número máximo de iterações definido foi 1000.

3.4.3 Critérios de Avaliação

A medida F^1 foi utilizada para comparar os resultados da classificação. F^1 é a média harmônica das medidas de *Precisão* e *Cobertura*, em que ambas as medidas têm o mesmo peso, isto é:

$$F^1 = 2 * \frac{\text{Precisão} * \text{Cobertura}}{\text{Precisão} + \text{Cobertura}}. \quad (3.5)$$

Precisão e *Cobertura* são computadas para cada classe em uma avaliação multi-classe. A *Precisão* e *Cobertura* de cada classe c_i são:

$$\text{Precisão}_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FP_{c_i}}, \quad (3.6a) \quad \text{Cobertura}_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FN_{c_i}}, \quad (3.6b)$$

em que *TP* (Verdadeiro Positivo) é o número de documentos de teste que classificaram corretamente na classe c_i , *FP* (Falso Positivo) é o número de documentos de teste que são da classe c_j ($c_j \neq c_i$), mas foram classificados como c_i , e *FN* (Falso Negativo) é o número de documentos de teste da classe c_i mas que são da classe c_j ($c_j \neq c_i$). A medida de *Precisão* retorna a porcentagem de documentos que são corretamente classificados como c_i considerando todos os documentos classificados como c_i . A medida de *Cobertura* retorna a porcentagem de documentos corretamente classificados como c_i considerando todos os documentos que são atualmente da classe c_i .

Duas estratégias para sumarizar os resultados dos valores de precisão e de cobertura são computados para cada classe da coleção de textos: (i) ***micro-averaging*** e ***macro-averaging*** (SOKOLOVA; LAPALME, 2009). A estratégia de *micro-averaging* é um somatório de todos as medidas de avaliação. Portanto, a precisão e cobertura usando *micro-averaging* são:

² Weka: <<http://www.cs.waikato.ac.nz/ml/weka/>>

$$Precisão^{Micro} = \frac{\sum_{c_i \in \mathcal{C}} VP_{c_i}}{\sum_{c_i \in \mathcal{C}} (VP_{c_i} + FP_{c_i})}, \quad (3.7)$$

$$Cobertura^{Micro} = \frac{\sum_{c_i \in \mathcal{C}} VP_{c_i}}{\sum_{c_i \in \mathcal{C}} (VP_{c_i} + FN_{c_i})}. \quad (3.8)$$

A estratégia de *Macro-averaging* considera as médias de todas medidas de avaliação para cada classe. Portanto, *Precisão* e *Cobertura* usando estratégia de *Macro-averaging* são:

$$Precisão^{Macro} = \frac{\sum_{c_i \in \mathcal{C}} Precisão_{c_i}}{|\mathcal{C}|}, \quad (3.9)$$

$$Cobertura^{Macro} = \frac{\sum_{c_i \in \mathcal{C}} Cobertura_{c_i}}{|\mathcal{C}|}. \quad (3.10)$$

Micro-averaging são dominados pelo número de *VP*. Portanto, as grandes classes dominam pequenas classes em pontuações de micro-média. Por outro lado, a *Macro-averaging* dá peso igual a cada classe. Nesse caso, o número de *VP* em classes pequenas é enfatizado nas pontuações *Macro-averaging*. Essas duas estratégias dão diferentes pontuações e são complementares entre si. Nesta avaliação, F^1 é computada por meio da *Micro-averaging* de *Precisão* e *Cobertura* por *Micro- F^1* , e por *Macro-averaging* por *Macro- F^1* .

Para obter *Micro- F^1* e *Macro- F^1* , primeiro foi realizado um processo de validação cruzada de 10-fold. Para cada conjunto de treinamento (9 folds) foram realizadas 10 execuções para induzir um modelo de classificação considerando N documentos rotulados selecionados aleatoriamente em cada execução. Foi considerado $N = \{1, 10, 20, 30, 40, 50, 60, 70\}$. Esta variação no número de documentos rotulados permitiu: (i) demonstrar melhor o comportamento dos algoritmos para diferentes números de documentos rotulados; (ii) um *trade-off* entre o número de documentos rotulados e o desempenho de classificação; e (iii) verificar as diferenças entre os algoritmos de aprendizado supervisionados indutivos e semissupervisionados à medida que aumenta-se o número de documentos rotulados.

Os exemplos de treinamento restantes foram considerados como exemplos não-rotulados para algoritmos de aprendizado semissupervisionado. Assim, foram realizadas 100 execuções e, em cada execução, obteve-se um valor de *Precisão* e *Cobertura*. Os valores finais *Micro- F^1* e *Macro- F^1* apresentados na próxima seção foram uma média dos 100 valores obtidos na validação cruzada com 10 partes (*10-cross-validation*). Além da análise direta comparando os desempenhos de classificação, os algoritmos também foram comparados considerando o teste de significância estatística de Friedman e o teste *post hoc* de Nemenyi com 95% de nível de confiança para avaliar diferenças estatisticamente significativas entre os resultados da classificação (DEMSAR, 2006).

3.4.4 Análise e Discussão dos Resultados

Os resultados foram analisados considerando cada uma das etapas para realizar Mineração de Opiniões baseada em Aspectos individualmente, ou seja, uma análise sobre extração de

aspectos e outra sobre classificação da polaridade do sentimento.

3.4.4.1 Extração de Aspectos

Na Figuras 14a e 14b são apresentados, respectivamente, o desempenho da classificação de aspectos para as coleções de *revisões* de laptop e de restaurante. Considerando as duas medidas $Micro-F^1$ e $Macro-F^1$, o TCBHN apresentou os valores mais altos para todo o número de exemplos rotulados usados na avaliação experimental. O algoritmo TM apresentou o segundo melhor valor de $Micro-F^1$ para todo o número de exemplos rotulados.

Figura 14 – Resultados para extração de aspectos.

(a) Coleção de revisões de laptop.

(b) Coleção de revisões de restaurante.

Fonte: Elaborada pelo autor.

Figura 15 – Diagramas de diferença crítica para extração de aspectos.

Fonte: Elaborada pelo autor.

Na Figura 15 são apresentados os diagramas de Diferença Crítica entre os diferentes algoritmos de aprendizado aplicados para a extração de aspectos considerando os resultados das medidas de $Micro-F^1$ e $Macro-F^1$ apresentados nos gráficos da Figura 14. Nestes diagramas, cada algoritmo é classificado de acordo com o ranking do teste significância estatística de Friedman (DEMSAR, 2006). Os métodos conectados por uma linha não apresentam diferença estatística significativa entre eles.

De acordo com os diagramas de diferença crítica, o TCBHN obteve o melhor ranking médio para $Micro-F^1$ e $Macro-F^1$. Além disso, o TCBHN apresentou melhores resultados com diferenças estatisticamente significativas do que todos os algoritmos de aprendizado supervisionados. Em geral, não houve diferenças estatisticamente significativas entre os algoritmos de aprendizado semissupervisionados. No entanto, enfatiza-se que os algoritmos baseados em redes bipartidas apresentam um menor custo de computação do que os algoritmos baseados no modelo de espaço-vetorial (ROSSI; LOPES; REZENDE, 2016). Assim, pode-se concluir que o uso de exemplos não rotulados também é útil para a extração de aspectos por meio de aprendizado de máquina.

3.4.4.2 Classificação de Sentimentos de Aspectos

Na Figura 17a é apresentado o desempenho da classificação das polaridades do sentimento para a coleção de *revisões* de Laptop. Considerando a medida $Micro-F^1$, o algoritmo EM apresentou os valores mais altos ao considerar 1 e 10 exemplos rotulados. O algoritmo TM apresentou os valores mais elevados de $Micro-F^1$ ao considerar 20 ou mais exemplos rotulados

por classe. Os algoritmos TCBHN e Self-Training apresentaram resultados próximos de TM ao usar 20 ou mais exemplos rotulados por classe. Considerando a medida $Macro-F^1$, novamente, o algoritmo EM apresentou os valores mais altos ao considerar 1 e 10 exemplos rotulados. Os algoritmos TM e TCBHN apresentaram os valores mais elevados quando consideraram 20 ou mais exemplos rotulados.

Na Figura 17b é apresentado o desempenho dos algoritmos para coleção de *revisões* de restaurante. Considerando a medida $Micro-F^1$, o algoritmo GnetMine apresentou os valores mais altos para 1 exemplo rotulado por classe, o EM apresentou os valores mais altos para 10 exemplos rotulados e o TCBHN os valores mais altos ao considerar 20 ou mais exemplos rotulados para cada classe. Considerando a medida $Macro-F^1$, o TCBHN apresentou os valores mais altos para todo o número de exemplos rotulados. Para ambas coleções de *revisões* de laptop e de restaurante e para ambas medidas $Micro-F^1$ e $Macro-F^1$, os algoritmos de aprendizado semissupervisionados apresentaram melhor desempenho do que os algoritmos de aprendizado supervisionado. A exceção é o algoritmo LPBHN que apresentou resultados inferiores aos algoritmos de aprendizado supervisionado para a coleção de *revisões* de restaurante na medida $Macro-F^1$.

Na Figura 16 são apresentados os diagramas de diferença crítica entre os diferentes algoritmos de aprendizado utilizados para a classificação dos sentimentos do aspectos, considerando os resultados das medidas de $Micro-F^1$ e $Macro-F^1$ apresentados na Figura 17. Pode-se observar que o TCBHN obteve o melhor ranking médio e o TM obteve o segundo melhor ranking considerando as medidas de $Micro-F^1$ e $Macro-F^1$. Destaca-se que ambos os algoritmos são baseados em redes heterogêneas bipartidas.

Os algoritmos TCBHN e TM apresentaram melhores resultados com diferenças estatisticamente significativas do que o algoritmo de aprendizado supervisionado. Assim, pode-se concluir que a modelagem bipartida e o uso de exemplos não rotulados foram novamente úteis para melhorar a classificação do sentimento do aspecto.

Figura 16 – Diagramas de diferença crítica para classificação de sentimentos dos aspectos.

(a) Micro- F^1 . (b) Macro- F^1 .

Fonte: Elaborada pelo autor.

Figura 17 – Resultados para classificação de sentimentos de aspectos.

(a) Coleção de revisões de laptop

(b) Coleção de revisões de restaurante

Fonte: Elaborada pelo autor.

3.5 Considerações Finais

Neste capítulo foi apresentada a abordagem proposta nesta tese de doutorado: a ASPHN (*Aspect-Based Sentiment Propagation on Heterogeneous Networks*). Esta proposta realiza a extração de aspectos e a classificação de sentimento do aspecto por meio de aprendizado semissupervisionado baseado em redes heterogêneas bipartidas. Nessa proposta é demonstrado como representar as coleções de *revisões* para a realizar a Mineração de Opiniões Baseada em Aspectos usando redes heterogêneas bipartidas e como usar algoritmos de aprendizado semissupervisionados baseados em redes heterogêneas para realizar a extração de aspectos e a

classificação de sentimentos dos aspectos.

Foi realizada uma avaliação experimental que demonstrou que a ASPHN obteve melhores resultados do que os algoritmos tradicionais de aprendizado semissupervisionado ou de aprendizado supervisionado com base no modelo de espaço-vetorial. Portanto, esta proposta possibilita que o usuário rotule menos exemplos e obtenha desempenho igual ou melhor do que os algoritmos tradicionais de aprendizado de máquina que usam grandes conjuntos de dados rotulados. Com essa avaliação experimental pode-se comprovar duas hipóteses: (i) o uso de redes bipartidas para representação e (ii) o uso de aprendizado semissupervisionado melhoram o desempenho da extração de aspectos e da classificação de sentimentos dos aspectos com uso de menos dados rotulados. A validação foi comprovada com diferença estatística significativa dos resultados. O desenvolvimento e avaliação dessa proposta resultaram na publicação de um artigo em conferência nacional (MATSUNO *et al.*, 2015) e um artigo em periódico (MATSUNO *et al.*, 2016).

EXTRAÇÃO DE ASPECTOS NÃO SUPERVISIONADA E INDEPENDENTE DE DOMÍNIO PARA MINERAÇÃO DE OPINIÕES

4.1 Considerações Iniciais

Uma das tarefas desafiadoras para realizar Mineração de Opiniões baseada em Aspectos é a identificação (extração) dos termos (palavras ou conjunto de palavras) que representam os aspectos. Várias propostas para essa tarefa são encontradas na literatura, como as baseadas nos léxicos para o domínio de contexto, características linguísticas, relações sintáticas e modelos de tópicos (ZHANG; LIU, 2014). A principal desvantagem desses métodos é a dependência de domínio, por exemplo, a necessidade de especialistas para construir um léxico ou para definir as regras de extração de aspectos para diferentes idiomas e aplicações. Embora os métodos dependentes de domínio sejam interessantes, sua aplicação aos problemas do mundo real é limitada - especialmente em cenários dinâmicos como redes sociais e fóruns online - uma vez que muitos termos que são aspectos em um domínio podem não ser aspectos em outro. Por exemplo, o termo “fila de espera” é aspecto no domínio de *restaurante* e não é aspecto no domínio de *produto eletrônico*. Para tratar essa questão de dependência de domínio são utilizados métodos de aprendizado de máquina que não necessitam de um léxico predefinido ou da construção manual de regras. Nesta tese, foram investigados métodos de extração de aspectos com base em critérios estatísticos e estratégias propostas para eliminar o uso de léxicos predefinidos e a construção de regras de sintaxe supervisionada.

Considerando o desafio da extração de aspectos, foi realizada uma avaliação comparativa dos métodos de extração de aspectos em Mineração de Opiniões. Conforme revisão da literatura apresentada no Capítulo 2, este é o primeiro estudo experimental sobre métodos independentes de domínio para extração de aspectos. Foi realizada uma análise abrangente de quatro abordagens

principais para extração de aspectos: (i) baseada em frequência; (ii) baseada em combinações; (iii) baseada em medidas de associação, e (iv) baseada em modelos de tópicos. Além disso, é apresentada uma visão geral de avaliação da extração de aspectos, em particular, discutindo como medir a *Precisão* e a *Cobertura* para aspectos que são compostos por vários termos, bem como suas variações. Para suprir a falta de procedimentos padrões para avaliar os métodos de extração de aspectos, nesta tese, foi proposto um *framework* de avaliação e foram disponibilizadas as ferramentas computacionais (de código aberto) desenvolvidas para essa tarefa. A avaliação experimental foi realizada em sete coleções de revisões de referência e a discussão dos resultados é um guia promissor para determinar quais cenários cada método é mais apropriado.

O restante deste capítulo está organizado como segue. Na Seção 4.2 são apresentados os métodos de extração de aspectos, bem como as variações propostas para cenários independentes do domínio. Na Seção 4.3 é apresentado um *framework* proposto neste projeto de doutorado para avaliar métodos de extração de aspectos. Nesta seção também é apresentada uma breve descrição das ferramentas computacionais desenvolvidas para facilitar esse processo. Uma avaliação experimental é apresentada e discutida na Seção 4.4, com descrição das coleções de revisões, das avaliações e análise de desempenho de cada método de extração de aspectos. Por fim, na Seção 4.5 são apresentadas as considerações finais sobre o conteúdo deste capítulo.

4.2 Métodos para Extração de Aspectos Não-supervisionados e Independentes de Domínio

A principal questão sobre a extração de aspectos a partir de documentos textuais é como definir os termos (palavras-chave e expressões) que correspondem a um aspecto de uma determinada revisão ou opinião. Vários métodos, técnicas e estratégias foram propostos na literatura para tratar esta questão de pesquisa. Foram consideradas quatro tipos de abordagens: (i) baseada em frequência; (ii) baseada em combinações; (iii) baseada em medidas de associação, e (iv) baseada em tópicos. Essas quatro abordagens foram investigadas neste trabalho devido à grande popularidade na literatura que envolve a Mineração de Opiniões e à facilidade de propor variações para cenários independentes do domínio (uma vez que se baseiam em critérios estatísticos).

4.2.1 Métodos Baseados em Frequência

Os aspectos podem ser expressos nos textos por uma ou mais palavras. Embora tais palavras sejam de diferentes classes gramaticais, os aspectos geralmente são formados por substantivos e frases nominais, de acordo com Liu *et al.* (2013). Além disso, as pessoas são mais propensas a falar sobre os aspectos mais relevantes (de um produto ou serviço) em comentários de opinião (revisões). Isso sugere que, em geral, os aspectos devem ser substantivos frequentes.

Os métodos para extrair aspectos com base na frequência usam uma ferramenta para identificar as classes gramaticais de cada termo no texto. Essa ferramenta utilizada é um *Part-Of-Speech Tagger* (POS Tagger) que é um software que dado um texto escrito em língua natural, identifica a classe gramatical de cada palavra, como por exemplo verbos, adjetivos, substantivos e preposições. Posteriormente, os termos mais frequentes que pertencem a uma classe gramatical, geralmente substantivos, são selecionados como termo candidato a aspecto.

A extração de aspectos envolve regras gramaticais de substantivos frequentes, por exemplo, “se uma palavra frequente t , cuja classe gramatical é um substantivo singular, é modificada por uma palavra de opinião, então t é um aspecto” (LIU *et al.*, 2013). No entanto, a construção dessas regras gramaticais é uma tarefa dependente de domínio e requer um esforço humano significativo.

Nesta tese de doutorado é proposto o uso de um método baseado em frequência simples e eficaz para combinar unigramas frequentes e bigramas (por exemplo, duas palavras consecutivas). Primeiro, todos os termos (unigramas e bigramas) são extraídos e suas respectivas frequências são computadas. A partir deste conjunto, os termos compostos por um ou mais substantivos são selecionados como candidatos a aspectos. Para o processo de extração de aspectos, foram selecionados termos com uma frequência maior ou igual a um limite definido pelo usuário. Por exemplo, os termos “food” e “staff” são substantivos muito frequentes nas revisões de restaurantes e foram extraídos como aspectos. Do mesmo modo, essa estratégia permite extrair os aspectos “lens” e “flash” em revisões sobre câmeras digitais.

4.2.2 Métodos baseados em Combinações

Os aspectos nem sempre são compostos de substantivos frequentes. Existem diferentes técnicas de seleção que distinguem se um substantivo pode ser um aspecto ou não (HU; LIU, 2004; ZHANG; LIU, 2014). Neste caso, os métodos que dependem da frequência dos termos geralmente aplicam um conjunto de restrições sobre substantivos ou frases nominais de alta frequência para identificar os termos que podem ser aspectos. Esses métodos tem por base combinar os termos frequentes e os termos infrequentes são definidos por uma heurística. A abordagem de Hu e Liu (2004) é uma das primeiras propostas para extrair aspectos frequentes e infrequentes. Esta abordagem pode ser resumida nas seguintes etapas:

- (i) Preparação dos textos: *PoS tagging* de cada termo, a remoção de *stopwords* e *stemming* são realizadas nesta etapa.
- (ii) Extração de aspectos frequentes: os substantivos e frases nominais mais frequentes são extraídos como termos candidatos a aspecto.
- (iii) Extração de aspectos infrequentes: as palavras que estão próximas aos termos de opinião, definidos por um Léxico de Opinião, são extraídas como aspectos.

O Léxico de Opinião na etapa (iii) é um recurso dependente do domínio. Portanto, para obter independência de domínio, nesta tese de doutorado é proposta a seguinte adaptação: as palavras que são adjetivos serão consideradas como termos de opinião. Assim, os termos que são substantivos e estão associados a adjetivos foram extraídos como aspectos. De fato, essa proposta usa uma regra gramatical fixa, que geralmente é verdadeira e costuma ocorrer em qualquer domínio conforme estudos empíricos e observações discutidas nesta tese.

A vantagem deste método é que além dos aspectos frequentes, os termos infrequentes que são aspectos também são extraídos. Por exemplo, na coleção de revisões sobre câmera digital, os termos “4mp camera”, “metering option”, “auto setting”, “feel”, “low light focus”, “exposure control”, “manual mode”, “casing”, “memory card” e “performance” são infrequentes, mas são extraídos como aspectos usando a adaptação aqui proposta.

4.2.3 Métodos baseados em Medidas de Associação

É possível empregar medidas de associação (também chamadas de medidas de qualidade ou similaridade) para calcular relações entre termos (TAN; KUMAR; SRIVASTAVA, 2004). Essas medidas calculam a associação entre os termos t_i e t_j considerando as informações contidas na matriz de contingência apresentada na Tabela 4. A matriz de contingência apresenta os seguintes elementos:

- w_{ij} é a frequência em que os termos t_i e t_j ocorrem juntos;
- $w_{i,\neg j}$ é a frequência em que o termo t_i ocorre, sem o termo t_j ;
- $w_{\neg i,j}$ é a frequência em que o termo t_j ocorre, sem o termo t_i ;
- $w_{\neg i,\neg j}$ é a frequência em que ambos os termos t_i e t_j não ocorrem juntos.

Tabela 4 – Matriz de contingência para os termos t_i e t_j .

	t_j	$\neg t_j$	Total
t_i	w_{ij}	$w_{i,\neg j}$	w_i
$\neg t_i$	$w_{\neg i,j}$	$w_{\neg i,\neg j}$	$w_{\neg i}$
Total	w_j	$w_{\neg j}$	N

A última linha e a última coluna da matriz de contingência apresentam a frequência total de cada elemento. Por exemplo, w_i indica a frequência total do termo t_i , que também pode ser obtido por $w_i = w_{ij} + w_{i,\neg j}$. O elemento N indica o tamanho total da amostra.

Uma vez que as matrizes de contingência são obtidas para cada par de termos, então é possível calcular várias medidas de associação. As medidas de associação referem-se a uma grande variedade de coeficientes para medir a força estatística da relação entre as variáveis de interesse, na qual é expressa por um número chamado coeficiente de associação. Essas medidas

de associação estatística podem ser usadas para extrair n -gramas como candidatos a aspecto, em que esses n -gramas representam correlação significativa entre palavras ou outro fenômeno interessante.

Ao contrário dos trabalhos existentes que exploram as medidas de associação em conjunto com regras sintáticas supervisionadas (dependentes do domínio), neste projeto de doutorado é proposto explorar os n -gramas significativos (de acordo com cada medida de associação) diretamente como termo de aspecto.

Nesta proposta, foram consideradas três medidas populares: *Odds Ratio* (OR), *Pointwise Mutual Information* (PMI) and *Chi-Square* (χ^2). Essas medidas são descritas nas próximas subseções.

4.2.3.1 Odds Ratio (OR)

A medida Odds Ratio (OR) calcula a proporção do número de vezes que as palavras em um n -grama ocorrem juntas (ou não) ao número de vezes que as palavras ocorrem individualmente. Esta medida é definida na Equação 4.1, que calcula a força da relação entre os termos t_i e t_j . Se w_{i-j} e/ou w_{-ij} for zero, então cada valor zero será suavizado para evitar a divisão por zero.

$$OR(t_i, t_j) = \frac{w_{ij} \times w_{-i-j}}{w_{-ij} \times w_{i-j}} \quad (4.1)$$

A medida é frequentemente usada para seleção de atributos em aplicações de Mineração de Textos e, mais recentemente, foi investigada em tarefas de Mineração de Opiniões (DENG; LUO; YU, 2014; MEDHAT; HASSAN; KORASHY, 2014; RAVI; RAVI, 2015). Em geral, Odds Ratio é utilizada para determinar a importância dos termos de opinião para melhorar o desempenho da classificação de sentimentos (DENG; LUO; YU, 2014).

4.2.3.2 Pointwise Mutual Information (PMI)

A medida *Pointwise Mutual Information* (PMI) é uma medida da área da Teoria da Informação e quantifica a probabilidade de duas palavras ocorrerem juntas (distribuição conjunta) em relação às suas distribuições individuais.

Matematicamente, se dois termos t_i e t_j tiveram probabilidades $P(i) = \frac{w_i}{N}$ e $P(j) = \frac{w_j}{N}$, esta medida é definida pela Equação 4.2, em que a probabilidade $P(i, j) = \frac{w_{ij}}{N}$.

$$PMI(t_i, t_j) = \log \frac{P(i, j)}{P(i) \times P(j)} \quad (4.2)$$

PMI é usada na Mineração de Opiniões por diversas abordagens e em tarefas diferentes. Em Turney (2002), PMI é utilizada para calcular a orientação semântica de uma frase para identificar a polaridade do sentimento. Essa polaridade é obtida pela diferença do valor de PMI da

frase com um termo positivo e do valor de PMI com um termo negativo. Popescu e Etzioni (2005) avaliaram cada frase nominal descoberta calculando uma pontuação de PMI entre a frase e alguns discriminadores de meronímias¹ que são associados com a classe de entidade. Por exemplo, os discriminadores da meronímia para a classe da câmera são “da câmera”, “a câmera”, “a câmera vem com”, e assim por diante, que foram usados para encontrar componentes da câmera por busca na web. A ideia é que as frases descobertas (aspectos candidatos) que frequentemente ocorrem com esses discriminadores de parte-da-relação são provavelmente os aspectos (RAVI; RAVI, 2015; DENG; LUO; YU, 2014).

4.2.3.3 Chi-square (χ^2)

A medida estatística Chi-square (χ^2) mede a falta de independência entre os termos t_i e t_j e pode ser comparada à distribuição χ^2 com um grau de liberdade para julgar extremidade (YANG; PEDERSEN, 1997). A medida χ^2 é definida Equação 4.3 e obtém um valor zero se i e j forem independentes.

$$\chi^2(i, j) = \frac{N \times (w_{ij} \times w_{\neg i \neg j} - w_{\neg ij} \times w_{i \neg j})^2}{(w_{ij} + w_{\neg ij}) \times (w_{ij} + w_{i \neg j}) \times (w_{\neg i \neg j} + w_{\neg ij}) \times (w_{\neg i \neg j} + w_{i \neg j})} \quad (4.3)$$

Como PMI e Odds-Ratio, a medida estatística χ^2 também é usada para seleção de atributos na Mineração de Textos e foi investigada em tarefas de Mineração de Opiniões (DENG; LUO; YU, 2014; MEDHAT; HASSAN; KORASHY, 2014; RAVI; RAVI, 2015). Kaji e Kitsuregawa (2007) utilizam χ^2 para determinar a força da polaridade para cada palavra e frases para construir um léxico automaticamente. Mubarak, Adiwijaya e Aldhi (2017) empregaram χ^2 para selecionar um subconjunto de termos relevantes a serem usados na construção de *bag-of-words* para classificação de aspectos e sentimentos.

4.2.4 Métodos baseados em Modelos de Tópicos

A ideia básica de métodos baseados em modelos de tópicos é que os documentos são representados como misturas de tópicos latentes, nos quais os tópicos são associados à distribuição de palavras dentro de um vocabulário fixo. O algoritmo principal deste tipo de método é o *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003), um modelo generativo probabilístico para coleções de dados discretos, como o corpus de documentos. Na Figura 18 é apresentada uma representação gráfica desse modelo. Seguindo o formalismo do modelo de representação padrão, os vértices representam variáveis aleatórias e as arestas indicam possíveis dependências. Os vértices com preenchimento cinza são variáveis aleatórias observadas e os vértices sem preenchimento são variáveis aleatórias latentes. O retângulo em torno de grupos de variáveis aleatórias é um subconjunto que denota replicação. O retângulo externo representa o

¹ Uma meronímia denota um constituinte *parte de* ou *membro de* algo.

conjunto de revisões e o retângulo interno representa as palavras. R indica o número de revisões na coleção e N indica o número de palavras em cada revisão.

Figura 18 – Modelo gráfico do LDA

Fonte: Blei, Ng e Jordan (2003).

O processo gerador do LDA é descrito a seguir (BLEI; NG; JORDAN, 2003):

1. Amostra $\theta \approx Dir(\alpha)$.
2. Para cada palavra $w_n, n \in \{1, 2, 3, \dots, N\}$
 - a) Obter um exemplo de um tópico $z_n \approx Mult(\theta)$
 - b) Obter um exemplo de uma palavra $w_n \approx P(w_n | z_n; \beta)$ com distribuição multinomial condicionada ao tópico z_n .

Dada uma revisão, o principal problema de inferência é calcular a distribuição posterior das variáveis latentes (Equações 4.4 e 4.5) de cada documento.

$$P(z, \theta | w, \alpha, \beta) = \frac{P(z, w, \theta | \alpha, \beta)}{P(w | \alpha, \beta)} \quad (4.4)$$

$$P(z, w, \theta | \alpha, \beta) = P(\theta | \alpha) \prod_{n=1}^N [P(z_n | \theta) P(w_n | z_n, \beta)] \quad (4.5)$$

Uma vez que esta distribuição é intratável computacionalmente (devido ao acoplamento entre θ e β), a inferência variacional é usada para aproximar esta distribuição (BLEI; NG; JORDAN, 2003).

Alguns trabalhos aplicam o LDA em textos de opiniões para extrair tópicos como aspectos de produtos (BRODY; ELHADAD, 2010; ZHAO *et al.*, 2010; TITOV; MCDONALD, 2008a; TITOV; MCDONALD, 2008b). Em Brody e Elhadad (2010) e Zhao *et al.* (2010), o LDA foi usado em nível da sentença. Zhao *et al.* (2010) aperfeiçoam o modelo considerando diferentes distribuições de palavras para aspectos, ranqueamento e conceitos das palavras. Em Titov e McDonald (2008a) e Titov e McDonald (2008b) o modelo básico de LDA é aperfeiçoado considerando diferentes distribuições para os tópicos locais e globais. Isto é aplicado em nível de documento.

Neste projeto de doutorado são propostas duas estratégias para explorar tópicos gerados usando o LDA para extração de aspectos em cenários independentes de domínio: (i) os substantivos de cada tópico são considerados como aspectos e (ii) os substantivos de cada tópico são combinados com outros substantivos consecutivos no texto original para identificar aspectos compostos por duas ou mais palavras. Nesta segunda estratégia, as preposições entre dois substantivos são desconsideradas. Para ilustrar, considere que os dois tópicos latentes “[camera, flash, functions, leading, avancing]” e “[resolution, camera, indoor, canon, extends]” foram extraídos pelo método LDA da coleção de revisões de câmera digital. Observe que os tópicos latentes contêm termos que podem ser usados para aspectos como “[camera flash]” e “[camera resolution]”

4.3 *Framework* para avaliação de métodos de extração de aspectos para Mineração de Opiniões

O processo de avaliação experimental de Mineração de Opiniões baseada em Aspectos apresenta alguns desafios. Primeiramente, existem poucas coleções de textos com aspectos rotulados. Além disso, não existem padrões para: (i) a representação dos aspectos anotados; (ii) a avaliação de termos considerados como aspectos; e (iii) a análise estatística dos resultados experimentais. Em geral, os pesquisadores coletam e rotulam seus próprios conjuntos de dados e avaliam propostas com critérios baseados em medidas de *Precisão* e *Cobertura*. Esses critérios são bem aceitos, mas existem diferentes estratégias para calcular se o aspecto foi corretamente identificado ou extraído, especialmente os aspectos compostos por dois ou mais termos. Embora esses processos de avaliação sejam relevantes e válidos, diferentes propostas de avaliação dificultam a reprodução dos experimentos e a comparação efetiva entre os métodos. Devido a essas dificuldades, nesta tese de doutorado é proposto um *framework* para avaliar métodos de extração de aspectos como uma solução para essas desvantagens. Tal *framework* é apresentado na Figura 19. As subseções a seguir descrevem detalhadamente cada etapa dessa proposta.

4.3.1 *Coleções de Revisões*

Diferentes anotações e representações de dados foram propostas para os *datasets* compostos por revisões que são utilizados para avaliar os métodos de extração de aspectos. Nesta tese, foi compilado o que é comum entre essas propostas para disponibilizar essas coleções de revisões no *framework* proposto.

As revisões contendo as avaliações dos usuários podem ser coletadas de vários sites e com diferentes estruturas de dados. Posteriormente, elas são pré-processadas para separar a parte textual das revisões de seus metadados (outras informações, como usuário, localização e data de publicação). Para cada parte textual das revisões, os termos que indicam aspectos são rotulados. Opcionalmente, os respectivos termos ou expressões que representam a polaridade do sentimento

Figura 19 – Framework para avaliação de métodos de extração de aspectos para mineração de opiniões.

Fonte: Elaborada pelo autor.

relacionadas aos aspectos também são rotulados como “*palavras de sentimentos*” e “*palavras de opinião*”.

Para realizar este passo de pré-processamento da avaliação experimental, foi desenvolvida a ferramenta RotuLabic-AspectSentiment que permite que um grupo de usuários participe facilmente no processo de anotação usando um navegador da web. Na Figura 20a é apresentado um exemplo com uma revisão anotada usando essa ferramenta proposta. Na Figura 20b é apresentada a estrutura de dados de saída das anotações para a revisão “*The price is reasonable although the service is poor*”. É importante notar que o *framework* de avaliação proposto, também fornece a ferramenta computacional de anotação de aspecto, denominada por RotuLabic-AspectSentiment (open source), disponível no endereço <<http://sites.labic.icmc.usp.br/ipm/tese/cap6/rotulabic-as>>.

As coleções de revisões pré-processadas são organizadas em dois conjuntos: (i) Coleção de Revisões $R = \{r_1, r_2, \dots, r_n\}$, em que cada revisão r_i representa um texto original sem anotações e n indica o número total de revisões e ; (ii) Lista de Aspectos de Referência $A_{Ref} = \{a_1, a_2, \dots, a_k\}$ em que cada aspecto a_j representa um aspecto presente em R e k indica o número de aspectos. Esses dois conjuntos são usados como entrada para o próximo estágio da avaliação experimental.

4.3.2 Extração de Aspectos

Os métodos de extração de aspecto propostos na literatura apresentam seus resultados (saída) em diferentes formatos. Assim, após a execução de cada método de extração, sua

Figura 20 – Visão geral do processo de anotação de um documento revisão.

Document : restaurant_review_001

The **price** is reasonable although the **service** is poor.

Add Aspect

Select part of the text, then choose the polarity

NEGATIVE
NEUTRAL
POSITIVE

Add Hidden Aspect Add General Aspect

Remove All Aspects

Aspect	Polarity	Sentiment Words
price	POSITIVE	reasonable x
service	NEGATIVE	poor x

Previous Document Skip Document Finish

Return to the instructions

- (a) Exemplo de anotação para uma revisão usando a ferramenta Rotulabic-AspectSentiment. Neste caso, a sentença é apresentada para o usuário em um navegador web. O usuário seleciona as palavras que considera aspectos. Opcionalmente, o usuário pode também marcar as palavras que considera que expressam o sentimento.

```
<noteTerms>
  <aspectTerm term="price" polarity="positive" from="4" to="9"/>
  <sentimentTerm term="reasonable" polarity="positive" from="13" to="22"/>

  <aspectTerm term="service" polarity="negative" from="37" to="44"/>
  <sentimentTerm term="poor" polarity="negative" from="48" to="51"/>
</noteTerms>
```

- (b) Exemplo das anotações para a revisão “The price is reasonable although the service is poor”.

Fonte: Elaborada pelo autor.

respectiva saída é processada para uma lista de aspecto padronizada, denominada Lista de Aspectos Inferidos $A_{Inf} = \{a'_1, a'_2, \dots, a'_m\}$, em que cada aspecto a'_i representa um aspecto inferido pelo método e m indica o número total de aspectos inferidos. Cada método a ser avaliado gerará sua respectiva A_{Inf} com seus respectivos aspectos inferidos que serão comparados com a

Lista de Aspectos de Referência A_{Ref} obtida na etapa anterior.

4.3.3 Critérios de Avaliação

Uma questão bem crítica que também é considerada de diferentes modos na literatura está relacionada ao quão um método acerta (ou erra) um termo inferido como aspecto. Por exemplo, suponha que dada uma coleção de revisões R com um conjunto de aspectos de referência $A_{Ref} = \{a_1, a_2, \dots, a_n\}$ e que um método M infere que o termo a_{inf} é um aspecto, quando será considerado que o método M acerta? Existem diversas maneiras para considerar que o método acerta, por exemplo:

- se a_i é exatamente igual a um dos aspectos em A_{Ref} ; ou/e,
- se a_{inf} é um sinônimo de um ou mais aspectos em A_{Ref} ; ou/e,
- se a_{inf} é parte de algum dos aspectos em A_{Ref} (ou *substring*).

Nesta proposta de avaliação, para calcular os acertos de um método M , é realizada uma comparação entre os elementos da Lista de Aspectos Inferidos A_{Inf} com os elementos da Lista de Aspectos de Referência A_{Ref} . Na sequência, são considerados dois modos para pontuar os acertos: (i) correspondência exata e (ii) correspondência aproximada. No primeiro modo, somente quando o aspecto inferido x é exatamente igual a um dos aspectos em A_{Ref} . Já no segundo modo, é considerada uma pontuação relativa quando a_{inf} é parte de algum aspecto em A_{Ref} . Por exemplo, se $a_{inf} = \text{"vinho"}$ e existe um aspecto "vinho tinto" , então uma pontuação de 50% de acerto é considerada, isto é 0.5. Outro exemplo, é se existe um aspecto de referência $\text{"wine list selection"}$, então é considerada uma pontuação de acerto 33.33%, isto é 0.33.

Dados o conjunto de aspectos inferidos A_{Inf} e o conjunto de aspectos de referência A_{Ref} , para avaliar um método de extração de aspectos M foi considerada a medida F^1 (Equação 5.3) que é a média harmônica das medidas de *Precisão* (Equação 5.2a) e de *Cobertura* (Equação 5.2b), em que:

- VP (Verdadeiro Positivo) indica o número de termos inferidos como aspectos que foram corretamente extraídos, ou seja $a_{inf} \in A_{Ref}$;
- FP (Falso Positivo) indica o número de termos inferidos como aspectos que não são aspectos reais, ou seja $a_{inf} \notin A_{Ref}$; e
- FN (Falso Negativo) indica o número de termos que são aspectos e que não foram inferidos como aspectos, ou seja $a_{ref} \notin A_{Inf}$.

$$Precisão = \frac{VP}{VP + FP}, \quad (4.6a)$$

$$Cobertura = \frac{VP}{VP + FN}, \quad (4.6b)$$

$$F^1 = 2 * \frac{Precisão * Cobertura}{Precisão + Cobertura}. \quad (4.7)$$

As medidas de avaliação são ponderadas de acordo com o tipo de correspondência definido (exato ou aproximado). A medida *Precisão* retorna a porcentagem de termos corretamente classificados como aspectos, considerando todos os termos classificados como aspectos. A medida *Cobertura* retorna a porcentagem de termos corretamente classificados como aspectos, considerando todos os aspectos que são realmente aspectos. Para processamento desses valores, foi desenvolvida uma ferramenta web disponível em <http://sites.labic.icmc.usp.br/ipm/tese/cap4/evaluation>, em que o usuário informa A_{Inf} e A_{Ref} , define o modo de correspondência (exata ou aproximada) para pontuar os acertos e submete para que seja calculado o valor das medidas de *Precisão*, *Cobertura* e F^1 .

4.3.4 Avaliação Comparativa

A comparação de desempenho de dois ou mais métodos é analisada considerando um teste de significância estatística com determinado nível de confiança (DEMSAR, 2006; GARCIA; HERRERA, 2008). Neste framework, para realizar a avaliação comparativa, além da comparação direta entre as medidas obtidas individualmente por cada método, são utilizados testes estatísticos não-paramétricos. Esses testes estatísticos não assumem uma distribuição prévia sobre os valores dos resultados experimentais. Os testes estatísticos utilizados são:

- Teste de Wilcoxon para comparação entre dois métodos de extração de aspectos em múltiplos *datasets* (DEMSAR, 2006).
- Teste de Friedman para comparação de três ou mais métodos de extração de aspectos em múltiplos *datasets* (GARCIA; HERRERA, 2008).

Os resultados da avaliação experimental, bem como a análise estatística, são apresentados no *framework* proposto por meio de tabelas, gráficos e diagramas de diferença crítica (análise de significância estatística). Na próxima seção, é apresentada uma comparação geral dos métodos independentes do domínio de extração de aspectos (apresentados na Seção 4.2) usando as ferramentas disponíveis no *framework* proposto.

4.4 Avaliação Experimental

Para avaliação experimental foram utilizadas 7 coleções de revisões disponíveis na literatura que possuem aspectos rotulados. Essas coleções são compostas por textos de comentários e

revisões em língua natural no idioma inglês que foram extraídas de sites de e-commerce e da competição SemEval². Na Tabela 9 é apresentada uma visão geral dessas coleções de revisões. As coleções D1 a D5 foram obtidas em Hu e Liu (2004) e D6 e D7 em Pontiki *et al.* (2014). Essas coleções e suas anotações foram processadas conforme os padrões definidos na subseção 4.3.1. As descrições e as Listas de Aspectos de Referência de cada coleção e outras informações estão disponíveis em <http://sites.labic.icmc.usp.br/ipm/tese/cap4/dataset>,

Tabela 5 – Descrição das coleções de revisões utilizadas no experimento

ID	Domínio	#Revisões	#Aspectos	#TotalTermos	#MédiaTermos
D1	Câmera Digital 1	45	68	1045	54.46
D2	Câmera Digital 2	34	47	734	48.74
D3	Telefone Celular	41	78	924	51.70
D4	MP3 Player	95	126	1614	63.95
D5	DVD Player	99	82	1074	33.02
D6	Laptop	3045	955	2559	5.26
D7	Restaurante	3045	1219	3083	5.34

4.4.1 Configuração Experimental

Os seguintes passos foram realizados para todos os métodos de extração de aspectos:

- (1) Os textos foram pré-processados para identificação da marcação *Part-of-Speech* de cada termo. Isso foi realizado por meio da ferramenta Stanford POS Tagger (MANNING *et al.*, 2014).
- (2) Remoção das *Stopwords*.
- (3) Os n -gramas foram extraídos, considerando $0 < n \leq 2$ e janela $0 < k \leq 2$.

Os métodos de extração de aspectos e suas respectivas definições de parâmetros utilizados na avaliação experimental são descritos a seguir.

– Métodos baseados em Frequência

- **Unigramas Frequentes que são substantivos (M1)**: os termos compostos por apenas um substantivo (ou frase nominal) e que são frequentes na coleção de revisões são extraídos como aspectos. Foram considerados os seguintes valores para frequência $f = \{1, 3, 5, 10\}$.
- **Bigramas Frequentes compostos por algum substantivo (M2)**: os termos compostos por duas palavras que são frequentes na coleção de revisões e que são compostos

² SemEval: <http://alt.qcri.org/semeval2016/>

por pelo menos uma palavra que seja um substantivo (ou frase nominal) são extraídos como aspectos. Foram considerados os seguintes valores para frequência $f = \{1, 3, 5, 10\}$.

- ***Bigramas Frequentes compostos somente por substantivos (M3)***: os termos compostos por duas palavras que são frequentes na coleção de revisões e que são compostos por apenas palavras que sejam substantivos (ou frases nominais) são extraídos como aspectos. Foram considerados os seguintes valores para frequência $f = \{1, 3, 5, 10\}$.
 - ***Unigramas Frequentes que são substantivos + Bigramas Frequentes compostos por algum termo que é substantivo (M4)***: os termos compostos por única palavra que são substantivos frequentes e os termos compostos por duas palavras que são frequentes e compostos por pelo menos uma palavra que é um substantivo (ou frase nominal) na coleção de revisões são extraídos como aspectos. Foram considerados os seguintes valores para frequência $f = \{1, 3, 5, 10\}$.
 - ***Unigramas Frequentes que são substantivos + Bigramas Frequentes compostos por todos os termos que são substantivos (M5)***: os termos compostos por apenas palavras que são substantivos frequentes e os termos compostos por duas palavras que são frequentes e compostos por apenas palavras que são substantivos na coleção de revisões são extraídos como aspectos. Foram considerados os seguintes valores para frequência $f = \{1, 3, 5, 10\}$.
- **Substantivos Frequentes + Substantivos Infrequentes+Adjetivos (M6)**: com base em [Hu e Liu \(2004\)](#) e nas adaptações propostas na Seção 2, os aspectos são extraídos considerando as seguintes regras: (i) unigramas frequentes que são substantivos, (ii) bigramas frequentes compostos por substantivos e (iii) termos que são infrequentes que são substantivos e associados a termos que são adjetivos.
- **Medidas de Associação (M7, M8, M9)**: os unigramas frequentes que são substantivos maior ou igual a f são extraídos como aspectos. Considerando uma medida de associação Y , os bigramas que possuem um valor maior ou igual ao limite l são extraídos como aspectos. Neste experimento foram utilizadas as medidas: Odds Ratio, χ^2 e PMI. Foram consideradas frequência $f = \{5\}$ e $l = \{m - 2d, m - 1d, m, m + 1d, m + 2d\}$ em que m é a média aritmética e d é o desvio padrão dos valores obtidos das medidas de associação utilizadas.
- **Baseada em Modelo de Tópicos**
- ***Modelo de Tópicos- Téc.I (M10)***: os tópicos obtidos pelo LDA são representados por uma lista das *top-k* palavras mais representativas. Assim, os substantivos pertencentes à lista *top-k* são extraídos como aspectos.

- **Modelo de Tópicos - Téc.II (M11)**: além dos aspectos extraídos como no item anterior, os termos compostos pelos substantivos apresentados no tópico e que também estão associados a outro substantivo no texto original são extraídos como aspectos.

Considerando essas configurações, os métodos foram utilizados para inferir uma lista de aspectos para cada coleção de revisões. Cada Lista de Aspectos Inferidos foi comparada à Lista de Aspectos de Referência e foram computados os valores de *Precisão*, *Cobertura* e F^1 .

4.4.2 Análise e Discussão dos Resultados

A análise dos resultados experimentais considera a eficácia (baseada nos resultados de F^1) dos métodos de extração de aspectos considerando os dois tipos de correspondências com as respectivas listas de aspectos de referências: (i) correspondência aproximada e (ii) correspondência exata. Para cada método foi considerado o melhor resultado dentre todas as possibilidades de configurações. Nas Tabelas 6 e 7 são apresentados os resultados de F^1 obtidos para cada método de extração em cada coleção de revisões conforme correspondência exata e aproximada respectivamente. Nessas tabelas, para facilitar a visualização dos dados, os melhores resultados por coleção estão em negrito. Os resultados obtidos considerando outras medidas de avaliação (*Precisão* e *Cobertura*) estão disponíveis no Anpêndice A

De acordo com o conhecido teorema de “*No Free Lunch*”, observa-se que nenhum método pode superar qualquer outro método quando o desempenho (F^1) é amortizado considerando todos os conjuntos de dados. No entanto, alguns métodos são mais apropriados de acordo com certas características das coleções de revisões.

Os métodos baseados em frequência apresentaram resultados gerais promissores, especialmente em conjuntos de dados pequenos (*D1*-Câmera digital 1, *D2*-Câmera digital 2, *D3*-Telefone celular, *D4*-MP3 Player e *D5*-DVD Player). Em particular, os métodos que exploram a frequência de unigramas que são substantivos combinados com bigramas frequentes que também contêm substantivos (*M4* -comb-1gram-2gram-some-noun e *M5* -comb-1gram-2gram-all-noun) apresentaram resultados melhores que os demais, porém, insatisfatórios em vários cenários. De fato, esses métodos exploram a premissa de que os aspectos tendem a ser compostos por substantivos, que ocorre independentemente do domínio.

Métodos baseados em medidas de associação apresentaram desempenho médio em todos os conjuntos de dados e também observou-se que esses métodos representam uma estratégia estável para extração de aspectos. Essa estabilidade é interessante para problemas independentes do domínio. Embora as três medidas de associação analisadas (*OR*, χ^2 , *PMI*) tenham inferido listas de aspectos muito semelhantes, enfatiza-se que o método *PMI* é um pouco superior. Além disso, o *PMI* não faz qualquer pressuposto sobre a distribuição de dados — que é um recurso que pode afetar o método χ^2 .

Métodos baseados em modelos de tópicos foram indicados como uma boa alternativa para a extração de aspectos na literatura. No entanto, ressalta-se que esses métodos são úteis apenas para conjuntos de dados que possuem uma grande quantidade de textos. Esse recurso é inerente ao processo de aprendizagem do algoritmo LDA, que depende do modelo estatístico generativo a partir de dados textuais. Assim, observa-se que o desempenho de métodos baseados em LDA para extração de aspectos foi satisfatório apenas para as coleções de revisões *D6-laptop* e *D7-restaurante*, que possuem características adequadas para este processo de aprendizado de máquina não supervisionado.

Foi realizada uma análise estatística para comparar a significância do desempenho da extração de aspectos considerando a medida F^1 e correspondência exata e aproximada das melhores configurações de cada método (análise do melhor caso). Foi utilizado o teste Friedman não-paramétrico com testes pós-hoc Nemenyi para comparar múltiplos métodos em vários conjuntos de dados. Este procedimento é considerado uma estratégia mais consistente para comparar vários métodos (DEMSAR, 2006; GARCIA; HERRERA, 2008). Os resultados foram analisados com 95% de nível de confiança ($\alpha = 0,05$).

Nas Figuras 21 e 22 são apresentadas análises por meio de gráficos do teste estatístico, denominado de diagrama de diferença crítica. Nesses diagramas, os métodos de extração de aspectos são classificados de acordo com sua posição no ranking geral de desempenho de classificação. Os métodos são conectados por uma linha se não houver diferença estatisticamente significativa entre eles.

Tabela 6 – Resultados da medida F^1 para cada método em cada coleção de revisões considerando correspondência exata

Métodos		Coleções de Revisões						
Descrição	Id	D1	D2	D3	D4	D5	D6	D7
freq-1gram-noun	M1	0.25	0.21	0.32	0.28	0.31	0.22	0.26
freq-2gram-some-noun	M2	0.14	0.17	0.12	0.14	0.10	0.21	0.21
freq-2gram-all-noun	M3	0.14	0.17	0.11	0.12	0.06	0.25	0.25
comb-freq-some-noun	M4	0.29	0.24	0.37	0.32	0.34	0.25	0.30
comb-freq-all-noun	M5	0.29	0.23	0.36	0.30	0.32	0.28	0.31
comb-freq-adj	M6	0.19	0.17	0.26	0.20	0.26	0.24	0.31
ODDS	M7	0.24	0.22	0.36	0.29	0.33	0.25	0.28
χ^2 (CHI2)	M8	0.25	0.19	0.32	0.28	0.31	0.25	0.28
PMI	M9	0.25	0.19	0.32	0.28	0.31	0.26	0.29
LDA-Tec. I	M10	0.08	0.11	0.14	0.11	0.11	0.17	0.21
LDA-Tec. II	M11	0.09	0.11	0.13	0.09	0.10	0.24	0.28

Considerando o ranking médio, os métodos que obtiveram o melhor desempenho são aqueles baseados em frequências que combinam unigramas e bigramas que contêm termos que são substantivos. Os métodos que consideram apenas unigramas ou apenas bigramas apresenta-

Figura 21 – Diagrama de diferença crítica para medida F^1 considerando correspondência exata

Fonte: Elaborada pelo autor.

Tabela 7 – Resultados da medida F^1 para cada método em cada coleção de revisões considerando correspondência aproximada

Método		Coleções de Revisões						
Descrição	Id	D1	D2	D3	D4	D5	D6	D7
freq-1gram-noun	M1	0.34	0.30	0.39	0.37	0.36	0.33	0.38
freq-2gram-some-noun	M2	0.18	0.25	0.19	0.21	0.16	0.30	0.29
freq-2gram-all-noun	M3	0.18	0.23	0.15	0.19	0.13	0.41	0.39
comb-freq-some-noun	M4	0.36	0.33	0.43	0.39	0.39	0.37	0.43
comb-freq-all-noun	M5	0.37	0.33	0.42	0.38	0.38	0.38	0.41
comb-freq-noun-adj	M6	0.24	0.23	0.32	0.24	0.30	0.33	0.40
ODDS	M7	0.32	0.31	0.42	0.37	0.38	0.36	0.37
CHI2	M8	0.34	0.30	0.39	0.33	0.36	0.36	0.38
PMI	M9	0.34	0.30	0.39	0.37	0.36	0.39	0.44
LDA-Tec. I	M10	0.11	0.17	0.16	0.16	0.15	0.26	0.28
LDA-Tec. II	M11	0.10	0.16	0.16	0.12	0.14	0.38	0.43

Figura 22 – Diagrama de diferença crítica para medida F^1 considerando correspondência aproximada

Fonte: Elaborada pelo autor.

ram os piores resultados. No entanto, não há significância estatística na comparação dos resultados obtidos pelos métodos que combinam unigramas e bigramas com aqueles que consideram apenas unigramas. Os métodos $M4$ (comb-1gram-2gram-some-noun) e $M5$ (comb-1gram-2gram-

all-noun) apresentam diferenças estatisticamente significativas em relação aos métodos baseados em LDA ($M10$ e $M11$) e métodos com base apenas em bigramas frequentes ($M2$ e $M3$).

Além da análise estatística dos resultados experimentais, também foi realizada uma análise de significância prática dos métodos de extração de aspectos, ou seja, identificando cenários em que cada método pode ser usado de forma mais apropriada. Na sequência, é apresentado um exemplo dessa análise prática.

Embora no ranking médio, os métodos baseados em modelos de tópicos terem apresentado resultados inferiores, esses métodos obtiveram resultados bons nas coleções $D6$ e $D7$ que possuem uma quantidade de revisões 30 vezes maior que nas demais coleções. Isso nos dá um indicativo que o uso desses métodos pode ser mais apropriado quando se tem uma grande quantidade de revisões. Para ilustrar é apresentado um exemplo.

Sejam r_1 , r_2 e r_3 três revisões do domínio do restaurante usadas na avaliação, em que os resultados do pré-processamento linguístico (POS) são apresentados respectivamente em Figuras 23a, 23b e 23c.

- r_1 : “We didn’t get **drink refills** and she didn’t even offer us the option of **dessert**.”
- r_2 : “We only ordered **desserts** and **drinks**, but no **refills** were offered.”
- r_3 : “No **refills** on **fountain drinks**, though.”

Figura 23 – Revisões com marcação Part-of Speech (POS)

PRP VBD RB VB VB NNS CC PRP VBD RB RB VB PRP DT NN IN NN .
We did n't get drink refills and she did n't even offer us the option of dessert.

(a) r_1 with POS tagger

PRP RB VBD NNS CC NNS , CC DT NNS VBD VBN .
We only ordered desserts and drinks , but no refills were offered .

(b) r_2 with POS tagger

DT NNS IN NN NNS , RB .
No refills on fountain drinks , though .

(c) r_3 with POS tagger

Fonte: Elaborada pelo autor.

Os aspectos na Lista de Aspectos de Referência são os termos: “**dessert**”, “**drink**”, “**drink refills**” e “**fountain drinks**”. Na Tabela 8 são apresentadas a relação entre os métodos e os aspectos extraídos para essas revisões (r_1 , r_2 e r_3), em que as linhas representam os

métodos e as colunas representam os aspectos. Cada célula c_{ij} é marcada (✓) quando o método da linha i extraiu o aspecto da coluna j . Observa-se que os substantivos frequentes “**drink**” e “**dessert**” foram extraídos como aspectos por todos os métodos (exceto os métodos que extraem apenas bigramas).

O aspecto “**drink refills**” foi extraído somente pelo método LDA usando a segunda estratégia. Os outros métodos excluíram este termo, pois é um bigrama composto por um substantivo e verbo. Em contraste, os métodos baseados em tópicos não extraíram o termo “**fountain refills**” como aspecto. Este aspecto (“**fountain refills**”) não foi inferido pelos métodos $M4$ (comb-1gram-2gram-some-noun) e $M6$ (comb-1gram-2gram-noun -adj) devido aos limiares definidos. Este resultado ilustra como esses métodos obtêm resultados práticos diferentes de acordo com seus parâmetros.

Tabela 8 – Exemplo de aspectos por método de extração

Methods		Aspects			
Description	ID	dessert	drink	drink refills	fountain drinks
freq-1gram-noun	M1	✓	✓		
freq-2gram-some-noun	M2				✓
freq-2gram-all-noun	M3				✓
comb-freq-some-noun	M4	✓	✓		
comb-freq-all-noun	M5	✓	✓		✓
comb-freq-noun-adj	M6	✓	✓		
ODDS	M7	✓	✓		✓
χ^2 (CHI2)	M8	✓	✓		✓
PMI	M9	✓	✓		✓
LDA-Tec. I	M10	✓	✓		
LDA-Tec. II	M11	✓	✓	✓	

4.5 Considerações Finais

Embora a Mineração de Opiniões baseada em Aspectos tenha recebido grande atenção na literatura nos últimos anos, ainda existem algumas limitações quanto à descrição detalhada dos métodos existentes, critérios de avaliação e técnicas de comparação. Em particular, os métodos independentes de domínio têm sido um desafio pouco explorado mesmo com os benefícios reconhecidos das aplicações baseadas em Mineração de Opiniões baseada em Aspectos no mundo real. Neste capítulo foi apresentado um extenso estudo sobre esta área em que (i) descreveu-se os trabalhos existentes e foram propostas algumas variações para possibilitar a independência de domínio, (ii) foi proposto um *framework* geral para avaliação e comparação de métodos de extração de aspectos, e (iii) foram desenvolvidos os métodos e ferramentas de código aberto, bem como foram disponibilizadas as coleções de revisões pré-processadas para apoiar pesquisadores na área de Mineração de Opiniões.

Os resultados da avaliação experimental realizada demonstraram que os métodos baseados em frequência são úteis para conjuntos de dados com poucas revisões. Além disso, os métodos baseados em medidas de associação obtêm resultados medianos, mas com o diferencial de ser estável (em relação à medida F^1) em vários domínios. Os métodos baseados em LDA (tópicos latentes) são alternativas promissoras para coleções com uma grande quantidade de revisões.

O *framework* disponibilizado usa critérios estatísticos e medidas apropriadas para comparar os métodos e para avaliar as Listas de Aspectos Inferidos. Deste modo, novos estudos podem ser comparados utilizando o mesmo protocolo, bem como prover meios para reprodutibilidade dos resultados experimentais. As adaptações propostas para independência de domínio dos métodos de extração de aspectos e o *framework* proposto resultaram em artigo (MATSUNO *et al.*, 2018) submetido para periódico.

Como principais trabalhos futuros pretende-se realizar (i) a adição de novas coleções de revisões de referência pré-processadas ao repositório (disponível publicamente para pesquisadores), e (ii) a avaliação de novos métodos de extração de aspectos para Mineração de Opiniões. Finalmente, novos estudos serão realizados para proporar uma estratégia baseada em consenso que visa combinar resultados de diferentes métodos não supervisionados.

EXTRAÇÃO DE ASPECTOS PARA MINERAÇÃO DE OPINIÕES USANDO APRENDIZADO POR TRANSFERÊNCIA ENTRE DOMÍNIOS

5.1 Considerações Iniciais

Mineração de opiniões tornou-se muito popular para automatizar a extração de conhecimento de revisões sobre produtos e serviços (LIU, 2015). Em geral, o objetivo é determinar a opinião do consumidor em algum tópico através da polaridade do sentimento expressada no texto das revisões (*por exemplo, positivo, negativo ou neutro*) (LIU; ZHANG, 2012). Enquanto os primeiros estudos sobre a mineração de opiniões tentaram extrair a polaridade de todo o texto das revisões (mineração de opiniões em nível de documento) ou das sentenças das revisões (mineração de opiniões em nível da sentença), estudos mais recentes investigam a mineração de opiniões baseada em aspectos (FELDMAN, 2013; LAU; LI; LIAO, 2014; AKHTAR *et al.*, 2017; SCHOUTEN; FRASINCAR, 2016). Neste caso, a opinião do consumidor é analisada em alto nível de detalhe, em que, a opinião sobre cada característica do produto ou serviço é considerada (por exemplo, qualidade das fotos de uma câmera, peso do notebook, preço da comida). Na realidade, as tarefas de Mineração de Opiniões baseada em Aspectos são consideradas mais complexas (MUKHERJEE; LIU, 2012; WANG *et al.*, 2014; MATSUNO *et al.*, 2016) porque envolvem uma questão desafiadora: como extrair automaticamente os aspectos de cada produto ou serviço das revisões feitas pelo consumidor?

Os trabalhos existentes para extração de aspectos que são baseados em padrões das características linguísticas e utilizam técnicas de processamento de língua natural identificar classes gramaticais e relações sintáticas nos textos das revisões (RANA; CHEAH, 2016). Nesses

trabalhos, um conjunto de aspectos previamente rotulados (conjunto de treinamento) é utilizado em uma tarefa de aprendizado supervisionado, que induz um modelo de classificação. Em seguida, os aspectos são extraídos das novas revisões de acordo com os atributos linguísticos apresentados ao classificador. Por exemplo, um classificador baseado em regras que aprendeu o seguinte padrão “SE uma palavra w é um substantivo E w está em uma relação com um verbo v ENTÃO w é extraído como um aspecto.” É capaz de identificar o aspecto chamado “camera” da revisão apresentada na Figura 24, onde w = “camera” e t = “is”. A relação entre w e v é definida pelo adjetivo “nice”.

Figura 24 – Part-of-Speech (PoS) e relações de dependência sintática extraídas da sentença “The camera is nice”.

O desempenho de sistemas de extração de aspectos depende da quantidade de aspectos rotulados (RANA; CHEAH, 2016; SCHOUTEN; FRASINCAR, 2016). O processo de rotulação requer intenso esforço humano, bem como o tratamento das características das fontes de dados. Por exemplo, as revisões coletadas das redes sociais podem conter textos curtos com linguagem informal (*por exemplo, gírias e jargões*), enquanto fóruns especializados podem conter linguagem mais técnica e textos mais longos. No entanto, é possível aproveitar os aspectos rotulados em um domínio para suportar a extração de aspecto de outro domínio, em que não há aspectos rotulados (HU; LIU, 2004; ZHANG *et al.*, 2016; RANA; CHEAH, 2017). Esta estratégia é conhecida como *Cross Domain Transfer Learning* e foi investigada como uma solução promissora para problemas em que rotulação frequente de dados é impossível ou custosa (PAN; YANG, 2010; SCHOUTEN; FRASINCAR, 2016).

Neste capítulo, é apresentada uma abordagem de extração de aspectos entre domínios baseada na aprendizagem transdutiva denominada por CD-ALPHN (*Cross-Domain Aspect Label Propagation through Heterogeneous Networks*). As abordagens de *cross-domain* existentes induzem os classificadores a partir de aspectos rotulados no domínio fonte e, em seguida, aplicam esses classificadores no domínio alvo, ou seja, dois estágios separados que podem causar inconsistência devido a diferentes espaços de atributos ou diferentes distribuições de probabilidade marginais. Para lidar com este problema, nesta tese de doutorado é proposto um processo de aprendizado transdutivo a partir de um espaço unificado com os atributos dos domínios fonte e alvo. Para isso foi proposta uma representação baseada em grafos, em que os aspectos rotulados do domínio fonte, os atributos linguísticos e os candidatos a aspectos do domínio alvo são todos mapeados como vértices em uma rede heterogênea. Durante o processo de aprendizado transdutivo proposto, os vértices rotulados e não rotulados são explorados em um processo de propagação de rótulos. Nesse caso, as informações do rótulo do domínio fonte (aspectos rotulados) são propagadas para os vértices dos atributos linguísticos. Em seguida, as

informações de rótulos dos atributos linguísticos são propagados para os vértices dos candidatos a aspecto; que depois propagam a informação de rótulo de volta para a rede. Este processo de propagação continua até a convergência, ou seja, quando a informação do rótulo dos vértices da rede não é alterada significativamente. Assim, o processo de aprendizado de transferência entre domínios é realizado de forma mais natural, com a grande vantagem de usar uma formalização matemática e garantias teóricas de convergência para soluções sub-ótimas.

Para comparar a abordagem proposta com outras abordagens do estado da arte foi realizada uma avaliação experimental com sete conjuntos de dados de referência. Os resultados experimentais indicam que a abordagem proposta é muito competitiva, especialmente em cenários onde há uma alta inconsistência (diferença entre os domínios) entre os domínios fonte e alvo que, em geral, representa o cenário mais comum em aplicações do mundo real. Além disso, também foi apresentada uma visão geral de como essa abordagem pode ser usada em problemas práticos por um Sistema de Inteligência Analítica para suportar o processo de tomada de decisão.

5.2 Transferência de Aprendizado entre Domínios

A transferência de aprendizado entre domínios visa utilizar dados rotulados de outros domínios para suportar tarefas de aprendizado em domínios sem dados rotulados (PAN; YANG, 2010). O domínio com dados rotulados é chamado de domínio fonte, enquanto o domínio sem dados rotulados é chamado de domínio alvo. Em aplicações do mundo real, diferentes domínios têm diferenças no espaço de atributos, bem como diferentes distribuições de dados subjacentes, exigindo assim abordagens de aprendizado apropriadas para tratar essas limitações (LONG *et al.*, 2014).

A maioria dos trabalhos existentes de transferência de aprendizado entre domínios são conhecidos como aprendizado por transferência indutiva (*inductive transfer learning*) (PAN; YANG, 2010; LU *et al.*, 2015). Nesse caso, o objetivo é identificar atributos que são úteis no domínio fonte para serem aplicados no domínio alvo, obtendo assim um espaço de atributos compartilhado para o problema de transferência de aprendizado entre domínios. Em alguns estudos, os melhores atributos para o domínio alvo são selecionados do domínio fonte e refletidos para melhorar a classificação final (CHEN; MUKHERJEE; LIU, 2014; LU *et al.*, 2015). O aprendizado por transferência indutiva também é apresentado como uma transferência de aprendizado entre domínios em duas etapas (WU; TAN, 2011; RANA; CHEAH, 2017). Na primeira etapa, um conjunto de atributos comuns a ambos os domínios são extraídos e, na segunda etapa, são selecionados atributos úteis específicos para o domínio alvo (WU; TAN, 2011; RANA; CHEAH, 2017).

Diferentes estratégias de aprendizado para modelos de classificação foram propostas para *cross-domain transfer learning* (LU *et al.*, 2015). A estratégia mais comum (com resultados promissores) é treinar classificadores bem conhecidos, como *Multilayer Perceptrons*, *SVM*, *Naïve-*

Bayes e *kNN*, no espaço de atributos que é comum a ambos os domínios. Outras estratégias são baseadas no consenso de vários classificadores, considerando a amostragem de dados diferentes tanto do domínio fonte quanto do domínio alvo (LUO *et al.*, 2008; ZHUANG *et al.*, 2010). Além disso, existem estratégias que usam o aprendizado ativo para melhorar o processo de transferência de aprendizado. Essas estratégias exigem anotação humana quando um conjunto de classificadores discorda sobre a classe de alguma instância no domínio alvo (LI *et al.*, 2013; WU; HUANG; YAN, 2017).

A grande maioria dos trabalhos existentes concentra-se principalmente na transferência de polaridade de sentimento entre diferentes domínios, em que o uso de transferência de aprendizado entre domínios para extração de aspectos é pouco explorado (AL-MOSLMI *et al.*, 2017). Algumas iniciativas recentes exploram padrões de atributos linguísticos extraídos dos aspectos para obter um espaço de atributos comum entre os dois domínios (ZHANG *et al.*, 2016; RANA; CHEAH, 2017). No entanto, esses trabalhos também utilizam a abordagem de transferência de aprendizado indutivo em duas etapas. Por outro lado, a hipótese desta tese é que o processo de transferência de aprendizado pode ser melhorado explorando uma representação que unifica os diferentes espaços de atributos de cada domínio, utilizando padrões de atributos linguísticos como uma ponte entre os domínios fonte e alvo. Nesse sentido, foi proposta uma representação unificada baseada em redes heterogêneas e o uso de aprendizado transdutivo em redes para a extração de aspectos entre domínios.

5.3 Aprendizado Transdutivo para Extração de Aspectos entre Domínios

A abordagem proposta nesta tese é denominada por Propagação de Rótulo de Aspecto entre Domínios por meio de Redes Heterogêneas (*Cross-Domain Aspect Label Propagation through Heterogeneous Networks* - CD-ALPHN). Essa técnica será apresentada em duas etapas. Na primeira etapa é descrita a construção da rede heterogênea e sua utilização como uma representação unificada entre os domínios fonte e alvo. Essa representação será apresentada na Subsubseção 5.3.1. Na segunda etapa, será apresentado o algoritmo proposto para o aprendizado transdutivo. Esse algoritmo utiliza a propagação de rótulos na rede heterogênea para transferir a informação dos rótulos dos aspectos rotulados do domínio fonte para os aspectos candidatos do domínio alvo, usando padrões das características linguísticas como ponte. Esse algoritmo será descrito na Subsubseção 5.3.2.

5.3.1 Representação de Dados entre Domínios para Aprendizado em Redes Heterogêneas

Seja R^S o conjunto de revisões do domínio fonte, em que há um conjunto de aspectos rotulados A^S . Seja R^T o conjunto de revisões do domínio alvo, em que há um conjunto de candidatos a aspectos A^T . Os candidatos a aspectos selecionados em A^T são compostos de todas as palavras extraídas das revisões R^T , após a remoção das *stopwords* (isto é, excluindo pronomes, preposições, advérbios e artigos).

Seja L todos os tipos de atributos linguísticos identificados (por exemplo, substantivos, verbos, adjetivos), que dependem da linguagem das revisões e das ferramentas linguísticas (por exemplo, *Part-of-Speech* (PoS) — como mostrado na Figura 24). Alguns atributos linguísticos são comuns aos domínios fonte e alvo, assim, eles podem ser interpretados como uma ponte entre os dois domínios.

Nesta tese é proposta uma representação de dados para aprendizado *cross-domain* com base em uma rede heterogênea $G = (V, E, W)$, em que $V = \{A^S \cup A^T \cup L\}$, ou seja, a rede heterogênea é formada por três tipos diferentes de vértices. O conjunto de arestas E conecta os atributos linguísticos aos conjuntos de aspectos fonte (A^S) e alvo (A^T). Assim, se dois aspectos participam da mesma relação sintática, os dois vértices que representam esses aspectos serão conectados ao vértice representando a característica linguística dessa relação sintática. De fato, as palavras de um aspecto costumam aparecer em muitas revisões e, portanto, o conjunto W indica o peso das arestas que é calculado pela frequência de ocorrência entre o aspecto e o atributo linguístico.

Na Figura 25 é ilustrado o esquema geral da rede heterogênea proposta nesta tese. Esse tipo de representação é uma estratégia intuitiva para combinar informações obtidas de diferentes domínios. Essa representação unificada, também elimina várias inconsistências das abordagens do aprendizado por transferência existentes, uma vez que o espaço de atributos é compartilhado por domínios fonte e alvo. O processo de aprendizado via propagação de rótulos não sofre o problema de diferentes distribuições de probabilidade, de acordo com a equação de regularização que será apresentada na próxima subseção.

5.3.2 Propagação de Rótulos de Aspectos por meio de Atributos Linguísticos

Essa proposta apresenta uma solução baseada na regularização de grafos para propagar informações de rótulos do domínio fonte para o domínio alvo, em que os atributos linguísticos são usados como uma ponte para essa propagação (ZHOU; CHEN; WANG, 2010). A abordagem de regularização de grafos tem que satisfazer duas premissas: (i) dois vértices conectados no grafo tendem a ser classificados com o mesmo rótulo e (ii) os rótulos de vértice devem estar próximos da informação de classe real do domínio fonte.

Figura 25 – Esquema geral da representação da rede heterogênea proposta para unificar a representação dos conjuntos de atributos entre domínios fonte (vértices com aspectos rotulados) e alvo (vértices dos candidatos a aspecto), em que os atributos linguísticos são usados como vértices ponte.

Fonte: Elaborada pelo autor.

A função de regularização a ser minimizada, que representa a propagação do rótulo de aspecto desta proposta, é definida na Equação 5.1. Os dois primeiros termos desta função de regularização, $\frac{1}{2} \sum_{a_s \in A^S} \sum_{l \in L} w_{a_s, l} (\mathbf{f}_{a_s} - \mathbf{f}_l)^2 + \frac{1}{2} \sum_{a_t \in A^T} \sum_{l \in L} w_{a_t, l}$, representam a primeira premissa. a_s é um vértice de A^S representando aspectos rotulados do domínio fonte, a_t é um vértice de A^T que representa os aspectos candidatos do domínio alvo e l é um vértice de L que representa um atributo linguístico. Os pesos das conexões entre os atributos linguísticos e os aspectos são representados por $w_{a_s, l}$ (aspectos rotulados) e $w_{a_t, l}$ (aspectos candidatos). Cada vértice do grafo tem um vetor de rótulos $\mathbf{f}$, que representa o nível de pertinência do vértice para uma das classes (Aspecto = Sim ou Aspecto = Não). A matriz $\mathbf{F}$ resume o nível de pertinência de todos os vértices do grafo para cada classe.

O último termo da função de regularização proposta representa a segunda premissa. Neste caso, o objetivo é garantir que a informação $\mathbf{f}$ dos vértices dos aspectos rotulados (proveniente do domínio de fonte) não seja modificada durante o processo de propagação do rótulo. Isso significa que a informação dos $\mathbf{f}$ vértices dos aspectos rotulados deve estar sempre próxima das informações reais do rótulo (do domínio fonte) representadas pelo vetor $\mathbf{y}$.

$$Q(\mathbf{F}) = \sum_{A^S, A^T, L \subset V} \frac{1}{2} \sum_{a_s \in A^S} \sum_{l \in L} w_{a_s, l} (\mathbf{f}_{a_s} - \mathbf{f}_l)^2 + \frac{1}{2} \sum_{a_t \in A^T} \sum_{l \in L} w_{a_t, l} (\mathbf{f}_{a_t} - \mathbf{f}_l)^2 + \lim_{\mu \rightarrow \infty} \mu \sum_{a_s \in A^S} (\mathbf{f}_{a_s} - \mathbf{y}_{a_s}) \quad (5.1)$$

A solução da função de regularização na Equação 5.1 pode ser obtida através de um algoritmo iterativo, uma vez que o uso de ferramentas de otimização matemática (*solvers*) para obter uma solução é computacionalmente caro e intratável para grandes números de vértices.

Deste modo, foi proposto um algoritmo iterativo para a propagação de rótulos de aspecto entre domínios por meio de rede heterogênea (CD-ALPHN), conforme descrito no algoritmo 9.

Algoritmo 9 – Cross-Domain Aspect Label Propagation through Heterogeneous Network (CD-ALPHN)

Entrada: $A^S, A^T, L, \mathbf{Y}, \mathbf{W}, \mathbf{D}$

Saída: $\mathbf{F}(\mathcal{O}^U)$

```

1 início
2    $\mathbf{P} \leftarrow (\mathbf{D}^{-1}) \cdot \mathbf{W}$ 
3   repita
4     para cada  $A^S, A^T, L \subset V$  faça
5        $\mathbf{F}(A^S, L) \leftarrow \mathbf{P}(A^S, L) \cdot \mathbf{F}(A^S, L)$ 
6        $\mathbf{F}(L, A^T) \leftarrow \mathbf{P}(L, A^T) \cdot \mathbf{F}(L, A^T)$ 
7        $\mathbf{F}(A^S) \leftarrow \mathbf{Y}(A^S)$ 
8     fim
9   até convergência ou número fixo de iterações
10 fim
11 retorna  $\mathbf{F}(A^T)$ 

```

A entrada de dados do algoritmo CD-ALPHN é composta por: (i) conjuntos de vértices A^S, A^T e L da rede heterogênea; (ii) uma matriz $\mathbf{Y}$ com a verdadeira informação de rótulo dos aspectos do domínio fonte, (iii) uma matriz $\mathbf{W}$ de pesos das conexões entre os vértices e (iv) uma matriz diagonal $\mathbf{D}$ com o grau de cada vértice.

Primeiramente, uma matriz de probabilidade de conexões $\mathbf{P}$ entre cada par de vértices é inicializada (Linha 2) pelo cálculo da relação entre o peso da conexão e o grau de vértices. Em seguida, é realizado o processo iterativo de propagação do rótulo (Linhas 4 - 8) o qual possui três etapas:

- Propagação de informações de rótulos dos vértices dos aspectos rotulados A^S (domínio fonte) para os vértices das características linguísticas L (ponte) – Linha 5;
- Propagação de informações de rótulos dos vértices das características linguísticas L (ponte) para os vértices dos aspectos do candidato A^T (domínio alvo) – Linha 6; e
- Reajuste do rótulo de informações dos vértices de aspectos A^S de acordo com os rótulos de domínio fonte – Linha 7.

A convergência do algoritmo CD-ALPHN é identificada quando a matriz $\mathbf{F}$ tem poucas modificações entre as iterações consecutivas ou quando um número máximo de iterações é alcançado. Após a convergência, cada vértice em A^T representando os candidatos a aspectos (domínio alvo) terá um nível de pertinência para as classes (Aspecto = Sim ou Aspecto = Não).

Esta informação será utilizada por sistemas para Mineração de Opiniões baseada em Aspectos e apoiar a tomada de decisões.

5.4 Avaliação Experimental

O objetivo da avaliação experimental é comparar se o aprendizado transdutivo com redes heterogêneas (CD-ALPHN) é mais eficaz que as abordagens indutivas para transferência de conhecimento entre domínios. O desempenho do algoritmo proposto, CD-ALPHN, foi avaliado experimentalmente utilizando conjuntos de dados de referência (revisões de produtos e serviços). Tais experimentos assim como os resultados obtidos são apresentados nesta seção. Além disso, um exemplo do Sistema de análise de dados chamado “Websensors-SentimentAnalysis” também é apresentado para demonstrar a relevância prática desta proposta em aplicações do mundo real.

5.4.1 Coleções de Revisões

Foram utilizadas sete coleções de revisões de referência que são amplamente citadas na literatura de mineração de opiniões baseada em aspectos. Na Tabela 9 é apresentada uma visão geral dessas coleções de revisões. Ressalta-se que cada coleção possui um conjunto de aspectos rotulados que permite avaliar o desempenho de métodos automáticos de extração de aspectos. As coleções de revisões *D1* a *D5* foram obtidas em [Hu e Liu \(2004\)](#). As coleções *D6* e *D7* foram obtidas em [Pontiki et al. \(2014\)](#).

Tabela 9 – Descrição das coleções de revisões utilizadas nos experimentos

ID	Coleção	#Revisões	#Aspectos	#TotalTermos	#MédiaTermos
<i>D1</i>	Câmera Digital 1	45	68	1045	54.46
<i>D2</i>	Câmera Digital 2	34	47	734	48.74
<i>D3</i>	Telefone Celular	41	78	924	51.70
<i>D4</i>	MP3 Player	95	126	1614	63.95
<i>D5</i>	DVD Player	99	82	1074	33.02
<i>D6</i>	Laptop	3045	955	2559	5.26
<i>D7</i>	Restaurante	3045	1219	3083	5.34

5.4.2 Configuração Experimental

Os experimentos foram configurados para comparar a abordagem proposta, CD-ALPHN, com as abordagens *cross-domain* do estado da arte para a extração de aspectos. As abordagens existentes de extração de aspectos entre domínios identificam atributos que são comuns aos domínios fonte e de alvo e treinam um classificador indutivo dos aspectos rotulados no domínio fonte. O modelo aprendido do classificador é aplicado para classificar aspectos no domínio alvo. Foram utilizados cinco classificadores populares ([ROSSI; MARCACINI; REZENDE, 2013](#)):

- **J48**: este classificador é baseado em árvore de decisão. Foi analisado o parâmetro de confiança c com os valores $c = \{0.15, 0.20, 0.25\}$.
- ***k*-Nearest Neighbors (*k*-NN)**: este classificador é baseado em instância. Foram analisados os valores para o parâmetro $k = \{1, 3, 5, 7, 9, 11, 13, 15\}$, em que k é o número de vizinhos mais próximos com similaridade de coseno.
- ***Multilayer Perceptron* (MLP)**: este classificador é baseado em rede neural usando um *Multilayer Perceptron* e o algoritmo de treinamento *Backpropagation*. Foram analisados os valores $\{8, 16, 32\}$ para o número de neurônios na camada oculta e a função sigmoid como função de ativação.
- ***Naïve-Bayes* (NB)**: este classificador é livre de parâmetros.
- ***Support Vector Machine* (SVM)**: foi considerado *kernel* Polinomial (expoente = 2) e como o parâmetro C é real, positivo e sem limite, utilizou-se esses valores como 10^Y e $Y = \{-3, -2, -1, 0, 1, 2, 3\}$.
- ***Cross-Domain Aspect Label Propagation through Heterogeneous Networks* (CD-ALPHN)** é proposta sem parâmetros.

A simulação de um processo de domínio cruzado para extração de aspectos foi realizada do seguinte modo: dada uma *coleção de revisões* para representar o domínio alvo, todas as outras *coleções* são combinadas para compor o domínio fonte. Por exemplo, se a *coleção D7* representa o domínio alvo, o domínio fonte é representado pelos aspectos rotulados (e seus atributos linguísticos) extraídos das *coleções D1, D2, D3, D4, D5 e D6*. Assim, a avaliação experimental proposta é semelhante a um cenário do mundo real em que os padrões linguísticos de aspectos rotulados de diferentes domínios fonte podem ser usados para extrair aspectos de um novo domínio alvo.

5.4.3 Critérios de Avaliação

Para avaliar o desempenho da extração de aspectos foi calculada a medida F^1 (Equação 5.3), que é a média harmônica das medidas de Precisão (Equação 5.2a) e Cobertura (Equação 5.2b), em que:

- ***VP* (Verdadeiro Positivo)** indica o número de termos inferidos como aspectos que foram corretamente extraídos;
- ***FP* (Falso Positivo)** indica o número de termos inferidos como aspectos que não são aspectos reais; e,
- ***FN* (Falso Negativo)** indica o número de termos que são aspectos e que não foram inferidos como aspectos.

$$Precisão = \frac{VP}{VP + FP}, \quad (5.2a) \quad Cobertura = \frac{VP}{VP + FN}, \quad (5.2b)$$

$$F^1 = 2 * \frac{Precisão * Cobertura}{Precisão + Cobertura}. \quad (5.3)$$

Tabela 10 – Level of inconsistency (β) considering the number of aspects of the target domain that occur in the source domain.

Source Domains	Target Domain	Level of Inconsistency
Digital Camera 2 (D2) Cellular Phone (D3) MP3 Player (D4) DVD Player (D5) Laptop (D6) Restaurant (D7)	Digital Camera 1 (D1)	Very Low-Level
Digital Camera 1 (D1) Cellular Phone (D3) MP3 Player (D4) DVD Player (D5) Laptop (D6) Restaurant (D7)	Digital Camera 2 (D2)	Very Low-Level
Digital Camera 1 (D1) Digital Camera 2 (D2) MP3 Player (D4) DVD Player (D5) Laptop (D6) Restaurant (D7)	Cellular Phone (D3)	Mid-Level
Digital Camera 1 (D1) Digital Camera 2 (D2) Cellular Phone (D3) DVD Player (D5) Laptop (D6) Restaurant (D7)	MP3 Player (D4)	Low-Level
Digital Camera 1 (D1) Digital Camera 2 (D2) Cellular Phone (D3) MP3 Player (D4) Laptop (D6) Restaurant (D7)	DVD Player (D5)	Low-Level
Digital Camera 1 (D1) Digital Camera 2 (D2) Cellular Phone (D3) MP3 Player (D4) DVD Player (D5) Restaurant (D7)	Laptop (D6)	High-Level
Digital Camera 1 (D1) Digital Camera 2 (D2) Cellular Phone (D3) MP3 Player (D4) DVD Player (D5) Laptop (D6)	Restaurant (D7)	High-Level

Tabela 11 – Resultados de F^1 para a extração de aspectos de cada abordagem.

ID	CD-ALPHN	J48	kNN	MLP	NB	SVM
D1	0.560	0.554	0.550	0.570	0.559	0.566
D2	0.579	0.579	0.572	0.581	0.572	0.574
D3	0.600	0.585	0.589	0.598	0.591	0.598
D4	0.589	0.583	0.581	0.587	0.585	0.595
D5	0.570	0.573	0.578	0.581	0.565	0.583
D6	0.610	0.480	0.547	0.581	0.570	0.504
D7	0.683	0.543	0.583	0.668	0.649	0.569

5.4.4 Análise e Discussão dos Resultados

Foram analisados os resultados experimentais considerando o nível de inconsistência entre o espaço de atributos dos domínios fonte e alvo. Quando processo de aprendizado por transferência entre domínios é realizado com baixo nível de inconsistência, o espaço de atributos do domínio alvo está bem representado no domínio fonte (há uma quantidade significativa de aspectos compartilhados entre ambos os domínios). Por exemplo, tem-se um baixo nível de inconsistência quando o domínio fonte contém aspectos sobre um modelo particular de uma Câmera Digital e o domínio alvo contém aspectos sobre outro modelo de Câmera Digital. Embora sejam modelos diferentes de Câmera Digital, os aspectos sobre esse tipo de produto são muito semelhantes. Por outro lado, quando o domínio alvo consiste em um produto ou serviço completamente novo com diferentes aspectos, tem-se um alto nível de inconsistência. Aliás, o último caso é mais interessante para aplicações do mundo real.

Na Tabela 11 são apresentados os resultados experimentais (Medida F^1) no processo de extração de aspectos. É destacado o melhor algoritmo para cada coleção de revisões com um texto em negrito nas células. Na sequência é apresentada uma análise dos resultados experimentais de acordo com o nível de inconsistência do processo entre domínios.

Muito baixo nível: As coleções de revisões D1 (Câmera digital 1) e D2 (Câmera digital 2) contêm quase os mesmos aspectos. Assim, um processo de domínio cruzado baseado em duas etapas foi suficiente para o processo de transferência de aprendizado, uma vez que não há praticamente nenhum problema de distribuição de espaço e probabilidades diferentes. O classificador de redes neurais (MLP) obteve os melhores resultados.

Baixo nível: Os *datasets* D4 (MP3 Player) e D5 (DVD Player) contêm muitos aspectos compartilhados, devido às funções semelhantes desses dois produtos. Do mesmo modo descrito acima (muito baixo nível), um processo de domínio cruzado baseado em duas etapas foi suficiente para o processo de transferência de aprendizagem. O correspondente classificador *Support Vector Machine* (SVM) obteve os melhores resultados.

Médio nível: As coleções de revisões D3 (Telefone Celular) compartilha alguns aspectos com

as coleções de revisões *D1* (Câmera digital 1), *D2* (Câmera digital 2) e *D4* (MP3 Player). No entanto, o domínio alvo contém uma quantidade significativa de novos aspectos. Nesse cenário, notou-se que o processo de aprendizado transdutivo proposto no algoritmo CD-ALPHN obtém resultados competitivos e obteve o melhor desempenho na extração de aspectos. No entanto, os resultados são semelhantes às abordagens tradicionais em duas fases, especialmente quando comparadas com os classificadores MLP e SVM.

Alto nível: As coleções de revisões *D6* (Laptop) e *D7* (Restaurante) representam cenários em que existe um grande conjunto de novos aspectos no domínio alvo. Neste caso, abordagens tradicionais baseadas em duas etapas produzem resultados inferiores ao algoritmo proposto CD-ALPHN para o aprendizado transdutivo. Esses resultados experimentais são uma forte evidência da importância de um modelo de representação unificado baseado em redes heterogêneas para aprendizado por transferência entre domínios.

Em resumo, o uso da aprendizado transdutivo a partir de uma representação unificada em redes heterogêneas produz resultados promissores para a aprendizado por transferências entre domínios, especialmente quando há um alto nível de inconsistências entre os domínios fonte e alvo — cenário mais comum em aplicações do mundo real. No entanto, enfatiza-se que o algoritmo proposto CD-ALPHN é competitivo mesmo quando há um baixo nível de inconsistência, sendo uma alternativa interessante para ser utilizado em diferentes cenários.

5.5 Aplicações

Para complementar a análise experimental baseada em coleções de referência, discute-se o uso prático da abordagem CD-ALPHN integrada aos sistemas de suporte à decisão. Assim, foi desenvolvido um Sistema de Inteligência Analítica para mineração de opiniões baseada em aspectos denominado por “*Websensors-SentimentAnalysis*”, que usa o conceito de “informação como sensor”. Neste caso, além das revisões, informações geográficas (*e.g.*, localização do consumidor) e informações de tempo (*e.g.*, data de publicação) também são coletadas para melhorar a visualização dos resultados (como dados do sensor).

5.5.1 Websensors

Com a popularização da Web e a interação dos usuários ativa e constantemente nesse ambiente, a Web tornou-se um ambiente virtual do reflexo dos acontecimentos no mundo real (ZHUGE, 2012). Devido a esse fato, pesquisas que investigam como extrair conhecimento desse conteúdo digital disponível na Web tem sido amplamente desenvolvidas. Um exemplo disso é o tema desta tese a Mineração de Opiniões. Uma das propostas que tem se destacado em diversos domínios de aplicações são os Websensors proposto por Marcacini *et al.* (2013). Os Websensors proporcionam a extração padrões e sensoriamento de conhecimento publicado na Web ao longo

do tempo. De forma geral, a construção dos *Websensors* é baseada na definição do problema e nas necessidades da aplicação de métodos de aprendizado de máquina para gerar modelos de aprendizado que são atualizados periodicamente para monitorar a evolução ao longo do tempo de acordo com aplicação. Alguns exemplos dessas aplicações foram desenvolvidos na área de agronegócios (SANTOS; ROSSI; MARCACINI, 2017) e segurança pública (FLORENCE; NOGUEIRA; MARCACINI, 2017). Nesta tese foi proposta o uso de Websensors para a Mineração de Opiniões baseada em Aspectos.

5.5.2 Websensors-SentimentAnalysis

O “Websensors-SentimentAnalysis” usa o conceito de “informação como sensor”. Neste caso, além das revisões, informações geográficas (*e.g.*, localização do consumidor) e informações de tempo (*e.g.*, data de publicação) também são coletadas para melhorar a visualização dos resultados (como dados do sensor).

Em um Sistema de Inteligência Analítica, o objetivo é que o usuário/analista tenha uma visão geral dos dados de interesse, compilados e representados de forma intuitiva. Na Figura 26 é apresentada uma das principais interfaces do *Websensors-SentimentAnalysis* a partir de 93 revisões reais sobre restaurantes, em que o modelo de extração de aspectos pode ser obtido usando vários outros domínios fonte com aspectos rotulados. Inicialmente, todos os aspectos de acordo com as polaridades positivas e negativas são resumidos na Figura 26-A. Isso fornece uma visão geral do produto ou serviço. Quando o usuário seleciona uma polaridade de interesse (Positivo ou Negativo), o sistema Websensors-SentimentAnalysis apresenta: as revisões com os aspectos destacados (Figura 26-B), uma evolução temporal da polaridade de acordo com a frequência de aspectos positivos ou negativos (Figura 26-C) e um mapeamento geográfico de acordo com a localidade dos consumidores que apresentaram as revisões (Figura 26-D).

No contexto deste trabalho, é importante enfatizar que na Figura 26-B, os aspectos são destacados de acordo com a matriz $F(A^T)$ obtida pelo algoritmo proposto CD-ALPHN. O usuário pode selecionar os aspectos extraídos e, em seguida, obter uma visão geral sobre quais comentários que este aspecto obteve polaridades positivas ou negativas. Assim, selecionando um aspecto de interesse e sua polaridade, pode-se combinar o aspecto selecionado com informações geográficas para construir um mapa de calor, como mostrado na Figura 27.

Uma vez que os aspectos são extraídos, várias outras funcionalidades de um sistema de inteligência analítica podem ser implementadas, como notificação de ocorrência de um novo aspecto, relatórios dos aspectos *top-k* mais comentados nas revisões, alertas com os aspectos que precisam de atenção (aumento de polaridade negativa) do serviço ou produto. Essas funcionalidades do sistema de inteligência analítica proposto podem ser utilizadas como diretrizes para o desenvolvimento de novos sistemas de suporte à decisão e mineração de opiniões.

Figura 26 – Visão geral da interface do sistema Websensors-SentimentAnalysis (Data Analytics) para mineração de opiniões baseada em aspectos.

Fonte: Elaborada pelo autor.

5.6 Considerações Finais

Neste capítulo, foi apresentado uma nova abordagem para a extração de aspecto de um determinado domínio usando aspectos rotulados de outros domínios diferentes. Esta abordagem inova na (i) proposta de representação unificada de espaços de atributos entre diferentes domínios por meio de redes heterogêneas e em (ii) proposta de processo de transferência de transferência entre domínios através da propagação de rótulos com aprendizado transdutivo.

O algoritmo iterativo proposto CD-ALPHN (*Cross-Domain Aspect Label Propagation through Heterogeneous Networks*) foi comparado com as abordagens do estado da arte em tarefas de domínio cruzado e obteve resultados competitivos. Além disso, o CD-ALPHN é potencialmente útil em cenários em que há uma inconsistência de alto nível entre atributos fonte e alvo. Aliás, as aplicações do mundo real geralmente apresentam inconsistência de alto nível, uma vez que as avaliações são coletadas de diferentes fontes de dados, com diferentes tamanhos de texto e diferentes estilos de escrita.

Além da avaliação experimental com as coleções de referência, foi apresentado como a abordagem proposta pode ser empregada em sistemas de suporte à decisão. Especificamente, é apresentado um Sistema de Análise de Dados para mineração de opiniões baseadas em aspectos. Algumas características básicas deste *Data Analytics System* foram descritas e uma análise que contém diretrizes gerais interessantes para propor novas soluções.

As orientações para o trabalho futuro envolvem o enriquecimento da representação com

Figura 27 – Exemplo de um mapa de calor que representa a ocorrência geográfica de um aspecto selecionado de acordo com a localização dos consumidores.

Fonte: Elaborada pelo autor.

base em redes heterogêneas. Incluir a polaridade do sentimento (negativo ou positivo) no processo de transferência entre domínios, permitindo assim extrair aspectos e suas polaridades com suporte de dados rotulados de diferentes domínios. O desenvolvimento da proposta CD-ALPHN, as respectivas avaliações experimentais e práticas resultaram no seguinte artigo científico aceito em periódico relevante para essa área de pesquisa ([MARCACINI et al., 2018](#)).

PROJETOS PILOTOS

6.1 Considerações Iniciais

Nos Capítulos 3, 4 e 5 foram propostas e avaliadas abordagens para a Mineração de Opiniões baseada em Aspectos com o objetivo de diminuir a necessidade de esforço humano para rotular dados e para definir recursos adicionais dependentes de domínio. Neste capítulo são apresentados os projetos pilotos para a Mineração de Opiniões baseada em Aspectos desenvolvidos em colaboração com outros pesquisadores do grupo de pesquisa. Mais especificamente, são descritos três projetos e suas respectivas avaliações experimentais: (i) Sistema para Anotação e Sugestão de Rótulos para Mineração de Opiniões baseada em Aspectos; (ii) *Framework* para Mineração de Opiniões baseada em Aspectos; e (iii) Comparativo de métodos para extração de aspectos baseados em Relações Sintáticas e Aprendizado de Máquina para Mineração de Opiniões. É importante destacar que esses projetos pilotos e as respectivas avaliações experimentais estão em suas primeiras versões.

6.2 Sistema para Anotação de Documentos e Sugestão de Rótulos para Mineração de Opiniões baseada em Aspectos

No cenário de Mineração de Opiniões é comum que técnicas de análise automática para extração de aspectos e classificação de sentimentos utilizem métodos de aprendizado supervisionado, isto é, que utilizam textos previamente rotulados como um conjunto de treinamento para “aprender” a fazer automaticamente a análise dos textos. Devido a esse fato, há a necessidade de anotar manualmente os conjuntos de documentos para que sirvam de base para o treinamento e aprendizado do modelo para a análise automática. No entanto, por ser uma área de pesquisa recente, existem poucas coleções de textos anotados para a Mineração de Opiniões baseada

em Aspectos (SMATANA *et al.*, 2013; CHEN; MUKHERJEE; LIU, 2014). Em geral, cada grupo de pesquisa constrói seu corpus manualmente e com padrões particulares. É importante destacar que não existem padrões de anotação e de formatação digital definidos. A anotação linguística de corpus é uma área importante da pesquisa em processamento de língua natural, pois o corpus anotado é a base para diversas áreas de pesquisa (ROBERTS *et al.*, 2009; XIA; YETISGEN-YILDIZ, 2012).

O uso de ferramentas de apoio podem contribuir para padronizar e agilizar o processo de anotação manual de textos. Uma ferramenta construída com esse propósito é o RotuLabic, uma ferramenta criada pelos pesquisadores do Laboratório de Inteligência Computacional (LABIC) do ICMC-USP. A primeira versão do RotuLabic oferece suporte apenas à anotação em nível de documento como um todo. Essa versão do sistema permite a anotação de textos manualmente e utiliza métodos de aprendizado de máquina para oferecer ao usuário sugestões de rótulo ao documento como um todo.

O objetivo deste projeto piloto foi desenvolver um sistema de anotação de documentos para Mineração de Opiniões baseada em Aspectos. Esse sistema, designado por RotuLabic-AspectSentiment (RotuLabic-AS), tem como principal diferencial permitir anotar diversos termos do documento que indicam aspectos, suas respectivas polaridades dos sentimentos e os termos que indicam sentimento. Além disso, o sistema pode utilizar algoritmos de extração de aspectos e classificação de sentimentos para sugerir rótulos e assim agilizar e facilitar a tarefa do usuário (anotador).

6.2.1 RotuLabic-AspectSentiment

Para permitir um processo de anotação dos aspectos e seus respectivos sentimentos o RotuLabic-AspectSentiment (RotuLabic-AS) possui três componentes principais: (i) administração; (ii) anotação e (iii) sugestão de rótulos. Cada componente pode possuir diversos outros componentes. Cada componente será explicado nas subseções seguintes.

6.2.1.1 RotuLabic-AS – Administração

O componente de administração possui as funcionalidades para criação de um processo de anotação, configuração do processo e sumarização do andamento do processo. Para criar esse processo, o administrador precisa fornecer os arquivos que contém as revisões que serão rotuladas e um arquivo com instruções a serem seguidas pelos seus anotadores. Além disso, deve-se definir as configurações do processo e adicionar os usuários que serão anotadores. Nas Figuras 28 e 29 são apresentadas, respectivamente, as interfaces para criação do processo e para associação de anotadores ao processo.

Figura 28 – Interface para criação do processo de anotação - configuração.

Create new Labeling Process

Process Name

Upload labeling instructions Nenhum arquivo selecionado

Upload documents to be labeled Nenhum arquivo selecionado

Hidden Aspect ☒ Yes ☐ No

General Aspect ☒ Yes ☐ No

Add Labels

NEGATIVE

POSITIVE

NEUTRAL

Aspect Suggestion Algorithm

Polarity Suggestion Algorithm

Language of the documents

Use Automatic Translator ☒ No ☐ Yes

Min Frequency

Fonte: Elaborada pelo autor.

Figura 29 – Interface para criação do processo de anotação - anotadores.

Select the taggers labeling process "Restaurant-EN"

Choose a tagger

Taggers	
Solange	<input type="button" value="X"/>
Ivone	<input type="button" value="X"/>
Darlan	<input type="button" value="X"/>
Leonardo	<input type="button" value="X"/>

Fonte: Elaborada pelo autor.

O administrador do processo de anotação também pode acompanhar o andamento da anotação. Esses resultados são sumarizados nas interfaces conforme as Figuras 30a, 30b e 30c.

Figura 30 – Interface para sumarização do processo de anotação.

Labeling Process Information

Labeling Process: restaurant-br-o3	
Min agreement rate	2
Status	In Progress
Number of documents	5
Number of taggers	3
Aspect Identification Algorithm	Frequency-based
Polarity Classification Algorithm	None
Using Automated Translator	No
Language of the documents	Portuguese

Add Taggers

(a) Sumarização da situação do processo.

Labeling Process Statistics

Total Progress(9/15);
60% complete

Progress by Tagger hide/show	
Jyone	5/5 (100%)
Leonardo	4/5 (80%)
Solange	0/5 (0%)

Aspect Agreement, not considering polarity (0/9);

Agreement on Aspect and Polarity (0/9);

All-agreed Documents(0/3);

Agreement by Document
show/hide

(b) Concordância entre anotadores.

Download of Annotations, SemEval-like XML files

Include sentiment words? ☒ Yes ☐ No

Agreement Rate

(c) Exportação das anotações.

Fonte: Elaborada pelo autor.

6.2.1.2 RotuLabic-AS – Sugestão de Rótulos

Após criar o processo de anotação com os seus respectivos anotadores, são executados os métodos para sugestão de rótulos. No RotuLabic-AS são utilizados métodos de extração de aspectos e de classificação de sentimentos dos aspectos. Nesta versão piloto, foram implementados os métodos mais simples e que não necessitassem de muitos parâmetros de configuração e de recursos adicionais.

Para extração de aspectos foi utilizado o método baseado em frequência de substantivos. Esse método extrai como aspecto os termos que são substantivos mais frequentes. O parâmetro deste método é o limite de frequência mínima para que um termo seja extraído como aspecto. Esse parâmetro é definido pelo usuário.

Para o método de classificação de sentimentos dos aspectos foi utilizado o método baseado em léxico. No RotuLabic-AS é utilizado um recurso adicional composto por arquivos que contém palavras que são consideradas positivas e outro arquivo com palavras que indicam sentimentos negativos. Nesta primeira versão, este léxico é pré-definido no sistema. Foi utilizado o léxico disponibilizado em <<https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>>.

Figura 31 – Interface de anotação com sugestões de rótulos.

Document : restaurant_review_001

The price is reasonable although the service is poor.

Add Aspect

Select part of the text, then choose the polarity

NEGATIVE
NEUTRAL
POSITIVE

Add Hidden Aspect Add General Aspect

Remove All Aspects

Aspect	Polarity	Sentiment Words
price	POSITIVE	reasonable x
service	NEGATIVE	poor x

Previous Document Skip Document Finish

Return to the instructions

Fonte: Elaborada pelo autor.

Após execução dos respectivos métodos, o sistema apresenta as revisões a serem anotadas com as respectivas sugestões. Para sugerir termos que indicam sentimentos, eles são escritos de outra cor. Os termos sugeridos como aspectos são marcados na respectiva cor da polaridade dos sentimentos. Os termos sugeridos são apresentados na tabela inferior. No sistema, a interface de sugestão, bem como a de anotação, é a mesma conforme a Figura 31. Os métodos da sugestão de rótulos podem utilizar outros recursos adicionais como léxicos, dicionários, exemplos rotulados, entre outros.

6.2.1.3 Rotulabic-AS – Anotação

A interface de anotação do Rotulabic-AS é bem intuitiva. Cada revisão é apresentada com as respectivas sugestões. Para anotar um aspecto, o usuário seleciona um ou mais termos e seleciona a respectiva classe (positivo, negativo ou neutro). Os termos selecionados serão adicionados na tabela e marcados com a respectiva cor da classe. Para anotar termos que indicam sentimentos, designados aqui por “termos de sentimentos”, o usuário seleciona os termos na revisão e adiciona-os na linha da tabela para o aspecto associado. A Tabela está localizada na parte inferior da interface, conforme a Figura 31. Uma restrição do sistema é que todos os termos de sentimentos estejam relacionados a pelo menos um aspecto.

Quando o processo de cada anotador for concluído, a sumarização da situação do processo, bem como, o percentual de concordância entre os anotadores podem ser obtidos na interface apresentada na Figura 30a. As anotações de cada revisão podem ser exportadas de acordo com a concordância mínima definida pelo usuário. Para cada revisão é criado um arquivo com as respectivas anotações conforme exemplo apresentado na Figura 32.

Figura 32 – Formato de arquivo de exportação das anotações da revisão: “*The price is reasonable although the service is poor*”.

```
<noteTerms>
  <aspectTerm term="price" polarity="positive" from="4" to="9"/>
  <sentimentTerm term="reasonable" polarity="positive" from="13" to="22"/>

  <aspectTerm term="service" polarity="negative" from="37" to="44"/>
  <sentimentTerm term="poor" polarity="negative" from="48" to="51"/>
</noteTerms>
```

Fonte: Elaborada pelo autor.

6.2.2 Anotação de Coleções de Revisões

Para testar e validar o sistema, foi realizada a anotação de três coleções de revisões de avaliação de produtos no idioma português, sendo uma de laptop¹ com 89 documentos, um

¹ Página de origem: <<http://www.buscape.com.br/dell-inspiron-14-serie-3000-intel-core-i3-4005u-1-7-ghz-4096-mb-1024-avaliacoes>> (avaliações obtidas em 25 de março de 2016)

restaurante² com 101 documentos e uma de um aparelho celular³ com 100 documentos. As coleções foram rotuladas por dois pesquisadores em Mineração de Opiniões.

6.2.3 Considerações Finais – Rotulabic-Aspect Sentiment

O sistema proposto para Anotação e Sugestão de Rótulos contribuiu para auxiliar a construção de coleções de revisões e aspectos com seus respectivos sentimentos e aspectos anotados. Outra contribuição é que os dados das anotações são armazenados e podem ser utilizados na construção de dicionários ou léxicos para Mineração de Opiniões. Ainda é necessária a realização de testes do sistema e uma avaliação experimental para verificar se os métodos de sugestão de rótulos estão contribuindo no processo. As coleções de textos e código-fonte do Rotulabic-AS estão disponíveis em <<http://sites.labc.icmc.usp.br/ipm/tese/cap6/rotulabic-as>>

6.3 Framework para Mineração de Opiniões baseada em Aspectos

Com a popularização da internet e sua consequente quebra de barreiras geográficas, a capacidade de troca de informação entre grupos de diferentes países ganha relevância. É frequente a utilização de um idioma comum, geralmente o inglês, para facilitar a comunicação e a utilização das informações geradas. No entanto, existem muitos conteúdos e opiniões já produzidos e disponíveis na Internet em diversos idiomas. Por isso, há esforços para o desenvolvimento de ferramentas que por meio de tradução automática tornem informações acessíveis em idiomas diferentes do que foram originalmente (BORGMAN, 1997; CHOWDHURY, 2003; JURAFSKY; MARTIN, 2000). A maioria das propostas de Mineração de Opiniões em diversos idiomas é realizada em nível de documento como um todo (ARAUJO *et al.*, 2016; GUO *et al.*, 2010). Este trabalho teve como objetivo a construção de um *framework* para realizar a Mineração de Opiniões baseada em Aspectos e suportar diversos idiomas.

Diversos métodos são propostos para a realização de Mineração de Opiniões baseada em Aspectos usando diferentes técnicas. Esses métodos apresentam suas vantagens e desvantagens, variando de acordo com o domínio, idioma ou tipo de texto (LIU, 2015). Em vista disso, neste trabalho é proposto um *framework*, denominado Opinion-meter, para realização de Mineração de Opiniões baseada em Aspectos em diversos idiomas por meio da combinação de métodos de extração de aspectos e classificação de sentimentos e o uso de tradução automática.

Para o desenvolvimento do *framework* Opinion-meter foram levantados os principais trabalhos referentes ao processo de Mineração de Opiniões baseada em Aspectos e tradução

² Fonte:<<https://kekanto.com.br/biz/outback-steakhouse-23>> (avaliações obtidas em 25 de março de 2016)

³ Fonte:<<http://www.buscapi.com.br/smartphone-motorola-moto-g-2-geracao-dtv-colors-xt1069-16-gb#avaliacoes>> (avaliações obtidas em 25 de março de 2016)

automática. Dentre os trabalhos que abordam a questão da Mineração de Opiniões em múltiplos idiomas, destacam-se [Araujo et al. \(2016\)](#) e [Guo et al. \(2010\)](#). Esse último enfatiza as vantagens na possibilidade de analisar como pessoas de culturas diferentes podem avaliar uma mesma entidade. Em [Araujo et al. \(2016\)](#), ao comparar diversos métodos para o idioma inglês com a utilização de tradutor automático e de métodos nativos para outros idiomas, os resultados demonstram que, ao utilizar tradução automática, pode-se obter resultados semelhantes ou até melhores do que ao utilizar métodos nativos.

Dentre as principais tarefas para realização de Mineração de Opiniões baseada em Aspectos, a literatura apresenta uma variedade de métodos para a extração de aspectos e para a classificação de sentimentos. Os métodos que são frequentemente utilizados como base de comparação são os de [Turney \(2002\)](#) e [Hu e Liu \(2004\)](#) que foram utilizados como base para os algoritmos implementados neste trabalho.

Em [Turney \(2002\)](#) é apresentado um algoritmo para classificação de sentimento em sentenças. No método proposto, obtém-se os rótulos identificando a classe gramatical das palavras do texto avaliado. Em seguida, as sequências desses rótulos são comparadas com um conjunto de padrões pré-definidos como, por exemplo, um adjetivo seguido de um substantivo, para extrair frases cujas polaridades serão avaliadas. Para o cálculo da polaridade de cada frase, utiliza-se uma medida baseada no *Point-wise Mutual Information* (PMI). A equação do PMI é dada a seguir:

$$PMI(p1, p2) = \log_2 \left(\frac{P(p1, p2)}{P(p1)P(p2)} \right) \quad (6.1)$$

Em que $P(p1, p2)$ é a probabilidade das palavras $p1$ e $p2$ ocorrerem no mesmo documento, enquanto $P(pX)$ é a probabilidade da dada palavra aparecer no documento. Uma estimativa do PMI é feita substituindo as probabilidades por número de resultados em um motor de busca. Por fim, a polaridade de cada frase (f) é estimada obtendo-se a diferença entre o PMI de f e uma palavra com conotação positiva e o PMI de f e uma palavra com conotação negativa, resultando na medida *Semantic Orientation* (SO).

Em [Hu e Liu \(2004\)](#) é apresentada uma abordagem baseada em frequência para a extração de aspectos e palavras de sentimento. Nessa abordagem extraem-se expressões frequentes, para encontrar os potenciais aspectos, focando em substantivos e sintagmas nominais. Para selecionar as expressões frequentes é utilizada a geração de *itemsets* pelo algoritmo Apriori. Após identificar os aspectos nas revisões, considera-se que ao expressarem as opiniões, as pessoas utilizam palavras de sentimento próximas aos aspectos. Deste modo, nesse método, para identificar as palavras de sentimentos utiliza-se qualquer adjetivo próximo a um *itemset* frequente. Por fim, uma abordagem baseada em Léxico é utilizada para estimar a polaridade de cada aspecto, considerando as palavras de sentimento identificadas próximas aos aspectos.

6.3.1 Opinion-meter

O *framework* Opinion-meter, proposto neste trabalho, tem o objetivo de integrar em um único ambiente diversos métodos de extração de aspectos e a classificação da polaridade dos sentimentos para Mineração de Opiniões baseada em Aspectos. O *framework* usa tradução automática para permitir que métodos e ferramentas originalmente criadas para a utilização com textos em idioma inglês, possam ser utilizados com textos em outros idiomas. Como alternativa, o *framework* também incorpora técnicas de Aprendizado de Máquina para a classificação de sentimento, de forma independente de idioma.

Na Figura 33 são representados os principais componentes do *framework* e o fluxo da informação desde o *upload* dos textos até a obtenção e visualização dos resultados.

Inicialmente, o usuário deve fazer o *upload* dos textos que deseja analisar, agrupados sob um mesmo processo de análise. Ao fazê-lo, o usuário também indica se deseja utilizar tradução automática ou o idioma original dos textos. Os textos são pré-processados por uma ferramenta de classes gramaticais das palavras (para extrair a marcação *Parch-of-Speech* (POS)) (BALDRIDGE, 2005). Na sequência, é realizada a análise dos textos com os métodos escolhidos pelo usuário. Após a execução dos algoritmos, um resumo dos resultados é apresentado em uma interface do *framework*. Os métodos utilizados para extração de aspectos e classificação de sentimentos são apresentados nas próximas subseções.

Figura 33 – Arquitetura geral do *framework* Opinon-meter

Fonte: Elaborada pelo autor.

6.3.1.1 Extração de Aspectos

Os métodos de extração de aspectos tem objetivo de identificar nos textos palavras que indicam características ou propriedades das referidas entidades. O método de extração de aspectos implementado neste trabalho é baseado em Hu e Liu (2004) com as propostas na Seção 4.2 desta tese. Nesse método, são selecionadas como candidatas a aspecto apenas as palavras classificadas como substantivos, em que as marcações POS são apresentadas na Tabela 1. Para cada palavra o algoritmo considera sua frequência no conjunto dos textos, e apenas palavras cuja frequência atinge um valor mínimo (definido pelo usuário) são consideradas como

aspecto. Essa abordagem parte de um princípio de que conjuntos de substantivos frequentes tendem a ser os aspectos mais relevantes avaliados [Hu e Liu \(2004\)](#).

6.3.1.2 Classificação de Sentimentos dos Aspectos

Após a extração de aspectos, é realizada a classificação da polaridade atribuída ao sentimento associado a cada um dos aspectos extraídos em cada texto. Cada polaridade é classificada como “positiva”, “negativa” ou “neutra”, caso haja (i) associações boas, (ii) associações ruins, ou (iii) nenhuma associação identificada ou associações igualmente boas e ruins, respectivamente. Os métodos implementados são detalhados a seguir.

Tabela 12 – Tags utilizadas para selecionar candidatos a aspecto

Inglês		Português	
Tag	Descrição	Tag	Descrição
NN	Substantivo singular	N	Substantivos
NNS	Substantivo plural		

6.3.1.3 Baseado em Léxico

Essa abordagem é uma adaptação de [Hu e Liu \(2004\)](#), em que para cada sentença e aspectos nela contidos, a polaridade é calculada somando-se as contribuições das palavras da sentença que se encontram dentro de uma distância de até 10 palavras do aspecto sob análise, como na equação 6.2. Se o valor obtido for menor que zero, atribui-se polaridade negativa ao aspecto, caso seja maior que zero, atribui-se polaridade positiva, e caso seja igual a zero, atribui-se a polaridade neutra.

$$p(i) = \sum_{j=\max(i-10,1), i \neq j}^{\min(i+10,n)} (l_{pos}(s(j)) \times (v) - l_{neg}(s(j)) \times (v)), \quad (6.2)$$

em que, i é a posição do aspecto na sentença s , $s(j)$ a j -ésima palavra da mesma, $v = n - |i - j|$ é uma medida de proximidade, e $l_{pos}(x)$ e $l_{neg}(x)$ são dados pela equação 6.3.

$$l_{pos \text{ (ou neg)}}(x) : \begin{cases} 1 & \text{se } x \text{ é positiva (ou negativa)} \\ 0 & \text{caso contrário} \end{cases} \quad (6.3)$$

as funções $l_{pos}(x)$ e $l_{neg}(x)$ utilizam o Léxicos de Sentimento ([SOUZA; VIEIRA, 2012](#); [WILSON; WIEBE; HOFFMANN, 2005](#); [HU; LIU, 2004](#)) correspondente ao idioma utilizado para obter a informação de polaridade das palavras.

6.3.1.4 Baseado em PMI

Esse método é baseado na proposta de [Turney \(2002\)](#). Neste trabalho, para identificar as frases consideradas no cálculo da polaridade para textos em inglês, foram utilizados os padrões

baseados em marcações POS como definidos em [Turney \(2002\)](#). Para textos em português, foram utilizados os mesmos padrões, mas considerando as marcações POS correspondentes utilizadas pelo POS *tagger* no idioma em português. Para a extração das frases para os textos traduzidos, além dos padrões originalmente propostos, foram incluídos padrões que preveem a inversão da ordem das duas primeiras palavras dos padrões para os quais essa inversão não estava prevista. Por fim, para cada frase extraída é calculado a Orientação Semântica (*Semantic Orientation* (SO)). A polaridade de cada sentença é calculada com base na polaridade dominante e, para tanto, os valores de SO são somados.

6.3.1.5 Baseado em Classificação por Aprendizado de Máquina

Dentre os diversos métodos existentes de classificação, neste trabalho optou-se por utilizar um dos métodos mais simples, o algoritmo de Naïve Bayes ([RISH, 2001](#)). Esse método baseia-se no teorema de Bayes ([KOCH, 1990](#)) para identificar a classe para a qual um exemplo x tem a maior probabilidade de estar associado, como dado pela Equação 6.4.

$$y = \max_{y_i} P(y_i|x) \quad (6.4)$$

Neste trabalho, foi utilizada a implementação do classificador Naïve Bayes disponível na ferramenta WEKA ([HOLMES; DONKIN; WITTEN, 1994](#)).

6.3.2 Avaliação Experimental

Neste experimento buscou-se avaliar a viabilidade do *framework* proposto para a extração de aspectos e classificação dos sentimentos associados. Nas próximas subseções são apresentados a configuração experimental, os dados das bases utilizadas, os critérios de avaliação e os resultados obtidos.

6.3.2.1 Configuração Experimental

Para avaliar os métodos implementados, foram obtidas bases com aspectos e sentimentos rotulados para os idiomas inglês, espanhol e francês do SemEval 2016⁴. Além disso, por não haver uma base rotulada disponível no idioma português, textos de avaliação de produtos em português foram rotulados por dois especialistas de domínio. Foram considerados apenas os aspectos coincidentes e identificados com polaridades concordantes entre ambos. Todas as bases rotuladas utilizadas são do mesmo domínio e estão disponíveis em <http://sites.labc.icmc.usp.br/tese/cap6/opinion-meter>. Na Tabela 13 são apresentados o número de sentenças, aspectos, aspectos positivos, negativos e neutros de cada base.

Em relação ao método de extração de aspectos baseado em frequência, utilizou-se, para todas as bases, uma frequência mínima igual a 2, exceto para a base em português, que contém mais de 6 vezes o número de aspectos que a média das demais, para essa utilizou-se o valor 8.

⁴ SemEval: <http://alt.qcri.org/semeval2016/> (acesso em 27 de maio de 2016)

Tabela 13 – Dados das bases utilizadas

	Inglês	Português	Francês	Espanhol
Sentenças	48	397	40	58
Aspectos	55	281	35	38
Positivos	35	244	20	29
Negativos	17	35	13	6
Neutros	3	2	2	3

6.3.2.2 Análise e Discussão dos Resultados

Para avaliar os acertos na extração de aspectos, foi considerado que um aspecto foi corretamente identificado caso o mesmo esteja contido em algum dos aspectos identificados nas bases rotuladas. Para avaliar as polaridades dos sentimentos, por sua vez, foram consideradas apenas aquelas associadas aos aspectos corretamente identificados, considerando como corretas quando as polaridades atribuídas pelo algoritmo e pela base rotulada coincidem. Foi calculada a média ponderada da precisão entre as classes (positiva, negativa e neutra), dada pela razão entre o número de exemplos corretamente classificados para uma classe (positivos verdadeiros) e o total de exemplos classificados como dessa classe (positivos falsos e positivos verdadeiros).

Na Tabela 14 é apresentada a precisão na extração de aspectos e na Tabela 15 é apresentada a precisão na classificação de sentimentos. A precisão do método baseado em AM foi calculada por meio de validação cruzada, sendo a média de 4 execuções (MITCHELL, 1997). Em ambas as tabelas, a sigla ST indica que não houve uso de tradução automática e CT indica o seu uso. Vale ressaltar que, apesar de utilizar a mesma base rotulada, os resultados para o idioma português com e sem tradução podem variar dependendo da qualidade da tradução e da classificação feita pelos POS *taggers* e dos léxicos para inglês (texto traduzido) e português (texto sem tradução).

Ao comparar os resultados obtidos para português com e sem o uso de tradução automática, nota-se que os resultados obtidos com o uso de tradução são similares ou superiores aos obtidos sem o seu uso. Isso aponta para a possibilidade de que, com os devidos ajustes e aprimoramentos das técnicas utilizadas, resultados satisfatórios podem ser obtidos com o uso de tradução automática. Nota-se também que os resultados obtidos para português, na extração de aspectos, são consideravelmente superiores aos dos demais idiomas. Uma possível explicação é que o uso de uma frequência mínima maior pode ter colaborado para a remoção de substantivos ou palavras equivocadamente classificadas como tal mas pouco relevantes no conjunto dos textos.

Tabela 14 – Precisão na extração de aspectos por base rotulada

EN - ST	PT - ST	PT - CT	FR - CT	ES - CT
0.74	0.3	0.42	0.3	0.21

Os métodos baseados em PMI e léxico apresentam resultados bastante próximos para quase todas as bases, embora a abordagem baseada em léxico tenda a resultados melhores.

Tabela 15 – Precisão por classe na classificação de sentimento para cada método e base rotulada

	PMI	Léxico	AM
Inglês - ST	0.44	0.64	0.49
Português - ST	0.22	0.23	0.64
Português - CT	0.25	0.33	0.67
Francês - CT	0	0.086	0.94
Espanhol - CT	0.19	0.15	0.66

Os resultados obtidos com utilização de AM, por outro lado, parecem os mais promissores, obtendo os melhores resultados na maioria dos cenários. Isso é especialmente interessante quando considerado que o algoritmo implementado é um dos mais simples encontrados na bibliografia.

Na segunda etapa da análise, os resultados para o idioma francês são os que mais variam, possivelmente devido ao reduzido número de aspectos corretamente identificados e ao fato de não haverem aspectos de diferentes polaridades na mesma sentença entre aquelas avaliadas na segunda etapa.

6.3.3 Considerações Finais – Opinion-meter

A principal contribuição deste trabalho é agregar e integrar, em um único ambiente, diversos métodos de Mineração de Opiniões baseada em Aspectos com a utilização de tradução automática, possibilitando a análise de opiniões da *web* em diversos idiomas. Os resultados obtidos demonstram que o uso de tradução automática junto aos métodos de Mineração de Opiniões baseada em Aspectos é promissor, embora ainda haja necessidade de aprimoramento das técnicas utilizadas.

Outra contribuição é que o *framework* Opinion-meter está disponível ⁵ de forma que qualquer pessoa possa utilizá-lo para analisar textos nos idiomas suportados, utilizando os métodos implementados ou adicionando e estendendo o *framework* para utilizar seus próprios métodos. O *framework* Opinion-meter está em sua primeira versão e ainda deve ser ampliado para incorporar outros métodos de Mineração de Opiniões baseada em Aspectos e de aprendizado de máquina da literatura, além de ferramentas adicionais de processamento de língua natural.

6.4 Estudo Comparativo de métodos para extração de aspectos baseados em Relações Sintáticas e Aprendizado de Máquina para Mineração de Opiniões

As abordagens de Extração de Aspectos para Mineração de Opiniões baseada em Relações Sintáticas apresentam resultados promissores, porém exigem a necessidade de um es-

⁵ Todo o *framework*, código-fonte e dados estão disponíveis em <<http://sites.labc.icmc.usp.br/tese/cap6/opinion-meter>>

pecialista para estudar o corpus e definir o conjunto de regras sintáticas a serem utilizados. Com intenção de diminuir a necessidade de especialistas, foi investigado o uso de métodos de aprendizado de máquina para gerar um modelo que defina esse conjunto de regras sintáticas.

Neste projeto piloto, o objetivo foi comparar o desempenho de dois tipos de abordagens para métodos de extração de aspectos: aprendizado supervisionado e relações sintáticas.

6.4.1 Extração de Aspectos baseada em Relações Sintáticas

A abordagem baseada em Relações Sintáticas é considerada um método não supervisionado e tem como objetivo analisar a relação entre aspectos e termos opinativos. Consiste em identificar relações sintáticas e gramaticais em uma sentença para definir os termos que representam aspectos. Qiu *et al.* (2009) utilizou as dependências sintáticas para comparar sequência de palavras que contém aspectos ou palavras já identificadas, enquanto a abordagem de Bancken, Alfarone e Davis (2014) com o algoritmo *Aspectator* considerou os caminhos de dependência para descobrir novos aspectos, o que permite obter maior cobertura. Em uma abordagem similar, Wang *et al.* (2011) usou aspectos e termos relacionados à aspectos pré-definidos como entrada para o algoritmo e, em seguida, empregou o algoritmo *bootstrap*, para descobrir novos aspectos baseado em recorrência de palavras utilizadas como sementes e novos termos candidatos. Para a avaliação comparativa, foi utilizado o método baseado em relações sintáticas proposto por Bancken, Alfarone e Davis (2014), cujo experimento obteve resultados superiores em precisão e cobertura.

Aspectator

Este método utiliza como entrada uma coleção de revisões de um produto específico e extrai automaticamente os aspectos mais positivos e negativos desse produto acompanhado de uma lista de sentenças que contribuíram para a polaridade (positivo/negativo) do sentimento de cada aspecto. O algoritmo possui três passos principais:

- (1) Para cada sentença, o algoritmo extrai uma árvore de dependência sintática utilizando o analisador sintático ou *parser* por dependência da *Stanford NLP*⁶ (MANNING *et al.*, 2014; MARNEFFE; MANNING, 2008). Em seguida, extrai um par de opinião <M;A> composto de um Aspecto (A) e um modificador de sentimento (M) utilizando caminhos de dependências sintáticas, conforme apresentado na Figura 34.
- (2) Consolida diferentes menções dos aspectos extraídos no primeiro passo e verifica a similaridade semântica entre eles utilizando a biblioteca *WordNet* (WS4J) com o objetivo de identificar diferentes palavras que são relacionadas ao mesmo aspecto.

⁶ <<http://nlp.stanford.edu:8080/corenlp>>

- (3) Utilizando a *SentiWordNet* como o recurso léxico, atribui um *score* de sentimento a cada palavra extraída e ao consolidar o *score* final de cada aspecto, é possível identificar quais tiveram a maior e menor nota.

O *Aspectator* não exige exemplos rotulados e pode ser executado em qualquer coleção de texto sobre um produto ou serviço específico. A abordagem aproveita os caminhos de dependência sintática para dois propósitos: (i) detectar menções de aspecto e modificadores de sentimento, e (ii) descobrir relações entre eles. Esta é uma diferença significativa comparando com outras abordagens baseadas em dependências sintáticas.

Figura 34 – Dependências sintáticas do *Aspectator*

Tipo	Caminho de Dependência	Exemplo
Adjectival modifier	$M \xleftarrow{amod} A$
Direct object	$A \xleftarrow{nsubj} * \xrightarrow{dobj} M$
Adjectival complement	$A \xleftarrow{nsubj} * \xrightarrow{acomp} M$
Complement of a copular verb	$A \xleftarrow{nsubj} M \xrightarrow{cop} *$
Adverbial modifier to a passive verb	$A \xleftarrow{nsubjpass} * \xrightarrow{advmod} M$

Fonte: Adaptada de Bancken, Alfarone e Davis (2014).

6.4.2 Extração de Aspectos Baseada em Aprendizado de Máquina Supervisionado

A abordagem baseada em aprendizado supervisionado utiliza um conjunto de exemplos rotulados (dados de treinamento) para construir um classificador que será utilizado para prever a classe dos novos elementos. O objetivo da aprendizagem supervisionada é construir um modelo para mapear a relação entre a entrada e saída de um determinado conjunto de dados baseando-se em dados de treinamento já existentes. Dessa forma, a partir de resultados anteriores, é possível gerar a saída de novos conjuntos de dados.

6.4.2.1 Representação de Textos para Classificação de Aspectos

As coleções de textos foram representadas por uma matriz atributo-valor como proposto em (MATSUNO *et al.*, 2016). Para exemplificar essa representação, a sentença $s = \text{"The food is good, especially their more basic dishes, and the drinks are delicious."}$. Inicialmente é construída a árvore de dependência sintática como apresentado na Figura 35. Na sequência, uma matriz

atributo-valor é construída em que cada linha representa uma palavra da coleção e as colunas são os atributos linguísticos. Cada célula contém a frequência com que cada palavra aparece na coleção com o respectivo atributo linguístico e a coluna Aspecto indica se a palavra é um aspecto ou não, de acordo com a classificação original disponibilizada com o conjunto de dados de teste. As *stopwords* (preposições, artigos, conjunções e pronomes) são removidas da matriz. A Tabela 16 representa a matriz atributo-valor construída para a sentença *s*.

Figura 35 – Árvore sintática da sentença *s*.

Fonte: Elaborada pelo autor.

Tabela 16 – Matriz atributo-valor da sentença da Figura 35

PALAVRA	POS_JJ	POS_NN	POS_NNS	POS_RB	DEP_IN_ADVMOD	DEP_IN_DEP	DEP_IN_NSUBJ	DEP_OUT_ADVMOD	DEP_OUT_CC	DEP_OUT_CONJ	DEP_OUT_COP	DEP_OUT_DEP	DEP_OUT_DET	DEP_OUT_NMOD	DEP_OUT_NSUBJ	Aspecto?
<i>food</i>		1					1									S
<i>good</i>	1							1	1	1		1				N
<i>especially</i>				1	1											N
<i>basic</i>	1				1											N
<i>dishes</i>			1			1		2						1		S
<i>drinks</i>			1				1						1			S
<i>delicious</i>	1									1	1				1	N

6.4.2.2 Classificadores

Nesta avaliação experimental inicial, para a extração dos aspectos, foram utilizados os seguintes classificadores: Naïve Bayes (NB), J48 e *k*-Nearest Neighbour (KNN). Foram utilizadas as implementações disponíveis na ferramenta WEKA⁷.

6.4.3 Avaliação Experimental

Para comparar o desempenho das duas abordagens de extração de aspectos foi realizada uma avaliação experimental utilizando as coleções de revisões citadas no Capítulo 4.

⁷ <<https://www.cs.waikato.ac.nz/ml/weka/>>

6.4.3.1 Configurações dos Classificadores

Para a classificação dos aspectos, foram necessários dois conjuntos de textos, separados em: (i) dados de teste compostos pela matriz atributo-valor das palavras do conjunto de textos a ser avaliado, e (ii) dados de treinamento gerados da compilação de todos os conjuntos de textos, excluindo o conjunto de teste a ser avaliado. Para avaliação da execução dos algoritmos supervisionados foram utilizadas as implementações disponibilizadas na ferramenta *WEKA* com a configuração de parâmetros padrão de cada algoritmo. Os parâmetros e considerações dos algoritmos de aprendizado supervisionado são:

- **Naïve Bayes (NB)**: configuração padrão, não há parâmetros.
- **J48**: neste algoritmo de indução de árvores de decisão, foi utilizado o valor 0,25 para parâmetro *confidence factor*.
- **k-NN**: Foi considerado $k = 3$ e foi utilizado cosseno como medida de similaridade.

6.4.3.2 Configuração do Algoritmo Aspectator

O método baseado em relações, representado pelo *Aspectator* também utiliza como base o *parser* de dependência *StanfordNLP* para análise de árvore sintática. No entanto, o método define os caminhos de dependência mais importantes para sua execução, conforme mostrado na Figura 34. Além disso, o programa utiliza outros métodos de pré-processamento de texto, como por exemplo funções *stemmers*, da API *Lucene Snowball*⁸. Esse pré-processamento facilita a consolidação dos aspectos, já que palavras similares são agrupadas. A lista de aspectos final é disponibilizada em formato texto após a execução do programa. Os aspectos mais mencionados (positivamente ou negativamente) são extraídos, bem como exemplos de sentenças que foram utilizadas para a avaliação. Na Figura 36 é apresentado o formato de saída do programa.

6.4.3.3 Medidas de Avaliação

Para avaliar o desempenho, foram utilizadas as medidas de Precisão, Cobertura e F^1 como descrito no Capítulo 4. Para obter os valores de cada coleção, utilizou-se a Lista de Aspectos de Referência que continha todos os aspectos da respectiva coleção e a Lista de Aspectos Inferidos por cada método. Para realizar o cálculo, utilizou-se a ferramenta que realiza a correspondência exata e aproximada entre os aspectos de referência e os inferidos disponível em <<http://sites.labc.icmc.usp.br/ipm/tese/cap4/evaluation>>.

6.4.3.4 Análise e Discussão dos Resultados

Os resultados de Precisão, Cobertura e F^1 obtidos em cada coleção de revisões são apresentados nos gráficos das Figuras 37, 38 e 39 respectivamente para correspondência exata

⁸ <http://lucene.apache.org/core/3_0_3/api/contrib-snowball/>

Figura 36 – Arquivo de saída do *Aspectator*

```

Most positive
-----
Entity: food
Positive mentions: 264 (strength: 0.4251481167602952)
Mention 0 strength: 1.0 original entity: excellent#a food
    While the food was excellent, it wasn't cheap (though not extremely
    expensive either).

Mention 1 strength: 1.0 original entity: excellent#a food
    The vibe is very relaxed and cozy, service was great and the food was
    excellent!

Negative mentions: 82 (strength: 0.32009166129047384)
Mention 0 strength: 0.875 original entity: unfortunately#r food
    Unfortunately, the food was NOT something to get worked up about.

Mention 1 strength: 0.8125 original entity: absolutely#r_horrible#a food
    The food was absolutely horrible!

Total score: 85.9915865988991

```

entre os aspectos de referência e inferidos. É possível notar o método de extração baseado em relações sintáticas (*Aspectator*) possui maior precisão em relação aos classificadores. O resultado satisfatório está relacionado ao modo de processamento, especialmente a atribuição de *scores* para cada aspecto, que permite selecionar os mais relevantes para cada conjunto de textos. Porém, se o método acertar apenas um aspecto ele também terá precisão máxima (1.0 ou 100%).

Figura 37 – Resultados para medida de precisão.

Fonte: Elaborada pelo autor.

Isso também impacta negativamente o indicador de *Recall* e F^1 uma vez que apenas os aspectos mais relevantes são selecionados pelo *Aspectator*. Os classificadores apresentaram um desempenho baixo (valores inferiores a 40%), porém apresentaram um comportamento estável. Considerando os resultados de F^1 nenhum dos métodos foi melhor em todas as coleções de *reviews*. O classificador J48 apresentou melhores resultados em um maior número de coleções, mas não em todas.

Figura 38 – Resultados para medida de cobertura.

Fonte: Elaborada pelo autor.

Figura 39 – Resultados para medida de F^1 .

Fonte: Elaborada pelo autor.

6.4.4 Considerações Finais – Métodos baseado em Relações e Aprendizado Máquina

Neste trabalho comparou-se duas abordagens para extração de aspectos para Mineração de Opiniões que possuem diferentes estratégias para inferir se um termo é aspecto. Os resultados demonstraram uma melhor precisão utilizando a abordagem baseada em relações sintáticas, representada pelo algoritmo *Aspectator*. Os principais aspectos de cada conjunto de textos foram extraídos corretamente, apesar de não considerar aspectos que foram menos citados. Essa abordagem ainda possui a vantagem de ser independente do domínio, não precisar de exemplos rotulados, podendo ser aplicada rapidamente em novos conjuntos de dados. Conforme mostrado nos resultados, a abordagem manteve uma precisão adequada mesmo em diferentes tipos de dados (produtos e serviços). Entretanto, essa abordagem depende de um analisador de dependência sintática específico para cada idioma.

Os resultados de precisão dos classificadores para extração de aspectos foram inferiores, mas apresentaram melhores resultados nos valores de Cobertura e F^1 . Os resultados de ambos, considerando F^1 , foram inferiores a 40% o que demonstra que ainda existem muitos desafios para melhorar o desempenho dos métodos de extração de aspectos para Mineração de Opiniões.

CONCLUSÕES

Neste capítulo são apresentadas as conclusões desta tese de doutorado. Para tal, as questões de pesquisa definidas na introdução são retomadas e relacionadas com as contribuições científicas alcançadas durante o desenvolvimento deste trabalho. Na sequência, é apresentada uma síntese dos resultados desta tese. Por fim, são apresentadas as considerações sobre as limitações ocorridas no desenvolvimento desta tese, bem como as direções para trabalhos futuros.

7.1 Contribuições Científicas

Nesta tese, as soluções propostas tiveram por objetivo diminuir a necessidade de grande esforço humano na Mineração de Opiniões baseada em Aspectos. As questões de pesquisa foram divididas em dois grupos: (I) reduzir a necessidade de exemplos rotulados para construir um modelo para as tarefas de extração de aspectos e classificação de sentimentos dos aspectos e (II) diminuir a necessidade de recursos adicionais dependentes de domínio para extração de aspectos.

As questões de pesquisa do grupo I estão relacionadas a como reduzir a quantidade de exemplos rotulados para construir um modelo para as tarefas de extração de aspectos e classificação de sentimentos de aspectos. A seguir são discutidas essas questões.

Em relação às questões de pesquisa **QP-I.1 – Como as representações de textos das revisões de produtos e serviços podem impactar no processo de extração de aspectos por meio de aprendizado semissupervisionado baseado em redes?** e **QP-I.2 – Como as representações de textos das revisões de produtos e serviços podem impactar no processo de classificação de sentimentos dos aspectos por meio de aprendizado semissupervisionado baseado em redes?**, nesta tese foi realizada uma revisão da literatura que demonstrou que a maioria das abordagens é baseada em aprendizado supervisionado, paradigma que exige uma grande quantidade de exemplos rotulados o que pode limitar o desempenho em situações práticas. Além disso, as representações de textos de revisões para Mineração de Opiniões baseada em Aspectos que

utilizam o aprendizado de máquina são pouco exploradas na literatura. Em geral, utilizam-se representações dos textos no modelo espaço-vetorial devido à facilidade para extração dos termos e baixo custo de implementação. Entretanto, esse tipo de representação não permite explorar outras relações entre os termos e documentos, além da classe do documento.

Como solução para diminuir a necessidade de grande quantidade exemplos rotulados para a extração de aspectos e classificação de sentimentos foi proposta, nesta tese, a *Aspect-Based Sentiment Propagation for Heterogeneous Networks* (ASPHN) que representa a coleção de revisões por meio de redes bipartidas heterogêneas e que permite representar relações e tipos de objetos, como aspectos rotulados, aspectos não rotulados, atributos linguísticos e outras informações disponíveis sobre o problema. Assim, propicia o uso de algoritmos de propagação de rótulos para o aprendizado semissupervisionado. Essa solução foi abordada no Capítulo 3.

Para validar a proposta e responder as questões de pesquisa **QP-I.3 – Qual a influência da quantidade de exemplos rotulados no processo de extração de aspectos e classificação dos sentimentos?**, **QP-I.4O uso de métodos de aprendizado semissupervisionados baseado em redes pode aperfeiçoar o processo de extração de aspectos?** e **QP-I.5 – O uso de métodos de aprendizado semissupervisionados baseado em redes pode aperfeiçoar o processo de classificação de sentimentos dos aspectos?** foi realizada uma avaliação experimental que considerou (i) algoritmos de aprendizado supervisionados e semissupervisionados; (ii) algoritmos semissupervisionados baseados em dois tipos de representações de textos: modelo espaço-vetorial e redes bipartidas heterogêneas; e, (iii) para avaliar a influência da quantidade de exemplos rotulados os experimentos foram realizados variando a quantidade de exemplos rotulados de 1 a 70.

Tanto para a extração de aspectos quanto para a classificação de sentimentos dos aspectos, os algoritmos de aprendizado semissupervisionado apresentaram melhor desempenho que os algoritmos de aprendizado supervisionado. Além disso, os algoritmos semissupervisionados baseados em redes apresentaram melhor desempenho do que os algoritmos semissupervisionados que usaram representação no modelo espaço-vetorial. Em relação a quantidade de exemplos rotulados para o aprendizado foi observado que os algoritmos semissupervisionados apresentaram desempenho superior utilizando 10 ou 20 exemplos rotulados, enquanto que os algoritmos de aprendizado supervisionados apresentaram resultados satisfatórios a partir de 50 exemplos rotulados.

A solução proposta ASPHN demonstra como as representações de textos de revisões podem impactar nas tarefas de extração de aspectos e classificação de sentimentos que a torna promissora para suprir a necessidade de grande número de exemplos rotulados. Os detalhes da avaliação experimental e os resultados obtidos foram apresentados na Seção 3.4 desta tese. O desenvolvimento e avaliação da proposta da ASPHN resultaram na publicação de um artigo em conferência nacional e um artigo em periódico:

timent Analysis using Semi-Supervised Learning in Bipartite Heterogeneous Networks. Journal of Information and Data Management - JIDM, 2016.

MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M. ; REZENDE, S. O. . Análise de Sentimentos baseada em Aspectos usando Aprendizado Semissupervisionado em Redes Heterogêneas. In: Symposium on Knowledge Discovery, Mining and Learning, 2015, Petrópolis. Proceeding of the 3rd Symposium on Knowledge Discovery, Mining and Learning, 2015. p. 90-93.

As questões de pesquisa do grupo II estão relacionadas a como minimizar a necessidade de esforço humano para definir recursos adicionais dependentes de domínio. A seguir são discutidas essas questões.

Para responder a questão de pesquisa **QP-II.1 – Como a ausência de informações de domínio pode impactar os métodos de extração de aspectos para Mineração de Opiniões?** foram investigados e avaliados os principais métodos não supervisionados e independentes de domínio. Nesta avaliação, foi constatado que os métodos baseados em frequência que combinam unigramas e bigramas compostos por substantivos apresentaram melhor desempenho do que os baseados em modelos de tópicos e em medidas de associação. Porém, os resultados dos métodos baseados em frequência são inferiores quando comparados aos resultados obtidos por métodos que utilizam recursos adicionais e de aprendizado semissupervisionados e supervisionados. Portanto, a ausência de informações adicionais dependentes de domínio diminuem o desempenho dos métodos de extração de aspectos.

Os métodos analisados e as estratégias utilizadas são apresentadas no Capítulo 4 desta tese. O desenvolvimento desta avaliação resultou em um artigo científico que foi submetido para um periódico relevante para essa área de pesquisa:

MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Towards domain-independent aspect extraction methods for sentiment analysis. Expert Systems with Applications, 2018. [submetido]

Em relação à questão de pesquisa **QP-II.2 – Como as informações de outros domínios podem ser utilizadas para extrair aspectos de um domínio não conhecido?** foi verificado que estratégias de Aprendizado de Máquina por Transferência entre Domínios tem sido uma solução promissora para problemas em que o processo de rotulação é muito custoso ou não há exemplos rotulados. Nesta estratégia utiliza-se as características dos exemplos rotulados de um domínio (domínio origem) para aprender um modelo para outro domínio (domínio alvo). Normalmente, são utilizados métodos de aprendizado indutivos. Tais métodos possuem uma limitação relacionada à diferenças do espaço de atributos entre os domínios fonte e alvo. Como resposta a esse desafio, nesta tese, a solução proposta foi o uso da estratégia de aprendizado por

transferência entre domínios utilizando aprendizado semissupervisionado transdutivo baseado em redes heterogêneas denominada *Cross-Domain Aspect Label Propagation through Heterogeneous Networks* (CD-ALPHN). O método CD-ALPHN explora redes heterogêneas para unificar o espaço de atributos do domínio fonte e alvo, obtendo resultados competitivos para extração de aspectos quando comparado com métodos existentes. Em especial, o método CD-ALPHN obtém resultados superiores quando o domínio alvo contém novos aspectos (não cobertos no domínio fonte). Essa solução foi descrita no Capítulo 5 desta tese.

Quanto à questão de pesquisa **QP-II.3 – Como representações de textos de revisões podem impactar no aprendizado por transferência para extração de aspectos?**, na CD-ALPHN foi proposta uma representação de textos das revisões por meio de redes heterogêneas que usa atributos linguísticos como objetos comuns entre os domínios origem e alvo. Também foi proposto um algoritmo de propagação de rótulos baseado nessa representação para classificar os rótulos do domínio alvo. A representação de textos propostas para o aprendizado entre domínios é descrito na Subseção 5.3.1 desta tese. O algoritmo para propagar rótulos baseada nessa representação é apresentado na Subseção 5.3.2.

Para responder a questão **QP-II.4 – Qual o impacto das informações de outros domínios no processo de extração de aspectos de um domínio não conhecido?**, foi realizada uma avaliação experimental para comparar a solução proposta CD-ALPHN com as abordagens tradicionais que identificam atributos comuns aos domínios (origem e alvo) e treinam um classificador indutivo supervisionado dos aspectos rotulados no domínio de origem. O modelo aprendido é aplicado para classificar aspectos no domínio alvo. Os algoritmos CD-ALPHN, SVM e *Multilayer Perceptron* obtiveram os melhores desempenho em cenários em que o domínio alvo e domínio fonte compartilham muitos aspectos. Além da análise quantitativa, foi feita uma análise de inconsistência entre os domínios fonte e alvo. O CD-ALPHN apresentou resultados superiores quando há alto nível de inconsistência entre os domínios, e esse cenário é o mais comum em aplicações do mundo real. No entanto, enfatiza-se que o algoritmo proposto é competitivo mesmo quando há baixo nível de inconsistência. Essa avaliação experimental e seus resultados são apresentados na Seção 5.4.

Os recursos adicionais dependentes de domínio podem melhorar a extração de aspectos para Mineração de Opiniões, porém exigem especialistas para definir esses recursos que são custosos para o processo; e as cenários de frequente atualização dos dados pode ser impraticável. A proposta nesta tese para extração de aspectos por meio de aprendizado por transferência entre domínios é pioneira e apresentou resultados promissores. Essa estratégia de aprendizado apresenta muitas possibilidades de melhorias e assim pode contribuir efetivamente para minimizar o esforço humano por meio do aproveitamento de informações de aspectos já rotuladas em outros domínios.

Para complementar a avaliação experimental foi discutido o uso prático da solução CD-ALPHN em Sistemas de Suporte à Decisão. Para isso foi desenvolvido um Sistema de Análise

de Dados para Mineração de Opiniões baseada em Aspectos denominado por “Websensors-SentimentAnalysis”, que usa o conceito de “informação como sensor”. Neste caso, além das revisões, também foram coletadas informações geográficas (por exemplo, localização do consumidor) e informações de tempo (por exemplo, data de publicação) para melhorar a visualização dos resultados (como dados do sensor).

O desenvolvimento da proposta CD-ALPHN, as respectivas avaliações experimentais e práticas resultaram no seguinte artigo científico submetido para um periódico relevante para essa área de pesquisa:

MARCACINI, R. M.; MATSUNO, I. P.; ROSSI, R. G.; REZENDE, S. O. Cross-domain aspect extraction for sentiment analysis: a transductive learning approach. *Decision Support Systems*, 2018. [em revisão para publicação]

Por fim, outros trabalhos relacionados com esta tese foram desenvolvidos em parcerias com outros pesquisadores e apresentaram resultados promissores e que podem ser expandidos para novas pesquisas na área. Esses trabalhos foram denominados nesta tese por Projetos Pilotos e foram apresentados e discutidos no Capítulo 6.

O primeiro projeto piloto, apresentado na Seção 6.2, foi o Sistema de Anotação de Documentos para Mineração de Opiniões baseada em Aspectos que permite rotular termos que são aspectos, suas respectivas polaridades e termos que indicam os sentimentos. Para realização de testes e validação da proposta foi realizada a anotação de duas coleções de revisões sobre restaurante e laptop no idioma “Português-BR”. Este sistema é uma contribuição importante pois existem poucas coleções de revisões com aspectos e seus respectivos sentimentos anotados e por disponibilizar uma ferramenta para auxiliar a comunidade científica no processo de geração de novas coleções. O desenvolvimento deste projeto piloto resultou em dois artigos científicos. O primeiro já foi publicado em um evento de Iniciação Científica e o outro está em desenvolvimento:

FARIAS, D. S.; MATSUNO, I. P.; REZENDE, S. O. Annotation of Corpus for Aspect-Based Sentiment Analysis. In: *Simpósio Internacional de Iniciação Científica e Tecnológica da USP – SIICUSP*, 2016.

MATSUNO, I. P.; FARIAS, D. S.; MARCACINI, R. M.; REZENDE, S. O. Rotulabic-Aspect Sentiment: Sistema de Anotação com Sugestão de Rótulos para Mineração de Opiniões baseada em Aspectos. [em desenvolvimento]

O segundo projeto piloto foi a proposta e desenvolvimento de um *framework* para Mineração de Opiniões baseada em Aspectos. Neste *framework* foram implementados alguns métodos para extração de aspectos e para classificação dos sentimentos dos aspectos em um único ambiente. Também foi utilizado tradução automática como alternativa para suportar vários

idiomas. Este *framework* foi apresentado na Seção 6.3. O *framework* deve evoluir para facilitar a incorporação de novos métodos e ferramentas. O desenvolvimento deste projeto contribui para realização de novos trabalhos na área de Mineração de Opiniões baseada em Aspectos. O desenvolvimento deste projeto resultou na publicação seguinte artigo científico:

FARIAS, D. S.; MATSUNO, I. P.; MARCACINI, R. M.; REZENDE, S. O. Opinion-meter: um Framework para Análise de Sentimentos Baseada em Aspectos. In: Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2016, Teresina-PI.

No terceiro projeto piloto foi realizada uma avaliação comparativa inicial entre os métodos para extração de aspectos baseados em aprendizado de máquina e baseados em relações sintáticas. As especificações dos métodos e os resultados obtidos foram apresentados na Seção 6.4. O objetivo foi explorar o uso de aprendizado considerando atributos linguísticos. Este projeto resultou em um artigo científico em desenvolvimento:

ANZAI, R.; MATSUNO, I. P.; Dos SANTOS, A. C.; MARCACINI, R. M.; REZENDE, S. O. Estudo Comparativo de Métodos de Extração de Aspectos para Mineração de Opiniões. [em desenvolvimento].

Nas seções seguintes são apresentadas uma síntese dos resultados, as limitações encontradas e direções para trabalho futuro.

7.2 Síntese dos Resultados e Relação de Trabalhos Publicados

Baseado nos desenvolvimentos das propostas para atingir os objetivos definidos nesta tese de doutorado, foram obtidas as seguintes contribuições científicas:

- Proposta de uma nova Representação de Textos para coleções de revisões para produtos e serviços para extração de aspectos;
- Proposta de uma nova Representação de Textos para coleções de revisões para produtos e serviços para classificação de sentimentos de aspectos;
- Proposta de um novo uso (ou aplicação) de aprendizado semissupervisionado para Mineração de Opiniões baseada em Aspectos;
- Proposta de novas estratégias para métodos extração de aspectos independentes de domínio;
- Proposta de uma nova Representação de Textos para coleções de revisões para produtos e serviços para aprendizado por transferência entre domínios;

- Proposta de um novo método para propagação de rótulos para aprendizado por transferência entre domínios;
- Padronização, pré-processamento e disponibilização de um conjunto de coleções de revisões de produtos e serviços ;
- Proposta e desenvolvimento de novo sistema de rotulação de aspectos e respectivos sentimentos com algoritmos de sugestão de rótulos (*Rotulabic-Aspect Sentiment*);
- Divulgação dos resultados científicos diretos desta tese por meio de artigos:
- Artigos científicos publicados e submetidos em outras áreas relacionadas:

MARCACINI, R. M.; SINOARA, R. A.; MATSUNO, I. P.; REZENDE, S. O. Aprendizado não Supervisionado de Websensors. In: Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2013, São Carlos-SP. Proceeding of the 1st Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2013.

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 (IGDP), v. 13, p. 57-66, 2015.

MARQUES, C. A. N.; MATSUNO, I. P.; SINOARA, R. A.; REZENDE, S. O.; ROZENFELD, H. An exploratory study to evaluate the practical application of PSS methods and tools based on text mining. In: International Conference on Engineering Design, 2015, Milan. Proceedings of the 20th International Conference on Engineering Design (ICED15), 2015. v. 7. p. 7-311-7-320.

BARBOSA, J. R.; MATSUNO, I. P.; GUIMARAES, E. R.; REZENDE, S. O.; VINCENZI, A. M. R.; DELAMARO, M. E. Mineração de Textos para Apoiar a Predição de Severidade de Relatórios de Incidentes: um Estudo de Viabilidade. In: Simpósio Brasileiro de Qualidade de Software, 2017, Rio de Janeiro. Simpósio Brasileiro de Qualidade de Software, 2017. p. 89-103.

BARBOSA, J. R.; MATSUNO, I. P.; MARCACINI, R. M.; REZENDE, S. O.; VINCENZI, A. M. R.; DELAMARO, M. E. Predicting the Severity of Bug Reports with Semi-Supervised Learning: A Case Study, The Journal of Systems and Software, 2018 [submetido]

7.3 Limitações e Trabalhos Futuros

Durante o desenvolvimento das soluções propostas, foram definidos critérios de avaliação considerando os padrões usados nos respectivos tipos de abordagem separadamente. Ainda não

foi realizado uma avaliação que uniu todos os resultados encontrados encontrados, devido à dificuldade de comparar diretamente abordagens não supervisionadas e semissupervisionadas. Uma direção para trabalho futuro é uma avaliação experimental comparativa das abordagens propostas (aprendizado semissupervisionado, independente de domínio, aprendizado por transferências) usando mesmos métodos de avaliação apresentados e coleções de revisões.

As soluções propostas para as representações das coleções de revisões para extração de aspectos para o aprendizado semissupervisionado e para o aprendizado por transferência entre domínios explorou como atributos linguísticos os tipos de marcações PoS (Part of Speech) e relações sintáticas entre os termos. Um direcionamento para trabalhos futuros é exploração de outros atributos linguísticos que considerem também a semântica de modo automático, por exemplo as representações que utilizam Expressões de Domínio propostas em [Scheicher et al. \(2016\)](#), [Marques et al. \(2015\)](#) e Papéis Semânticos proposta em [Sinoara, Rossi e Rezende \(2016\)](#). Nesta tese foram propostas representações das coleções de revisões para classificação de sentimentos por meio de redes bipartidas heterogêneas que também podem ser estendida e aperfeiçoadas. Representações de coleções de revisão para classificação de sentimentos dos aspectos para aprendizado por transferência entre domínios não foram exploradas nesta tese e são direções para trabalhos futuros.

Em relação aos métodos independentes de domínio, é importante avaliar qual o nível de resultados obtidos com independência e qual o ganho com o uso de recursos com informações adicionais dependentes de domínio. Assim é possível avaliar a relação custo-benefício para a criação de novos recursos dependentes de domínio. Um direcionamento para trabalhos futuros é a investigação de métodos híbridos. Do mesmo modo, avaliar qual o ganho do uso desses recursos adicionais nas soluções proposta nesta tese por meio de aprendizado semissupervisionado e por transferência entre domínios.

Por fim, é importante ressaltar que a Mineração de Opiniões baseada em Aspectos é um tema de pesquisa relativamente novo e além de apresentar diversas questões em aberto possui poucas ferramentas implementadas disponíveis. Nesta tese, foram realizadas investigações, desenvolvimento de propostas e avaliações de métodos para extração de aspectos e classificação de sentimentos. Além disso, teve-se a preocupação de contribuir com as questões práticas dessa área por meio de desenvolvimento ferramentas para auxiliar em projetos futuros que embora estejam e suas versões iniciais são também direcionamento para trabalhos futuros.

REFERÊNCIAS

AGGARWAL, C. C.; WANG, H. **Managing and Mining Graph Data**. 1. ed. [S.l.]: Springer Publishing Company, 2010. ISBN 1441960449, 9781441960443. Citado nas páginas 48 e 55.

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 47.

AGGARWAL, C. C.; ZHAO, P. Towards graphical models for text processing. **Knowledge and information systems**, Springer, v. 36, n. 1, p. 1–21, 2013. Citado na página 48.

AKHTAR, M. S.; GUPTA, D.; EKBAL, A.; BHATTACHARYYA, P. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. **Knowledge-Based Systems**, Elsevier, v. 125, p. 116–135, 2017. Citado na página 109.

AL-MOSLMI, T.; OMAR, N.; ABDULLAH, S.; ALBARED, M. Approaches to cross-domain sentiment analysis: A systematic literature review. **IEEE Access**, IEEE, v. 5, p. 16173–16192, 2017. Citado na página 112.

ALPAYDIN, E. **Introduction to Machine Learning**. [S.l.]: MIT Press, 2010. Citado na página 49.

ANGELOVA, R.; KASNECI, G.; WEIKUM, G. Graffiti: Graph-based classification in heterogeneous networks. **World Wide Web**, Kluwer Academic Publishers, Hingham, MA, USA, v. 15, n. 2, p. 139–170, mar. 2012. Citado na página 55.

ANNEX, S. The new wave of user generated content: 7 concepts for 2017 and beyond. 2016. Citado na página 25.

ARAÚJO, M.; REIS, J.; PEREIRA, A.; BENEVENUTO, F. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In: **SAC '16 - Proceedings of the 31st Annual ACM Symposium on Applied Computing**. [S.l.: s.n.], 2016. p. 1140–1145. Citado nas páginas 131 e 132.

ARCHACK, A. G. N.; IPEIROTIS, P. G. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In: **KDD '07 - Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2007. p. 56–65. Citado na página 36.

AUE, A.; GAMON, M. Customizing sentiment classifiers to new domains: A case study. In: **CITESEER. Proceedings of recent advances in natural language processing (RANLP)**. [S.l.], 2005. v. 1, n. 3.1, p. 2–1. Citado na página 47.

BALDRIDGE, J. The opennlp project. URL: <http://opennlp.apache.org/index.html>, 2005. Citado na página 133.

- BANCKEN, W.; ALFARONE, D.; DAVIS, J. Automatically detecting and rating product aspects from textual customer reviews. In: **DMNLP '14 - Proceedings of the 1st International Conference on Interactions Between Data Mining and Natural Language Processing - Volume 1202**. [S.l.]: CEUR-WS.org, 2014. p. 1–16. Citado nas páginas 138 e 139.
- BELKIN, M.; MATVEEVA, I.; NIYOGI, P. Regularization and semi-supervised learning on large graphs. In: SPRINGER. **International Conference on Computational Learning Theory**. [S.l.], 2004. p. 624–638. Citado na página 48.
- BLAIR-GOLDENSOHN, S.; HANNAN, K.; MCDONALD, R.; NEYLON, T.; REIS, G. A.; REYNAR, J. Building a sentiment summarizer for local service reviews. In: **NLPIX '08 - WWW Workshop on NLP in the Information Explosion Era**. [S.l.: s.n.], 2008. Citado nas páginas 27, 42 e 46.
- BLANCO, R.; LIOMA, C. Graph-based term weighting for information retrieval. **Information Retrieval**, Kluwer Academic Publishers, Hingham, MA, USA, v. 15, n. 1, p. 54–92, 2012. Citado na página 55.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, JMLR.org, v. 3, p. 993–1022, 2003. Citado nas páginas 44, 68, 94 e 95.
- BLITZER, J.; DREDZE, M.; PEREIRA, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: **Proceedings of the 45th annual meeting of the association of computational linguistics**. [S.l.: s.n.], 2007. p. 440–447. Citado na página 47.
- BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: **ACM. Proceedings of the eleventh annual conference on Computational learning theory**. [S.l.], 1998. p. 92–100. Citado na página 53.
- BOIY, E.; MOENS, M.-F. A machine learning approach to sentiment analysis in multilingual web texts. In: . [S.l.]: Springer Netherlands, 2009. v. 12, n. 5, p. 526–558. ISSN 1386-4564. Citado na página 46.
- BOLLEGALA, D.; WEIR, D.; CARROLL, J. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In: **HLT '11 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: ACL, 2011. p. 132–141. Citado na página 68.
- BORGMAN, C. L. Multi-media, multi-cultural, and multilingual digital libraries: Or how do we exchange data in 400 languages? **D-Lib Magazine**, v. 3, n. 6, 1997. Citado na página 131.
- BREVE, F. A.; ZHAO, L.; QUILES, M. G.; PEDRYCZ, W.; LIU, J. Particle competition and cooperation in networks for semi-supervised learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 24, n. 9, p. 1686–1698, 2012. Citado nas páginas 48, 50 e 55.
- BRODY, S.; ELHADAD, N. An unsupervised aspect-sentiment model for online reviews. In: **HLT '10 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2010. p. 804–812. Citado na página 95.
- CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In: **ACM. Proceedings of the 23rd international conference on Machine learning**. [S.l.], 2006. p. 161–168. Citado na página 79.

- CHAKRABARTI, S.; DOM, B.; INDYK, P. Enhanced hypertext categorization using hyperlinks. **SIGMOD Record**, ACM, v. 27, n. 2, p. 307–318, 1998. ISSN 0163-5808. Citado na página 56.
- CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. (Ed.). **Semi-Supervised Learning**. [S.l.]: MIT Press, 2006. Citado nas páginas 27, 53 e 72.
- CHEN, Z.; MUKHERJEE, A.; LIU, B. Aspect extraction with automated prior knowledge learning. In: **ACL '14 - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. p. 347–358. Citado nas páginas 28, 67, 68, 111 e 126.
- CHOI, Y.; CARDIE, C. Hierarchical sequential learning for extracting opinions and their attributes. In: **ACLShort '10 - Proceedings of the Association for Computational Linguistics Conference Short Papers**. [S.l.]: ACL, 2010. p. 269–274. Citado nas páginas 28 e 44.
- CHOWDHURY, G. G. Natural language processing. **Annual review of information science and technology**, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003. Citado na página 131.
- COZMAN, F. G.; COHEN, I.; CIRELO, M. C. Semi-supervised learning of mixture models. In: **Proceedings of the 20th International Conference on Machine Learning (ICML-03)**. [S.l.: s.n.], 2003. p. 99–106. Citado na página 56.
- CULP, M.; MICHAILIDIS, G. An iterative algorithm for extending learners to a semi-supervised setting. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 17, n. 3, p. 545–571, 2008. Citado nas páginas 50 e 53.
- DAS, S.; CHEN, M. Yahoo! for amazon: Sentiment extraction from small talk on the web. In: **Management Science**. [S.l.: s.n.], 2007. v. 53(9). Citado na página 36.
- DELALLEAU, O.; BENGIO, Y.; ROUX, N. L. Efficient non-parametric function induction in semi-supervised learning. In: **AISTATS**. [S.l.: s.n.], 2005. v. 27, n. 28, p. 100. Citado na página 57.
- DELLAROCAS, X. Z. C.; AWAD, N. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. In: **Journal of Interactive Marketing**. [S.l.: s.n.], 2007. v. 21(4). Citado na página 36.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society. Series B (methodological)**, JSTOR, p. 1–38, 1977. Citado na página 54.
- DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006. Citado nas páginas 81, 83, 100 e 104.
- DENG, Z.-H.; LUO, K.-H.; YU, H.-L. A study of supervised term weighting scheme for sentiment analysis. **Expert Systems with Applications**, v. 41, n. 7, p. 3506 – 3513, 2014. Citado nas páginas 93 e 94.
- DING, X.; LIU, B.; YU, P. S. A holistic lexicon-based approach to opinion mining. In: **Proceedings of the 2008 International Conference on Web Search and Data Mining**. [S.l.]: ACM, 2008. p. 231–240. ISBN 978-1-59593-927-2. Citado nas páginas 45 e 46.

- ELWORTHY, D. Does baum-welch re-estimation help taggers? In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the fourth conference on Applied natural language processing**. [S.l.], 1994. p. 53–58. Citado na página 56.
- FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, ACM, v. 56, n. 4, p. 82–89, 2013. Citado na página 109.
- FERREIRA, R.; FREITAS, F.; CABRAL, L. d. S.; LINS, R. D.; LIMA, R.; FRANÇA, G.; SIMSKEZ, S. J.; FAVARO, L. A four dimension graph model for automatic text summarization. In: IEEE COMPUTER SOCIETY. **Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01**. [S.l.], 2013. p. 389–396. Citado na página 48.
- FLACH, P. **Machine learning: the art and science of algorithms that make sense of data**. [S.l.]: Cambridge University Press, 2012. Citado na página 27.
- FLORENCE, R.; NOGUEIRA, B.; MARCACINI, R. Constrained hierarchical clustering for news events. In: **IDEAS '17 - Proceedings of the 21st International Database Engineering & Applications Symposium**. [S.l.]: ACM, 2017. p. 49–56. Citado na página 121.
- GAMMERMAN, A.; VOVK, V.; VAPNIK, V. Learning by transduction. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence**. [S.l.], 1998. p. 148–155. Citado na página 53.
- GANAPATHIBHOTLA, M.; LIU, B. Mining opinions in comparative sentences. In: **Proceedings of the 22Nd International Conference on Computational Linguistics**. [S.l.]: ACL, 2008. p. 241–248. Citado na página 46.
- GANESHBHAI, S. Y.; SHAH, B. K. Feature based opinion mining: A survey. In: **IACC '15 - IEEE International Advance Computing Conference**. [S.l.]: IEEE, 2015. p. 919–923. Citado na página 68.
- GANTZ, J.; REINSEL, D. The digital universe decade – are you ready? In: **External Publication of IDC (Analyse the Future) Information and Data**. [S.l.: s.n.], 2010. Citado na página 35.
- GARCIA, S.; HERRERA, F. An extension on “statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons. **Journal of Machine Learning Research**, v. 9, p. 2677–2694, 2008. Citado nas páginas 100 e 104.
- GEE, K. R.; COOK, D. J. Text classification using graph-encoded linguistic elements. In: **FLAIRS Conference**. [S.l.: s.n.], 2005. p. 487–492. Citado na página 48.
- GHANI, R.; PROBST, K.; LIU, Y.; KREMA, M.; FANO, A. Text mining for product attribute extraction. **SIGKDD Explor. Newsl.**, ACM, v. 8, n. 1, p. 41–48, jun. 2006. Citado na página 44.
- GHOSE, P. G. I. A.; SUNDARARAJAN, A. Opinion mining using econometrics: A case study on reputation systems. In: **ACL '07 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2007. Citado na página 36.
- GUO, H.; ZHU, H.; GUO, Z.; ZHANG, X.; SU, Z. Product feature categorization with multilevel latent semantic association. In: **Proceedings of the 18th Conference on Information and Knowledge Management**. [S.l.: s.n.], 2009. p. 1087–1096. Citado na página 44.

_____. Opinionit: A text mining system for cross-lingual opinion analysis. In: **CIKM '10 - Proceedings of the 19th ACM International Conference on Information and Knowledge Management**. [S.l.: s.n.], 2010. p. 1199–1208. ISBN 978-1-4503-0099-5. Citado nas páginas [131](#) e [132](#).

HAFFARI, G.; SARKAR, A. Analysis of semi-supervised learning with the yarowsky algorithm. In: AUAI PRESS. **Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence**. [S.l.], 2007. p. 159–166. Citado na página [53](#).

HAI, Z.; CHANG, K.; CONG, G. One seed to find them all: mining opinion features via association. In: ACM. **Proceedings of the 21st ACM international conference on Information and knowledge management**. [S.l.], 2012. p. 255–264. Citado na página [43](#).

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.]: ACM, 2012. p. 2–3. ISBN 978-1-4503-1462-6. Citado na página [55](#).

HENSMAN, S. Construction of conceptual graph representation of texts. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Student Research Workshop at HLT-NAACL 2004**. [S.l.], 2004. p. 49–54. Citado na página [48](#).

HOBBS, J. R.; RILOFF, E. Information extraction. **Handbook of natural language processing**, CRC Press Boca Raton, FL, v. 2, 2010. Citado na página [43](#).

HOFMANN, T. Probabilistic latent semantic indexing. In: **SIGIR '99 - Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.]: ACM, 1999. p. 50–57. Citado nas páginas [44](#) e [68](#).

HOLMES, G.; DONKIN, A.; WITTEN, I. Weka: A machine learning workbench. In: **AN-ZIIS '94 - Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems**. [S.l.: s.n.], 1994. p. 357–361. Citado na página [135](#).

HU, M.; LIU, B. Mining and summarizing customer reviews. In: **KDD '04 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2004. p. 168–177. Citado nas páginas [27](#), [37](#), [42](#), [46](#), [91](#), [101](#), [102](#), [110](#), [116](#), [132](#), [133](#) e [134](#).

HU, N.; PAVLOU, P. A.; ZHANG, J. Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. In: **EC '06 - Proceedings of the 7th ACM Conference on Electronic Commerce**. [S.l.]: ACM, 2006. p. 324–330. Citado na página [36](#).

JAKOB, N.; GUREVYCH, I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: **EMNLP '10 - Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. [S.l.]: ACL, 2010. p. 1035–1045. Citado nas páginas [28](#) e [43](#).

JENSEN, D.; NEVILLE, J.; GALLAGHER, B. Why collective inference improves relational classification. In: **KDD '04 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2004. p. 593–598. Citado na página [56](#).

- JI, M.; SUN, Y.; DANILEVSKY, M.; HAN, J.; GAO, J. Graph regularized transductive classification on heterogeneous information networks. In: **ECML PKDD'10 - Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part I**. [S.l.]: Springer-Verlag, 2010. p. 570–586. Citado nas páginas [48](#), [60](#), [62](#), [70](#), [72](#) e [76](#).
- JIANG, L.; YU, M.; ZHOU, M.; LIU, X.; ZHAO, T. Target-dependent twitter sentiment classification. In: **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: ACL, 2011. p. 151–160. Citado na página [46](#).
- JIMÉNEZ-ZAFRA, S. M.; MARTÍN-VALDIVIA, M. T.; MARTÍNEZ-CÁMARA, E.; UREÑA-LÓPEZ, L. A. Combining resources to improve unsupervised sentiment analysis at aspect-level. **Journal of Information Science**, SAGE Publications, v. 42, n. 2, p. 213–229, 2016. Citado nas páginas [26](#), [67](#) e [68](#).
- JIN, W.; HO, H. H.; SRIHARI, R. K. A novel lexicalized hmm-based learning framework for web opinion mining. In: **ICML '09 - Proceedings of the 26th Annual International Conference on Machine Learning**. [S.l.]: ACM, 2009. p. 465–472. Citado nas páginas [28](#) e [44](#).
- JINDAL, N.; LIU, B. Mining comparative sentences and relations. In: **Proceedings of AAAI**. [S.l.: s.n.], 2006. v. 22, p. 1331–1336. Citado nas páginas [28](#) e [44](#).
- JOACHIMS, T. Transductive inference for text classification using support vector machines. In: **ICML**. [S.l.: s.n.], 1999. v. 99, p. 200–209. Citado na página [54](#).
- JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. [S.l.]: Prentice Hall, 2000. Citado na página [131](#).
- KAJI, N.; KITSUREGAWA, M. Building lexicon for sentiment analysis from massive collection of html documents. In: **EMNLP-CoNLL '07 - Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.: s.n.], 2007. p. 1075–1083. Citado na página [94](#).
- KESSLER, J. S.; NICOLOV, N. Targeting sentiment expressions through supervised ranking of linguistic configurations. In: **ICWSM**. [S.l.: s.n.], 2009. Citado na página [46](#).
- KIM, S.-M.; HOVY, E. Determining the sentiment of opinions. In: **Proceedings of the 20th International Conference on Computational Linguistics**. [S.l.: s.n.], 2004. Citado na página [46](#).
- KLINGER, R.; CIMIANO, P. Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model. In: **ACL'13 (2)**. [S.l.: s.n.], 2013. p. 848–854. Citado na página [44](#).
- KOBAYASHI, N.; INUI, K.; MATSUMOTO, Y. Extracting aspect-evaluation and aspect-of relations in opinion mining. In: **EMNLP-CoNLL '07 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.: s.n.], 2007. p. 1065–1074. Citado nas páginas [27](#), [43](#) e [44](#).
- KOCH, K.-R. Bayes' theorem. In: **Bayesian Inference with Geodetic Applications**. [S.l.]: Springer, 1990. p. 4–8. Citado na página [135](#).

KOVELAMUDI, S.; RAMALINGAM, S.; SOOD, A.; VARMA, V. Domain independent model for product attribute extraction from user reviews using wikipedia. In: **IJCNLP**. [S.l.: s.n.], 2011. p. 1408–1412. Citado na página 44.

KSCHISCHANG, F. R.; FREY, B. J.; LOELIGER, H.-A. Factor graphs and the sum-product algorithm. **Information Theory, IEEE Transactions on**, IEEE, v. 47, n. 2, p. 498–519, 2001. Citado na página 44.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. Citado na página 43.

LAU, R. Y.; LI, C.; LIAO, S. S. Social analytics: learning fuzzy product ontologies for aspect-oriented sentiment analysis. **Decision Support Systems**, Elsevier, v. 65, p. 80–94, 2014. Citado na página 109.

LI, F.; HAN, C.; HUANG, M.; ZHU, X.; XIA, Y.-J.; ZHANG, S.; YU, H. Structure-aware review mining and summarization. In: **Proceedings of the 23rd International Conference on Computational Linguistics**. [S.l.]: ACL, 2010. p. 653–661. Citado na página 43.

LI, S.; XUE, Y.; WANG, Z.; ZHOU, G. Active learning for cross-domain sentiment classification. In: **IJCAI '13 - Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2013. p. 2127–2133. Citado na página 112.

LIU, B. **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 3540378812. Citado nas páginas 37 e 42.

_____. Sentiment analysis and subjectivity. **Handbook of natural language processing**, Chapman & Hall Boca Raton, FL, v. 2, p. 627–666, 2010. Citado na página 45.

_____. Sentiment analysis and opinion mining. **Morgan & Claypool Publishers**, 2012. Citado nas páginas 25, 35, 36, 40, 41, 45, 46, 50 e 68.

_____. **Sentiment Analysis - Mining Opinions, Sentiments, and Emotions**. [S.l.]: Cambridge University Press, 2015. Citado nas páginas 26, 36, 67, 68, 71, 109 e 131.

LIU, B.; HU, M.; CHENG, J. Opinion observer: Analyzing and comparing opinions on the web. In: **WWW '05 - Proceedings of the 14th International Conference on World Wide Web**. [S.l.]: ACM, 2005. p. 342–351. ISBN 1-59593-046-9. Citado nas páginas 28 e 44.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: **Mining text data**. [S.l.]: Springer, 2012. p. 415–463. Citado nas páginas 25 e 109.

LIU, Q.; GAO, Z.; LIU, B.; ZHANG, Y. A logic programming approach to aspect extraction in opinion mining. In: **IEEE. IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (2013)**. [S.l.], 2013. v. 1, p. 276–283. Citado nas páginas 90 e 91.

LONG, C.; ZHANG, J.; ZHUT, X. A review selection approach for accurate feature rating estimation. In: **Proceedings of the 23rd International Conference on Computational Linguistics**. [S.l.]: ACL, 2010. p. 766–774. Citado na página 42.

- LONG, M.; WANG, J.; DING, G.; PAN, S. J.; PHILIP, S. Y. Adaptation regularization: A general framework for transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 26, n. 5, p. 1076–1089, 2014. Citado na página 111.
- LU, J.; BEHBOOD, V.; HAO, P.; ZUO, H.; XUE, S.; ZHANG, G. Transfer learning using computational intelligence: a survey. **Knowledge-Based Systems**, Elsevier, v. 80, p. 14–23, 2015. Citado na página 111.
- LU, Q.; GETOOR, L. Link-based classification. In: **Proceedings of the 20th International Conference on Machine Learning (ICML-03)**. [S.l.: s.n.], 2003. p. 496–503. Citado na página 48.
- LU, Y.; ZHAI, C.; SUNDARESAN, N. Rated aspect summarization of short comments. In: **Proceedings of the 18th International Conference on World Wide Web**. [S.l.]: ACM, 2009. p. 131–140. Citado na página 44.
- LUO, P.; ZHUANG, F.; XIONG, H.; XIONG, Y.; HE, Q. Transfer learning from multiple source domains via consensus regularization. In: **CIKM '08 - Proceedings of the 17th ACM Conference on Information and Knowledge Management**. [S.l.]: ACM, 2008. p. 103–112. Citado na página 112.
- MANEVITZ, L. M.; YOUSEF, M. One-class svms for document classification. **J. Mach. Learn. Res.**, JMLR.org, v. 2, p. 139–154, mar. 2002. Citado na página 44.
- MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J. R.; BETHARD, S.; MCCLOSKEY, D. The stanford corenlp natural language processing toolkit. In: . [S.l.]: The Association for Computer Linguistics, 2014. p. 55–60. Citado nas páginas 71, 101 e 138.
- MARCACINI, R. M.; MATSUNO, I. P.; ROSSI, R. G.; REZENDE, S. O. Cross-domain aspect extraction for sentiment analysis: a transductive learning approach. **Decision Support Systems**, 2018. Citado na página 123.
- MARCACINI, R. M.; SINOARA, R. A.; MATSUNO, I. P.; REZENDE, S. O. Aprendizado não supervisionado de websensors a partir de dados textuais. In: **KDMiLe '13 - Symposium on Knowledge Discovery, Mining and Learning**. [S.l.: s.n.], 2013. Citado na página 120.
- MARNEFFE, M.-C. D.; MANNING, C. D. **Stanford typed dependencies manual**. [S.l.], 2008. Citado na página 138.
- 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: **ICED '15 - Proceedings of the 20th International Conference on Engineering Design**. [S.l.: s.n.], 2015. Citado na página 152.
- MATSUNO, I. P.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Análise de sentimentos baseada em aspectos usando aprendizado semissupervisionado em redes heterogêneas. In: **KDMiLe '15 - Proceedings of the Symposium on Knowledge Discovery, Mining and Learning**. [S.l.: s.n.], 2015. p. 90–93. Citado na página 87.
- _____. Aspect-based sentiment analysis using semi-supervised learning in bipartite heterogeneous networks. **Journal of Information and Data Management**, v. 7, n. 2, p. 141, 2016. Citado nas páginas 87, 109 e 139.

_____. Towards domain-independent aspect extraction methods for sentiment analysis. **Expert Systems with Applications**, v. 7, n. 2, p. 141, 2018. Citado na página 108.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, Elsevier, v. 5, n. 4, p. 1093–1113, 2014. ISSN 2090-4479. Citado nas páginas 93 e 94.

MEI, Q.; LING, X.; WONDRA, M.; SU, H.; ZHAI, C. Topic sentiment mixture: Modeling facets and opinions in weblogs. In: **Proceedings of the 16th International Conference on World Wide Web**. [S.l.]: ACM, 2007. p. 171–180. Citado na página 44.

MISHRA, M.; HUAN, J.; BLEIK, S.; SONG, M. Biomedical text categorization with concept graph representations using a controlled vocabulary. In: ACM. **Proceedings of the 11th International Workshop on Data Mining in Bioinformatics**. [S.l.], 2012. p. 26–32. Citado na página 48.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw Hill, 1997. Citado nas páginas 27, 49, 50 e 136.

MOGHADDAM, S.; ESTER, M. On the Design of LDA Models for Aspect-based Opinion Mining. In: **CIKM '12 - Proceedings of the 21st ACM International Conference on Information and Knowledge Management**. [S.l.]: ACM, 2012. p. 803–812. ISBN 978-1-4503-1156-4. Citado na página 68.

MOONEY, R. J.; BUNESCU, R. Mining knowledge from text using information extraction. **SIGKDD Explor. Newsl.**, ACM, v. 7, n. 1, p. 3–10, jun. 2005. ISSN 1931-0145. Citado na página 43.

MUBAROK, M. S.; ADIWIJAYA; ALDHI, M. D. Aspect-based sentiment analysis to review products using naïve bayes. In: **AIP Conference Proceedings (2017)**. [S.l.]: AIP Publishing, 2017. p. 020060. Citado na página 94.

MUKHERJEE, A.; LIU, B. Aspect extraction through semi-supervised modeling. In: **ACL '12 - Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers**. [S.l.]: ACL, 2012. p. 339–348. Citado na página 109.

NEWMAN, M. **Networks: An Introduction**. [S.l.]: Oxford University Press, Inc., 2010. Citado nas páginas 48 e 55.

NIGAM, K.; GHANI, R. Analyzing the effectiveness and applicability of co-training. In: **Proceedings of the Ninth International Conference on Information and Knowledge Management**. [S.l.]: ACM, 2000. p. 86–93. Citado na página 54.

NIGAM, K.; MCCALLUM, A. K.; THRUN, S.; MITCHELL, T. Text classification from labeled and unlabeled documents using EM. **Machine Learning**, Kluwer Academic Publishers, v. 39, n. 2-3, p. 103–134, 2000. Citado na página 79.

OH, H.; MYAENG, S.; LEE, M. A practical hypertext categorization method using links and incrementally available class information. In: ACM. **SIGIR '00 - Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.], 2000. p. 264–271. Citado na página 56.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 22, n. 10, p. 1345–1359, 2010. Citado nas páginas 28, 110 e 111.

- PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Found. Trends Inf. Retr.**, Now Publishers Inc., v. 2, n. 1-2, p. 1–135, jan. 2008. Citado na página [25](#).
- PANG, L. L. B.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: **EMNLP '02 - Proceedings of Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2002. Citado na página [36](#).
- PARK, D.-H.; LEE, J.; HAN, I. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. **International J. Electron. Commerce**, M. E. Sharpe, Inc., Armonk, NY, USA, v. 11, n. 4, p. 125–148, jul. 2007. Citado na página [36](#).
- PONTIKI, M.; GALANIS, D.; PAVLOPOULOS, J.; PAPAGEORGIOU, H.; ANDROUTSOPOULOS, I.; MANANDHAR, S. Semeval-2014 task 4: Aspect based sentiment analysis. In: **Workshop on Semantic Evaluation**. [S.l.: s.n.], 2014. p. 27–35. Citado nas páginas [68](#), [77](#), [101](#) e [116](#).
- POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. In: **HLT '05 - Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing**. [S.l.]: ACL, 2005. p. 339–346. Citado nas páginas [42](#) e [94](#).
- QIU, G.; LIU, B.; BU, J.; CHEN, C. Expanding domain sentiment lexicon through double propagation. In: **IJCAI '09 - Proceedings of the 21st International Joint Conference on Artificial Intelligence**. [S.l.]: Morgan Kaufmann Publishers Inc., 2009. p. 1199–1204. Citado nas páginas [43](#), [44](#) e [138](#).
- _____. Opinion word expansion and target extraction through double propagation. **Computational linguistics**, MIT Press, v. 37, n. 1, p. 9–27, 2011. Citado nas páginas [27](#), [43](#) e [44](#).
- QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. Citado na página [52](#).
- RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, IEEE, v. 77, n. 2, p. 257–286, 1989. Citado na página [43](#).
- RANA, T. A.; CHEAH, Y.-N. Aspect extraction in sentiment analysis: comparative analysis and survey. **Artificial Intelligence Review**, Springer, v. 46, n. 4, p. 459–483, 2016. Citado nas páginas [109](#) e [110](#).
- _____. A two-fold rule-based model for aspect extraction. **Expert Systems with Applications**, Elsevier, v. 89, p. 273–285, 2017. Citado nas páginas [28](#), [110](#), [111](#) e [112](#).
- RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. **Knowledge-Based Systems**, Elsevier, v. 89, p. 14–46, 2015. Citado nas páginas [93](#) e [94](#).
- RISH, I. An empirical study of the naive bayes classifier. In: **IJCAI '01 - Workshop Empirical Methods in Artificial Intelligence**. [S.l.]: IBM New York, 2001. v. 3, n. 22, p. 41–46. Citado na página [135](#).
- ROBERTS, A.; GAIZAUSKAS, R.; HEPPLER, M.; DEMETRIOU, G.; GUO, Y.; ROBERTS, I.; SETZER, A. Building a semantically annotated corpus of clinical texts. **Journal of biomedical informatics**, Elsevier, v. 42, n. 5, p. 950–966, 2009. Citado na página [126](#).

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado nas páginas 49 e 56.

ROSSI, R. G.; LOPES, A. A.; REZENDE, S. O. A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In: **SAC '14 - Proceedings of the 29th Annual ACM Symposium on Applied Computing**. [S.l.]: ACM, 2014. p. 79–84. Citado nas páginas 56, 59, 72 e 76.

ROSSI, R. G.; LOPES, A. d. 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 48, 56, 63, 70, 72, 76, 78 e 83.

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**, Springer, v. 3, n. 29, p. 361–375, 2014. Citado nas páginas 48 e 55.

ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. Benchmarking text collections for classification and clustering tasks. **Institute of Mathematics and Computer Sciences, University of Sao Paulo**, 2013. Citado nas páginas 77 e 116.

ROSSI, R. G.; REZENDE, S. O.; LOPES, A. de A. Term network approach for transductive classification. In: SPRINGER. **International Conference on Intelligent Text Processing and Computational Linguistics**. [S.l.], 2015. p. 497–515. Citado na página 48.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986. Citado na página 52.

SALTON, G. **Automatic text processing: the transformation, analysis, and retrieval of information by computer**. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1989. Citado nas páginas 47 e 55.

SANTOS, B. N. d.; ROSSI, R. G.; MARCACINI, R. M. Transductive event classification through heterogeneous networks. In: **WebMedia '17 - Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web**. [S.l.]: ACM, 2017. p. 285–292. ISBN 978-1-4503-5096-9. Citado na página 121.

SARAWAGI, S. Information extraction. **Found. Trends databases**, Now Publishers Inc., v. 1, n. 3, p. 261–377, mar. 2008. Citado na página 43.

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 na página 152.

SCHOUTEN, K.; FRASINCAR, F. Survey on aspect-level sentiment analysis. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 28, n. 3, p. 813–830, 2016. Citado nas páginas 28, 109 e 110.

SILVA, T. C.; ZHAO, L. Stochastic competitive learning in complex networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 3, p. 385–398, 2012. Citado na página 48.

- 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 [152](#).
- SMATANA, M.; KONCZ, P.; SMATANA, P.; PARALIC, J. Active learning enhanced semi-automatic annotation tool for aspect-based sentiment analysis. In: IEEE. **Intelligent Systems and Informatics (SISY), 2013 IEEE 11th International Symposium on**. [S.l.], 2013. p. 191–194. Citado na página [126](#).
- SØGAARD, A. **Semi-Supervised Learning and Domain Adaptation in Natural Language Processing**. [S.l.]: Morgan & Claypool Publishers, 2013. (Synthesis Lectures on Human Language Technologies). Citado na página [50](#).
- SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, Elsevier, v. 45, n. 4, p. 427–437, 2009. Citado na página [80](#).
- SOLÉ, R. V.; COROMINAS-MURTRA, B.; VALVERDE, S.; STEELS, L. Language networks: Their structure, function, and evolution. **Complexity**, Wiley Online Library, v. 15, n. 6, p. 20–26, 2010. Citado na página [48](#).
- SOMASUNDARAN, S.; NAMATA, G.; GETOOR, L.; WIEBE, J. Opinion graphs for polarity and discourse classification. In: **TextGraphs-4 - Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing**. [S.l.]: ACL, 2009. p. 66–74. Citado nas páginas [27](#) e [43](#).
- SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 241–247. Citado na página [134](#).
- SUBRAMANYA, A.; TALUKDAR, P. **Graph-Based Semi-Supervised Learning**. [S.l.]: Morgan & Claypool Publishers, 2014. (Synthesis Lectures on Artificial Intelligence and Machine Learning). Citado na página [50](#).
- SUN, Y.; HAN, J. **Mining Heterogeneous Information Networks: Principles and Methodologies**. [S.l.]: Morgan & Claypool Publishers, 2012. Citado nas páginas [48](#) e [55](#).
- TAN, P.-N.; KUMAR, V.; SRIVASTAVA, J. Selecting the right objective measure for association analysis. **Information Systems**, Elsevier, v. 29, n. 4, p. 293–313, 2004. Citado na página [92](#).
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2005. ISBN 0-321-32136-7. Citado nas páginas [47](#) e [51](#).
- TAN, S.; WU, G.; TANG, H.; CHENG, X. A novel scheme for domain-transfer problem in the context of sentiment analysis. In: ACM. **Proceedings of the sixteenth ACM conference on Conference on information and knowledge management**. [S.l.], 2007. p. 979–982. Citado na página [47](#).
- TITOV, I.; MCDONALD, R. Modeling online reviews with multi-grain topic models. In: **WWW '08 - Proceedings of the 17th International Conference on World Wide Web**. [S.l.]: ACM, 2008. p. 111–120. Citado nas páginas [44](#), [68](#) e [95](#).

TITOV, I.; MCDONALD, R. T. A joint model of text and aspect ratings for sentiment summarization. In: **ACL '08 - Proceedings of the Association for Computational Linguistics Conference**. [S.l.]: ACL, 2008. v. 8, p. 308–316. Citado nas páginas [44](#), [68](#) e [95](#).

TURNEY, P. D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: **ACL '02 - Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2002. p. 417–424. Citado nas páginas [36](#), [93](#), [132](#), [134](#) e [135](#).

VAPNIK, V.; VASHIST, A. A new learning paradigm: Learning using privileged information. **Neural networks**, Elsevier, v. 22, n. 5-6, p. 544–557, 2009. Citado na página [54](#).

VAPNIK, V. N. The nature of statistical learning theory. Springer-Verlag New York, Inc., 1995. Citado na página [52](#).

WAN, X. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In: **Proceedings of the Conference on Empirical Methods in Natural Language Processing**. [S.l.]: ACL, 2008. p. 553–561. Citado na página [46](#).

WANG, B.; WANG, H. Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing. In: **IJCNLP**. [S.l.: s.n.], 2008. p. 289–295. Citado na página [43](#).

WANG, F.; ZHANG, C. Label propagation through linear neighborhoods. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 20, n. 1, p. 55–67, 2008. Citado na página [57](#).

WANG, H.; TSOU, B. K.; ZHU, M.; ZHU, J.; MA, M. Aspect-based opinion polling from customer reviews. **IEEE Transactions on Affective Computing**, IEEE Computer Society, Los Alamitos, CA, USA, v. 2, p. 37–49, 2011. Citado na página [138](#).

WANG, T.; CAI, Y.; LEUNG, H.-f.; LAU, R. Y.; LI, Q.; MIN, H. Product aspect extraction supervised with online domain knowledge. **Knowledge-Based Systems**, Elsevier, v. 71, p. 86–100, 2014. Citado na página [109](#).

WANG W. E ZHOU, Y. E-business websites evaluation based on opinion mining. In: **2009 International Conference on Electronic Commerce and Business Intelligence**. [S.l.]: IEEE, 2009. p. 87–90. Citado na página [35](#).

WEI, W.; GULLA, J. A. Sentiment learning on product reviews via sentiment ontology tree. In: **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2010. p. 404–413. Citado na página [46](#).

WIEBE, J. M.; BRUCE, R. F.; O'HARA, T. P. Development and use of a gold-standard data set for subjectivity classifications. In: **ACL '99 - Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 1999. p. 246–253. Citado na página [36](#).

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In: **HLT '05 - Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2005. p. 347–354. Citado na página [134](#).

- WITTEN, I.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques, Second Edition**. [S.l.]: Elsevier Science, 2005. (The Morgan Kaufmann Series in Data Management Systems). Citado na página 51.
- WU, F.; HUANG, Y.; YAN, J. Active sentiment domain adaptation. In: **ACL '17 - Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2017. v. 1, p. 1701–1711. Citado na página 112.
- WU, Q.; TAN, S. A two-stage framework for cross-domain sentiment classification. **Expert Systems with Applications**, Elsevier, v. 38, n. 11, p. 14269–14275, 2011. Citado na página 111.
- XIA, F.; YETISGEN-YILDIZ, M. Clinical corpus annotation: challenges and strategies. In: **Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey**. [S.l.: s.n.], 2012. Citado na página 126.
- XU, Z.; KING, I. **Introduction to Semi-Supervised Learning**. [S.l.]: CRC Press, 2015. (Chapman & Hall/CRC Machine Learning & Pattern Recognition). Citado na página 50.
- XUE, G.-R.; SHEN, D.; YANG, Q.; ZENG, H.-J.; CHEN, Z.; YU, Y.; XI, W. S.; MA, W.-Y. Irc: An iterative reinforcement categorization algorithm for interrelated web objects. In: **IEEE Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on**. [S.l.], 2004. p. 273–280. Citado na página 48.
- YANG, B.; CARDIE, C. Joint inference for fine-grained opinion extraction. In: **ACL '13 - Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**. [S.l.]: ACL, 2013. p. 1640–1649. Citado nas páginas 28 e 43.
- YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: **ICML '97 - Proceedings of the 14th International Conference on Machine Learning**. [S.l.: s.n.], 1997. v. 97, p. 412–420. Citado na página 94.
- YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. In: **ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the 33rd annual meeting on Association for Computational Linguistics**. [S.l.], 1995. p. 189–196. Citado na página 53.
- YIN, Z.; LI, R.; MEI, Q.; HAN, J. Exploring social tagging graph for web object classification. In: **KDD '09 - Proceedings of the International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2009. p. 957–966. Citado nas páginas 72 e 76.
- YU, J.; ZHA, Z.-J.; WANG, M.; CHUA, T.-S. Aspect ranking: Identifying important product aspects from online consumer reviews. In: **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: ACL, 2011. p. 1496–1505. Citado na página 44.
- ZHAI, Z.; LIU, B.; ZHANG, L.; XU, H.; JIA, P. Identifying evaluative opinions in online discussions. In: **Proceedings of AAAI**. [S.l.: s.n.], 2011. p. 7–11. Citado na página 43.
- ZHANG, L.; LIU, B. Extracting resource terms for sentiment analysis. In: **IJCNLP '11 - Proceedings of 5th International Joint Conference on Natural Language Processing**. [S.l.]: Asian Federation of Natural Language Processing, 2011. p. 1171–1179. Citado na página 68.

- _____. Aspect and entity extraction for opinion mining. In: **Data mining and knowledge discovery for big data**. [S.l.]: Springer, 2014. p. 1–40. Citado nas páginas 89 e 91.
- ZHANG, P.; WANG, J.; WANG, Y.; WANG, Y. A statistical approach to opinion target extraction using domain relevance. In: IEEE. **ICCC '16 - 2nd IEEE International Conference on Computer and Communications**. [S.l.], 2016. p. 273–277. Citado nas páginas 28, 110 e 112.
- ZHAO, W. X.; JIANG, J.; YAN, H.; LI, X. Jointly modeling aspects and opinions with a maxent-lda hybrid. In: **EMNLP '10 - Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. [S.l.]: ACL, 2010. p. 56–65. Citado nas páginas 45, 68 e 95.
- ZHOU, D.; BOUSQUET, O.; LAL, T. N.; WESTON, J.; SCHÖLKOPF, B. Learning with local and global consistency. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2004. v. 16, p. 321–328. Citado nas páginas 59 e 76.
- ZHOU, S.; CHEN, Q.; WANG, X. Active deep networks for semi-supervised sentiment classification. In: ACL. **Proceedings of the 23rd International Conference on Computational Linguistics**. [S.l.], 2010. p. 1515–1523. Citado na página 113.
- ZHU, J.; WANG, H.; TSOU, B. K.; ZHU, M. Multi-aspect opinion polling from textual reviews. In: **Proceedings of the 18th ACM Conference on Information and Knowledge Management**. [S.l.]: ACM, 2009. p. 1799–1802. Citado na página 46.
- ZHU, X. Semi-supervised learning literature survey. Citeseer, 2005. Citado na página 57.
- ZHU, X.; GHAAHRAMANI, Z.; LAFFERTY, J. Semi-supervised learning using gaussian fields and harmonic functions. In: **ICML '03 - Proceedings of the Twentieth International Conference on International Conference on Machine Learning**. [S.l.]: AAAI Press, 2003. p. 912–919. Citado nas páginas 58 e 76.
- ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado na página 53.
- ZHU, X.; GOLDBERG, A. B.; BRACHMAN, R.; DIETTERICH, T. **Introduction to semi-supervised learning**. [S.l.]: Morgan and Claypool Publishers, 2009. Citado nas páginas 50, 53, 56, 72 e 79.
- ZHUANG, F.; LUO, P.; XIONG, H.; XIONG, Y.; HE, Q.; SHI, Z. Cross-domain learning from multiple sources: A consensus regularization perspective. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 22, n. 12, p. 1664–1678, 2010. Citado na página 112.
- ZHUANG, L.; JING, F.; ZHU, X.-Y. Movie review mining and summarization. In: **CIKM '06 - Proceedings of the 15th ACM International Conference on Information and Knowledge Management**. [S.l.]: ACM, 2006. p. 43–50. Citado nas páginas 27 e 42.
- ZHUGE, H. **The Knowledge Grid: Toward Cyber-Physical Society**. [S.l.]: World Scientific, 2012. Citado na página 120.

RESULTADOS – EXTRAÇÃO DE ASPECTOS NÃO SUPERVISIONADA E INDEPENDENTE DE DOMÍNIO PARA MINERAÇÃO DE OPINIÕES

Este apêndice contém os resultados obtidos na avaliação experimental que foi apresentada no Capítulo 4. Na sequência são apresentados os resultados para as medidas de Precisão, Cobertura e F^1 para cada método de extração de aspecto avaliado com as respectivas variações de parâmetros. Esses resultados também são divididos pelo tipo de correspondência entre termos utilizado na avaliação: exato e aproximado.

Dataset

D1-CAM1

método	Precision- Aproximate Matching	Precision Exact Matching	RECALL- Aproximate Matching	RECALL Exact Matching	F1- Aproximate Matching	F1-Exact Matching
1gram-NN-freq01	0.06	0.05	0.78	0.57	0.12	0.10
1gram-NN-freq03	0.21	0.16	0.60	0.43	0.31	0.23
1gram-NN-freq05	0.29	0.22	0.40	0.29	0.34	0.25
1gram-NN-freq10	0.42	0.31	0.23	0.18	0.30	0.22
2gram-ALGUM-NN-freq01	0.03	0.02	0.28	0.21	0.05	0.04
2gram-ALGUM-NN-freq03	0.28	0.22	0.13	0.10	0.18	0.14
2gram-ALGUM-NN-freq05	0.50	0.38	0.09	0.07	0.15	0.12
2gram-ALGUM-NN-freq10	0.71	0.50	0.04	0.03	0.07	0.06
2gram-TODOS-NN-freq01	0.05	0.04	0.27	0.18	0.09	0.06
2gram-TODOS-NN-freq03	0.46	0.33	0.12	0.09	0.18	0.14
2gram-TODOS-NN-freq05	0.69	0.57	0.07	0.06	0.12	0.11
2gram-TODOS-NN-freq10	1.00	1.00	0.03	0.03	0.06	0.06
comb-1gram-2gram-ALGUM-NN-freq01	0.04	0.04	0.89	0.78	0.08	0.07
comb-1gram-2gram-ALGUM-NN-freq03	0.20	0.17	0.66	0.53	0.31	0.26
comb-1gram-2gram-ALGUM-NN-freq05	0.29	0.24	0.47	0.37	0.36	0.29
comb-1gram-2gram-ALGUM-NN-freq10	0.43	0.33	0.26	0.21	0.32	0.25
comb-1gram-2gram-TODOS-NN-freq01	0.05	0.05	0.89	0.75	0.10	0.09
comb-1gram-2gram-TODOS-NN-freq03	0.21	0.18	0.66	0.51	0.32	0.26
comb-1gram-2gram-TODOS-NN-freq05	0.31	0.24	0.47	0.35	0.37	0.29
comb-1gram-2gram-TODOS-NN-freq10	0.43	0.34	0.25	0.21	0.32	0.26
Comb-1gram-2gram-all-NN-JJ	0.15	0.13	0.54	0.43	0.24	0.19
ODDS-1gram-freq05-2gram-avg-1--3dp	0.05	0.04	0.59	0.50	0.09	0.08
ODDS-1gram-freq05-2gram-avg-2--2dp	0.05	0.04	0.59	0.50	0.09	0.08
ODDS-1gram-freq05-2gram-avg-3--1dp	0.05	0.04	0.59	0.50	0.09	0.08
ODDS-1gram-freq05-2gram-avg-4-0dp	0.08	0.06	0.50	0.32	0.14	0.10
ODDS-1gram-freq05-2gram-avg-5-1dp	0.09	0.07	0.47	0.31	0.16	0.11
ODDS-1gram-freq05-2gram-avg-6-2dp	0.26	0.20	0.42	0.31	0.32	0.24
ODDS-1gram-freq05-2gram-avg-7-3dp	0.26	0.20	0.42	0.31	0.32	0.24
CHI2-1gram-freq05-2gram-avg-1--3dp	0.05	0.04	0.60	0.50	0.09	0.08
CHI2-1gram-freq05-2gram-avg-2--2dp	0.05	0.04	0.60	0.50	0.09	0.08
CHI2-1gram-freq05-2gram-avg-3--1dp	0.05	0.04	0.60	0.50	0.09	0.08
CHI2-1gram-freq05-2gram-avg-4-0dp	0.09	0.07	0.51	0.34	0.15	0.11
CHI2-1gram-freq05-2gram-avg-5-1dp	0.13	0.10	0.44	0.31	0.20	0.15
CHI2-1gram-freq05-2gram-avg-6-2dp	0.29	0.22	0.41	0.29	0.34	0.25
CHI2-1gram-freq05-2gram-avg-7-3dp	0.29	0.22	0.41	0.29	0.34	0.25
PMI-1gram-freq05-2gram-avg-1--3dp	0.05	0.04	0.59	0.50	0.09	0.08
PMI-1gram-freq05-2gram-avg-2--2dp	0.05	0.04	0.59	0.50	0.09	0.08
PMI-1gram-freq05-2gram-avg-3--1dp	0.06	0.05	0.59	0.47	0.10	0.09
PMI-1gram-freq05-2gram-avg-4-0dp	0.07	0.04	0.53	0.31	0.12	0.08
PMI-1gram-freq05-2gram-avg-5-1dp	0.13	0.09	0.43	0.29	0.20	0.14
PMI-1gram-freq05-2gram-avg-6-2dp	0.29	0.22	0.40	0.29	0.34	0.25
PMI-1gram-freq05-2gram-avg-7-3dp	0.29	0.22	0.40	0.29	0.34	0.25
LDA-Tec1	0.07	0.05	0.32	0.21	0.11	0.08
LDA-Tec2	0.06	0.05	0.42	0.34	0.10	0.09

Dataset: D1-CAM1

MELHORES RESULTADOS

método	F1 Aproximate Matching	F1 Exact Matching
1gram-NN-freq	0.34	0.25
2gram-some-NN-freq	0.18	0.14
2gram-all-NN-freq	0.18	0.14
Comb-1gram-2gram-some-NN	0.36	0.29
Comb-1gram-2gram-all-NN	0.37	0.29
Comb-1gram-2gram-all-NN-JJ	0.24	0.19
ODDS	0.32	0.24
CHI2	0.34	0.25
PMI	0.34	0.25
LDA-Tec1	0.11	0.08
LDA-Tec2	0.10	0.09

Dataset

D2-CAM2

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1-Exact Matching
1gram-NN-freq01	0.06	0.04	0.76	0.43	0.11	0.08
1gram-NN-freq03	0.18	0.12	0.43	0.28	0.25	0.17
1gram-NN-freq05	0.29	0.18	0.30	0.19	0.30	0.19
1gram-NN-freq10	0.47	0.33	0.20	0.15	0.28	0.21
2gram-ALGUM-NN-freq01	0.05	0.04	0.44	0.32	0.09	0.07
2gram-ALGUM-NN-freq03	0.38	0.24	0.19	0.13	0.25	0.17
2gram-ALGUM-NN-freq05	0.50	0.29	0.13	0.09	0.21	0.13
2gram-ALGUM-NN-freq10	0.00	0.00	0.00	0.00	0.00	0.00
2gram-TODOS-NN-freq01	0.09	0.07	0.41	0.28	0.15	0.11
2gram-TODOS-NN-freq03	0.62	0.42	0.14	0.11	0.23	0.17
2gram-TODOS-NN-freq05	0.69	0.38	0.10	0.06	0.17	0.11
2gram-TODOS-NN-freq10	0.00	0.00	0.00	0.00	0.00	0.00
comb-1gram-2gram-ALGUM-NN-freq01	0.04	0.04	0.93	0.74	0.09	0.08
comb-1gram-2gram-ALGUM-NN-freq03	0.18	0.14	0.55	0.40	0.28	0.21
comb-1gram-2gram-ALGUM-NN-freq05	0.29	0.21	0.40	0.28	0.33	0.24
comb-1gram-2gram-ALGUM-NN-freq10	0.43	0.30	0.20	0.15	0.27	0.20
comb-1gram-2gram-TODOS-NN-freq01	0.06	0.05	0.93	0.70	0.11	0.09
comb-1gram-2gram-TODOS-NN-freq03	0.19	0.15	0.51	0.38	0.28	0.22
comb-1gram-2gram-TODOS-NN-freq05	0.30	0.21	0.37	0.26	0.33	0.23
comb-1gram-2gram-TODOS-NN-freq10	0.47	0.33	0.20	0.15	0.28	0.21
comb-1gram-freq05+2gram-todos-NN- freq03+1gram infreq com adj	0.15	0.11	0.51	0.36	0.23	0.17
ODDS-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-3--1dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-4-0dp	0.09	0.07	0.44	0.30	0.15	0.11
ODDS-1gram-freq05-2gram-avg-5-1dp	0.27	0.20	0.36	0.26	0.31	0.22
ODDS-1gram-freq05-2gram-avg-6-2dp	0.27	0.20	0.36	0.26	0.31	0.22
ODDS-1gram-freq05-2gram-avg-7-3dp	0.29	0.21	0.33	0.23	0.31	0.22
CHI2-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-3--1dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-4-0dp	0.10	0.07	0.44	0.30	0.16	0.11
CHI2-1gram-freq05-2gram-avg-5-1dp	0.13	0.09	0.41	0.26	0.20	0.13
CHI2-1gram-freq05-2gram-avg-6-2dp	0.29	0.18	0.30	0.19	0.30	0.19
CHI2-1gram-freq05-2gram-avg-7-3dp	0.29	0.18	0.30	0.19	0.30	0.19
PMI-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
PMI-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.12	0.10
PMI-1gram-freq05-2gram-avg-3--1dp	0.07	0.05	0.64	0.43	0.12	0.09
PMI-1gram-freq05-2gram-avg-4-0dp	0.08	0.05	0.54	0.28	0.15	0.09
PMI-1gram-freq05-2gram-avg-5-1dp	0.13	0.08	0.39	0.21	0.20	0.11
PMI-1gram-freq05-2gram-avg-6-2dp	0.29	0.18	0.30	0.19	0.30	0.19
PMI-1gram-freq05-2gram-avg-7-3dp	0.29	0.18	0.30	0.19	0.30	0.19
LDA-Tec1	0.11	0.07	0.50	0.30	0.17	0.11
LDA-Tec2	0.09	0.07	0.68	0.43	0.16	0.11

Dataset: D2-CAM2

MELHORES RESULTADOS

método	F1 Approximate Matching	F1 Exact Matching
1gram-NN-freq	0.30	0.21
2gram-some-NN-freq	0.25	0.17
2gram-all-NN-freq	0.23	0.17
Comb-1gram-2gram-some-NN	0.33	0.24
Comb-1gram-2gram-all-NN	0.33	0.23
Comb-1gram-2gram-all-NN-JJ	0.23	0.17
ODDS	0.31	0.22
CHI2	0.30	0.19
PMI	0.30	0.19
LDA-Tec1	0.17	0.11
LDA-Tec2	0.16	0.11

Dataset

D3-CEL

método	Precision- Aproximate Matching	Precision Exact Matching	RECALL- Aproximate Matching	RECALL Exact Matching	F1- Aproximate Matching	F1-Exact Matching
1gram-NN-freq01	0.09	0.07	0.84	0.59	0.16	0.13
1gram-NN-freq03	0.25	0.20	0.51	0.38	0.33	0.26
1gram-NN-freq05	0.39	0.32	0.39	0.32	0.39	0.32
1gram-NN-freq10	0.73	0.58	0.21	0.18	0.33	0.27
2gram-ALGUM-NN-freq01	0.04	0.03	0.29	0.22	0.06	0.05
2gram-ALGUM-NN-freq03	0.43	0.25	0.12	0.08	0.19	0.12
2gram-ALGUM-NN-freq05	1.00	0.83	0.07	0.06	0.13	0.12
2gram-ALGUM-NN-freq10	1.00	0.67	0.03	0.03	0.06	0.05
2gram-TODOS-NN-freq01	0.07	0.04	0.25	0.14	0.10	0.06
2gram-TODOS-NN-freq03	0.63	0.42	0.09	0.06	0.15	0.11
2gram-TODOS-NN-freq05	1.00	1.00	0.05	0.05	0.10	0.10
2gram-TODOS-NN-freq10	1.00	1.00	0.01	0.01	0.03	0.03
comb-1gram-2gram-ALGUM-NN-freq01	0.05	0.05	0.92	0.81	0.10	0.09
comb-1gram-2gram-ALGUM-NN-freq03	0.25	0.20	0.61	0.46	0.36	0.28
comb-1gram-2gram-ALGUM-NN-freq05	0.42	0.36	0.45	0.38	0.43	0.37
comb-1gram-2gram-ALGUM-NN-freq10	0.72	0.59	0.24	0.21	0.36	0.30
comb-1gram-2gram-TODOS-NN-freq01	0.07	0.06	0.91	0.73	0.13	0.11
comb-1gram-2gram-TODOS-NN-freq03	0.26	0.21	0.58	0.45	0.36	0.29
comb-1gram-2gram-TODOS-NN-freq05	0.41	0.35	0.43	0.37	0.42	0.36
comb-1gram-2gram-TODOS-NN-freq10	0.70	0.60	0.22	0.19	0.33	0.29
Comb-1gram-2gram-all-NN-JJ	0.21	0.18	0.61	0.47	0.32	0.26
ODDS-1gram-freq05-2gram-avg-1--3dp	0.07	0.06	0.63	0.54	0.12	0.11
ODDS-1gram-freq05-2gram-avg-2--2dp	0.07	0.06	0.63	0.54	0.12	0.11
ODDS-1gram-freq05-2gram-avg-3--1dp	0.07	0.06	0.63	0.54	0.12	0.11
ODDS-1gram-freq05-2gram-avg-4-0dp	0.14	0.12	0.57	0.47	0.22	0.19
ODDS-1gram-freq05-2gram-avg-5-1dp	0.37	0.32	0.47	0.40	0.42	0.35
ODDS-1gram-freq05-2gram-avg-6-2dp	0.37	0.32	0.46	0.40	0.41	0.36
ODDS-1gram-freq05-2gram-avg-7-3dp	0.39	0.33	0.41	0.35	0.40	0.34
CHI2-1gram-freq05-2gram-avg-1--3dp	0.07	0.06	0.63	0.54	0.12	0.11
CHI2-1gram-freq05-2gram-avg-2--2dp	0.07	0.06	0.63	0.54	0.12	0.11
CHI2-1gram-freq05-2gram-avg-3--1dp	0.07	0.06	0.63	0.54	0.12	0.11
CHI2-1gram-freq05-2gram-avg-4-0dp	0.15	0.12	0.57	0.46	0.23	0.20
CHI2-1gram-freq05-2gram-avg-5-1dp	0.17	0.14	0.46	0.36	0.25	0.20
CHI2-1gram-freq05-2gram-avg-6-2dp	0.39	0.32	0.39	0.32	0.39	0.32
CHI2-1gram-freq05-2gram-avg-7-3dp	0.39	0.32	0.39	0.32	0.39	0.32
PMI-1gram-freq05-2gram-avg-1--3dp	0.07	0.06	0.63	0.54	0.12	0.11
PMI-1gram-freq05-2gram-avg-2--2dp	0.07	0.06	0.63	0.54	0.13	0.11
PMI-1gram-freq05-2gram-avg-3--1dp	0.08	0.07	0.63	0.51	0.14	0.12
PMI-1gram-freq05-2gram-avg-4-0dp	0.10	0.08	0.58	0.42	0.17	0.13
PMI-1gram-freq05-2gram-avg-5-1dp	0.17	0.13	0.44	0.33	0.24	0.19
PMI-1gram-freq05-2gram-avg-6-2dp	0.39	0.32	0.39	0.32	0.39	0.32
PMI-1gram-freq05-2gram-avg-7-3dp	0.39	0.32	0.39	0.32	0.39	0.32
LDA-Tec1	0.10	0.09	0.43	0.36	0.16	0.14
LDA-Tec2	0.09	0.08	0.58	0.42	0.13	0.16

Dataset: D3-CEL

MELHORES RESULTADOS

método	F1 Aproximate Matching	F1 Exact Matching
1gram-NN-freq	0.39	0.32
2gram-some-NN-freq	0.19	0.12
2gram-all-NN-freq	0.15	0.11
Comb-1gram-2gram-some-NN	0.43	0.37
Comb-1gram-2gram-all-NN	0.42	0.36
Comb-1gram-2gram-all-NN-JJ	0.32	0.26
ODDS	0.42	0.36
CHI2	0.39	0.32
PMI	0.39	0.32
LDA-Tec1	0.16	0.14
LDA-Tec2	0.13	0.16

Dataset

D4-MP3

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1- Exact Matching
1gram-NN-freq01	0.07	0.05	0.75	0.51	0.12	0.09
1gram-NN-freq03	0.17	0.13	0.59	0.40	0.27	0.20
1gram-NN-freq05	0.25	0.18	0.51	0.37	0.33	0.25
1gram-NN-freq10	0.35	0.27	0.39	0.29	0.37	0.28
2gram-ALGUM-NN-freq01	0.02	0.02	0.29	0.20	0.04	0.03
2gram-ALGUM-NN-freq03	0.25	0.17	0.18	0.12	0.21	0.14
2gram-ALGUM-NN-freq05	0.42	0.27	0.10	0.07	0.16	0.11
2gram-ALGUM-NN-freq10	0.88	0.78	0.06	0.06	0.11	0.10
2gram-TODOS-NN-freq01	0.04	0.03	0.27	0.17	0.08	0.05
2gram-TODOS-NN-freq03	0.33	0.20	0.13	0.09	0.19	0.12
2gram-TODOS-NN-freq05	0.49	0.33	0.06	0.05	0.11	0.08
2gram-TODOS-NN-freq10	0.82	0.67	0.04	0.03	0.07	0.06
comb-1gram-2gram-ALGUM-NN-freq01	0.04	0.03	0.84	0.71	0.07	0.06
comb-1gram-2gram-ALGUM-NN-freq03	0.16	0.14	0.67	0.52	0.26	0.22
comb-1gram-2gram-ALGUM-NN-freq05	0.24	0.20	0.58	0.44	0.34	0.27
comb-1gram-2gram-ALGUM-NN-freq10	0.36	0.30	0.43	0.35	0.39	0.32
comb-1gram-2gram-TODOS-NN-freq01	0.05	0.04	0.84	0.67	0.09	0.08
comb-1gram-2gram-TODOS-NN-freq03	0.17	0.14	0.66	0.49	0.27	0.22
comb-1gram-2gram-TODOS-NN-freq05	0.25	0.19	0.55	0.41	0.34	0.26
comb-1gram-2gram-TODOS-NN-freq10	0.36	0.28	0.41	0.33	0.38	0.30
Comb-1gram-2gram-all-NN-JJ	0.15	0.12	0.63	0.48	0.24	0.20
ODDS-1gram-freq10-2gram-avg-1--3dp	0.05	0.04	0.61	0.49	0.08	0.07
ODDS-1gram-freq10-2gram-avg-2--2dp	0.05	0.04	0.61	0.49	0.08	0.07
ODDS-1gram-freq10-2gram-avg-3--1dp	0.05	0.04	0.61	0.49	0.08	0.07
ODDS-1gram-freq10-2gram-avg-4-0dp	0.10	0.07	0.51	0.35	0.16	0.12
ODDS-1gram-freq10-2gram-avg-5-1dp	0.12	0.09	0.48	0.33	0.19	0.14
ODDS-1gram-freq10-2gram-avg-6-2dp	0.33	0.26	0.40	0.32	0.37	0.29
ODDS-1gram-freq10-2gram-avg-7-3dp	0.33	0.26	0.40	0.31	0.36	0.28
CHI2-1gram-freq05-2gram-avg-1--3dp	0.05	0.04	0.69	0.56	0.09	0.08
CHI2-1gram-freq05-2gram-avg-2--2dp	0.05	0.04	0.69	0.56	0.09	0.08
CHI2-1gram-freq05-2gram-avg-3--1dp	0.05	0.04	0.69	0.56	0.09	0.08
CHI2-1gram-freq05-2gram-avg-4-0dp	0.10	0.08	0.63	0.44	0.17	0.13
CHI2-1gram-freq05-2gram-avg-5-1dp	0.15	0.11	0.56	0.40	0.23	0.18
CHI2-1gram-freq05-2gram-avg-6-2dp	0.15	0.11	0.55	0.37	0.23	0.17
CHI2-1gram-freq05-2gram-avg-7-3dp	0.25	0.18	0.51	0.37	0.33	0.25
PMI-1gram-freq10-2gram-avg-1--3dp	0.05	0.04	0.61	0.49	0.08	0.07
PMI-1gram-freq10-2gram-avg-2--2dp	0.05	0.04	0.61	0.49	0.09	0.07
PMI-1gram-freq10-2gram-avg-3--1dp	0.05	0.04	0.60	0.46	0.10	0.08
PMI-1gram-freq10-2gram-avg-4-0dp	0.07	0.05	0.55	0.33	0.12	0.08
PMI-1gram-freq10-2gram-avg-5-1dp	0.12	0.09	0.45	0.30	0.19	0.14
PMI-1gram-freq10-2gram-avg-6-2dp	0.35	0.27	0.39	0.29	0.37	0.28
PMI-1gram-freq10-2gram-avg-7-3dp	0.35	0.27	0.39	0.29	0.37	0.28
LDA-Tec1	0.10	0.07	0.44	0.29	0.16	0.11
LDA-Tec2	0.07	0.05	0.53	0.37	0.12	0.09

Dataset: D4-MP3

MELHORES RESULTADOS

método	F1 Aproximate Matching	F1 Exact Matching
1gram-NN-freq	0.37	0.28
2gram-some-NN-freq	0.21	0.14
2gram-all-NN-freq	0.19	0.12
Comb-1gram-2gram-some-NN	0.39	0.32
Comb-1gram-2gram-all-NN	0.38	0.30
Comb-1gram-2gram-all-NN-JJ	0.24	0.20
ODDS	0.37	0.29
CHI2	0.33	0.25
PMI	0.37	0.28
LDA-Tec1	0.16	0.11
LDA-Tec2	0.12	0.09

Dataset

D5-DVD

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1-Exact Matching
1gram-NN-freq01	0.07	0.06	0.64	0.48	0.13	0.10
1gram-NN-freq03	0.21	0.18	0.46	0.39	0.29	0.25
1gram-NN-freq05	0.34	0.29	0.39	0.33	0.36	0.31
1gram-NN-freq10	0.43	0.36	0.21	0.18	0.28	0.24
2gram-ALGUM-NN-freq01	0.05	0.02	0.36	0.16	0.08	0.04
2gram-ALGUM-NN-freq03	0.42	0.23	0.10	0.06	0.16	0.10
2gram-ALGUM-NN-freq05	0.83	0.57	0.06	0.05	0.11	0.09
2gram-ALGUM-NN-freq10	1.00	0.67	0.03	0.02	0.06	0.05
2gram-TODOS-NN-freq01	0.08	0.04	0.31	0.13	0.13	0.06
2gram-TODOS-NN-freq03	0.43	0.20	0.07	0.04	0.12	0.06
2gram-TODOS-NN-freq05	0.75	0.40	0.04	0.02	0.07	0.05
2gram-TODOS-NN-freq10	1.00	0.67	0.03	0.02	0.06	0.05
comb-1gram-2gram-ALGUM-NN-freq01	0.05	0.04	0.87	0.63	0.10	0.08
comb-1gram-2gram-ALGUM-NN-freq03	0.22	0.19	0.55	0.45	0.32	0.27
comb-1gram-2gram-ALGUM-NN-freq05	0.36	0.31	0.44	0.38	0.39	0.34
comb-1gram-2gram-ALGUM-NN-freq10	0.44	0.38	0.24	0.21	0.31	0.27
comb-1gram-2gram-TODOS-NN-freq01	0.06	0.05	0.85	0.61	0.12	0.09
comb-1gram-2gram-TODOS-NN-freq03	0.22	0.19	0.53	0.43	0.31	0.26
comb-1gram-2gram-TODOS-NN-freq05	0.35	0.30	0.42	0.35	0.38	0.32
comb-1gram-2gram-TODOS-NN-freq10	0.44	0.38	0.24	0.21	0.31	0.27
Comb-1gram-2gram-all-NN-JJ	0.21	0.18	0.53	0.44	0.30	0.26
ODDS-1gram-freq05-2gram-avg-1--3dp	0.08	0.06	0.71	0.49	0.15	0.11
ODDS-1gram-freq05-2gram-avg-2--2dp	0.08	0.06	0.71	0.49	0.15	0.11
ODDS-1gram-freq05-2gram-avg-3--1dp	0.08	0.06	0.71	0.48	0.15	0.11
ODDS-1gram-freq05-2gram-avg-4-0dp	0.12	0.09	0.56	0.39	0.20	0.15
ODDS-1gram-freq05-2gram-avg-5-1dp	0.14	0.11	0.52	0.39	0.22	0.17
ODDS-1gram-freq05-2gram-avg-6-2dp	0.35	0.30	0.43	0.37	0.38	0.33
ODDS-1gram-freq05-2gram-avg-7-3dp	0.34	0.29	0.41	0.34	0.37	0.31
CHI2-1gram-freq05-2gram-avg-1--3dp	0.08	0.06	0.71	0.49	0.15	0.11
CHI2-1gram-freq05-2gram-avg-2--2dp	0.08	0.06	0.71	0.49	0.15	0.11
CHI2-1gram-freq05-2gram-avg-3--1dp	0.08	0.06	0.71	0.49	0.15	0.11
CHI2-1gram-freq05-2gram-avg-4-0dp	0.14	0.12	0.53	0.40	0.23	0.18
CHI2-1gram-freq05-2gram-avg-5-1dp	0.17	0.14	0.45	0.34	0.25	0.20
CHI2-1gram-freq05-2gram-avg-6-2dp	0.34	0.29	0.39	0.33	0.36	0.31
CHI2-1gram-freq05-2gram-avg-7-3dp	0.34	0.29	0.39	0.33	0.36	0.31
PMI-1gram-freq05-2gram-avg-1--3dp	0.08	0.06	0.71	0.49	0.14	0.11
PMI-1gram-freq05-2gram-avg-2--2dp	0.08	0.06	0.71	0.48	0.15	0.11
PMI-1gram-freq05-2gram-avg-3--1dp	0.08	0.06	0.67	0.45	0.15	0.11
PMI-1gram-freq05-2gram-avg-4-0dp	0.11	0.07	0.60	0.37	0.18	0.12
PMI-1gram-freq05-2gram-avg-5-1dp	0.17	0.14	0.43	0.33	0.24	0.19
PMI-1gram-freq05-2gram-avg-6-2dp	0.34	0.29	0.39	0.33	0.36	0.31
PMI-1gram-freq05-2gram-avg-7-3dp	0.34	0.29	0.39	0.33	0.36	0.31
LDA-Tec1	0.09	0.07	0.41	0.28	0.15	0.11
LDA-Tec2	0.08	0.06	0.58	0.35	0.14	0.10

Dataset: D5-DVD

MELHORES RESULTADOS

método	F1 Approximate Matching	F1 Exact Matching
1gram-NN-freq	0.36	0.31
2gram-some-NN-freq	0.16	0.10
2gram-all-NN-freq	0.13	0.06
Comb-1gram-2gram-some-NN	0.39	0.34
Comb-1gram-2gram-all-NN	0.38	0.32
Comb-1gram-2gram-all-NN-JJ	0.30	0.26
ODDS	0.38	0.33
CHI2	0.36	0.31
PMI	0.36	0.31
LDA-Tec1	0.15	0.11
LDA-Tec2	0.14	0.10

Dataset

D6-laptop2014

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1-Exact Matching
1gram-NN-freq01	0.24	0.16	0.52	0.32	0.33	0.21
1gram-NN-freq03	0.43	0.29	0.25	0.17	0.32	0.22
1gram-NN-freq05	0.51	0.33	0.17	0.12	0.26	0.18
1gram-NN-freq10	0.57	0.38	0.09	0.07	0.15	0.11
2gram-ALGUM-NN-freq01	0.21	0.15	0.51	0.34	0.30	0.21
2gram-ALGUM-NN-freq03	0.69	0.40	0.06	0.04	0.11	0.08
2gram-ALGUM-NN-freq05	0.73	0.50	0.02	0.02	0.04	0.03
2gram-ALGUM-NN-freq10	1.00	0.80	0.00	0.00	0.01	0.01
2gram-TODOS-NN-freq01	0.37	0.23	0.45	0.28	0.41	0.25
2gram-TODOS-NN-freq03	0.83	0.54	0.05	0.03	0.09	0.07
2gram-TODOS-NN-freq05	0.83	0.62	0.02	0.01	0.03	0.03
2gram-TODOS-NN-freq10	1.00	1.00	0.00	0.00	0.01	0.01
comb-1gram-2gram-ALGUM-NN-freq01	0.19	0.16	0.88	0.66	0.31	0.25
comb-1gram-2gram-ALGUM-NN-freq03	0.45	0.31	0.31	0.22	0.37	0.25
comb-1gram-2gram-ALGUM-NN-freq05	0.52	0.35	0.19	0.14	0.28	0.20
comb-1gram-2gram-ALGUM-NN-freq10	0.58	0.39	0.09	0.07	0.16	0.12
comb-1gram-2gram-TODOS-NN-freq01	0.24	0.19	0.85	0.60	0.38	0.28
comb-1gram-2gram-TODOS-NN-freq03	0.45	0.31	0.29	0.21	0.36	0.25
comb-1gram-2gram-TODOS-NN-freq05	0.52	0.35	0.19	0.14	0.27	0.20
comb-1gram-2gram-TODOS-NN-freq10	0.58	0.39	0.09	0.07	0.16	0.12
Comb-1gram-2gram-all-NN-JJ	0.42	0.30	0.27	0.20	0.33	0.24
ODDS-1gram-freq05-2gram-avg-1--3dp	0.23	0.18	0.65	0.46	0.34	0.25
ODDS-1gram-freq05-2gram-avg-2--2dp	0.23	0.18	0.65	0.46	0.34	0.25
ODDS-1gram-freq05-2gram-avg-3--1dp	0.23	0.18	0.65	0.46	0.34	0.25
ODDS-1gram-freq05-2gram-avg-4-0dp	0.34	0.20	0.38	0.22	0.36	0.21
ODDS-1gram-freq05-2gram-avg-5-1dp	0.35	0.21	0.32	0.20	0.33	0.20
ODDS-1gram-freq05-2gram-avg-6-2dp	0.49	0.33	0.18	0.13	0.27	0.19
ODDS-1gram-freq05-2gram-avg-7-3dp	0.49	0.32	0.18	0.13	0.27	0.19
CHI2-1gram-freq05-2gram-avg-1--3dp	0.23	0.18	0.65	0.46	0.34	0.25
CHI2-1gram-freq05-2gram-avg-2--2dp	0.23	0.18	0.65	0.46	0.34	0.25
CHI2-1gram-freq05-2gram-avg-3--1dp	0.23	0.18	0.65	0.46	0.34	0.25
CHI2-1gram-freq05-2gram-avg-4-0dp	0.35	0.21	0.36	0.22	0.36	0.21
CHI2-1gram-freq05-2gram-avg-5-1dp	0.38	0.23	0.25	0.16	0.30	0.19
CHI2-1gram-freq05-2gram-avg-6-2dp	0.38	0.23	0.24	0.15	0.29	0.18
CHI2-1gram-freq05-2gram-avg-7-3dp	0.51	0.33	0.17	0.12	0.26	0.18
PMI-1gram-freq05-2gram-avg-1--3dp	0.23	0.18	0.65	0.46	0.34	0.25
PMI-1gram-freq05-2gram-avg-2--2dp	0.24	0.18	0.64	0.45	0.35	0.26
PMI-1gram-freq05-2gram-avg-3--1dp	0.26	0.19	0.63	0.42	0.37	0.26
PMI-1gram-freq05-2gram-avg-4-0dp	0.31	0.18	0.52	0.28	0.39	0.22
PMI-1gram-freq05-2gram-avg-5-1dp	0.35	0.21	0.30	0.18	0.32	0.19
PMI-1gram-freq05-2gram-avg-6-2dp	0.51	0.33	0.17	0.12	0.26	0.18
PMI-1gram-freq05-2gram-avg-7-3dp	0.51	0.33	0.17	0.12	0.26	0.18
LDA-Tec1	0.37	0.23	0.20	0.13	0.26	0.17
LDA-Tec2	0.35	0.23	0.41	0.26	0.38	0.24

Dataset: D6-laptop2014

MELHORES RESULTADOS

método	F1 Approximate Matching	F1 Exact Matching
1gram-NN-freq	0.33	0.22
2gram-some-NN-freq	0.30	0.21
2gram-all-NN-freq	0.41	0.25
Comb-1gram-2gram-some-NN	0.37	0.25
Comb-1gram-2gram-all-NN	0.38	0.28
Comb-1gram-2gram-all-NN-JJ	0.33	0.24
ODDS	0.36	0.25
CHI2	0.36	0.25
PMI	0.39	0.26
LDA-Tec1	0.26	0.17
LDA-Tec2	0.38	0.24

Dataset

D7-restaurant2014

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1-Exact Matching
1gram-NN-freq01	0.28	0.19	0.58	0.36	0.38	0.25
1gram-NN-freq03	0.42	0.30	0.31	0.23	0.36	0.26
1gram-NN-freq05	0.50	0.39	0.23	0.18	0.31	0.25
1gram-NN-freq10	0.61	0.47	0.16	0.13	0.25	0.21
2gram-ALGUM-NN-freq01	0.21	0.15	0.48	0.33	0.29	0.21
2gram-ALGUM-NN-freq03	0.59	0.31	0.12	0.07	0.20	0.12
2gram-ALGUM-NN-freq05	0.69	0.31	0.04	0.02	0.07	0.04
2gram-ALGUM-NN-freq10	0.75	0.30	0.01	0.01	0.02	0.01
2gram-TODOS-NN-freq01	0.38	0.24	0.40	0.25	0.39	0.25
2gram-TODOS-NN-freq03	0.75	0.50	0.07	0.06	0.14	0.10
2gram-TODOS-NN-freq05	0.65	0.49	0.02	0.02	0.04	0.03
2gram-TODOS-NN-freq10	0.86	0.56	0.01	0.00	0.01	0.01
comb-1gram-2gram-ALGUM-NN-freq01	0.21	0.17	0.89	0.69	0.33	0.28
comb-1gram-2gram-ALGUM-NN-freq03	0.43	0.30	0.43	0.30	0.43	0.30
comb-1gram-2gram-ALGUM-NN-freq05	0.51	0.38	0.27	0.21	0.35	0.27
comb-1gram-2gram-ALGUM-NN-freq10	0.61	0.46	0.17	0.14	0.26	0.21
comb-1gram-2gram-TODOS-NN-freq01	0.27	0.21	0.85	0.62	0.41	0.31
comb-1gram-2gram-TODOS-NN-freq03	0.43	0.32	0.38	0.28	0.40	0.30
comb-1gram-2gram-TODOS-NN-freq05	0.50	0.39	0.24	0.20	0.33	0.26
comb-1gram-2gram-TODOS-NN-freq10	0.61	0.48	0.16	0.13	0.25	0.21
Comb-1gram-2gram-all-NN-JJ	0.45	0.35	0.36	0.28	0.40	0.31
ODDS-1gram-freq05-2gram-avg-1--3dp	0.25	0.19	0.68	0.51	0.36	0.28
ODDS-1gram-freq05-2gram-avg-2--2dp	0.25	0.19	0.68	0.51	0.36	0.28
ODDS-1gram-freq05-2gram-avg-3--1dp	0.25	0.19	0.68	0.51	0.36	0.28
ODDS-1gram-freq05-2gram-avg-4-0dp	0.38	0.25	0.37	0.25	0.37	0.25
ODDS-1gram-freq05-2gram-avg-5-1dp	0.40	0.27	0.31	0.22	0.35	0.24
ODDS-1gram-freq05-2gram-avg-6-2dp	0.50	0.39	0.23	0.19	0.32	0.25
ODDS-1gram-freq05-2gram-avg-7-3dp	0.50	0.39	0.23	0.19	0.32	0.25
CHI2-1gram-freq05-2gram-avg-1--3dp	0.25	0.19	0.68	0.51	0.36	0.28
CHI2-1gram-freq05-2gram-avg-2--2dp	0.25	0.19	0.68	0.51	0.36	0.28
CHI2-1gram-freq05-2gram-avg-3--1dp	0.25	0.19	0.68	0.51	0.36	0.28
CHI2-1gram-freq05-2gram-avg-4-0dp	0.36	0.24	0.40	0.27	0.38	0.25
CHI2-1gram-freq05-2gram-avg-5-1dp	0.40	0.27	0.31	0.22	0.35	0.24
CHI2-1gram-freq05-2gram-avg-6-2dp	0.40	0.27	0.31	0.22	0.35	0.24
CHI2-1gram-freq05-2gram-avg-7-3dp	0.50	0.39	0.23	0.18	0.31	0.25
PMI-1gram-freq05-2gram-avg-1--3dp	0.25	0.19	0.68	0.51	0.36	0.28
PMI-1gram-freq05-2gram-avg-2--2dp	0.25	0.20	0.68	0.50	0.37	0.28
PMI-1gram-freq05-2gram-avg-3--1dp	0.28	0.21	0.67	0.46	0.39	0.29
PMI-1gram-freq05-2gram-avg-4-0dp	0.35	0.23	0.60	0.38	0.44	0.29
PMI-1gram-freq05-2gram-avg-5-1dp	0.38	0.25	0.36	0.24	0.37	0.24
PMI-1gram-freq05-2gram-avg-6-2dp	0.50	0.39	0.23	0.18	0.31	0.25
PMI-1gram-freq05-2gram-avg-7-3dp	0.50	0.39	0.23	0.18	0.31	0.25
LDA-Tec1	0.48	0.35	0.20	0.15	0.28	0.21
LDA-Tec2	0.45	0.30	0.41	0.27	0.43	0.28

Dataset: D7-restaurant2014

MELHORES RESULTADOS

método	F1 Approximate Matching	F1 Exact Matching
1gram-NN-freq	0.38	0.26
2gram-some-NN-freq	0.29	0.21
2gram-all-NN-freq	0.39	0.25
Comb-1gram-2gram-some-NN	0.43	0.30
Comb-1gram-2gram-all-NN	0.41	0.31
Comb-1gram-2gram-all-NN-JJ	0.40	0.31
ODDS	0.37	0.28
CHI2	0.38	0.28
PMI	0.44	0.29
LDA-Tec1	0.28	0.21
LDA-Tec2	0.43	0.28

Dataset

D2-CAM2

método	Precision- Approximate Matching	Precision Exact Matching	RECALL- Approximate Matching	RECALL Exact Matching	F1- Approximate Matching	F1-Exact Matching
1gram-NN-freq01	0.06	0.04	0.76	0.43	0.11	0.08
1gram-NN-freq03	0.18	0.12	0.43	0.28	0.25	0.17
1gram-NN-freq05	0.29	0.18	0.30	0.19	0.30	0.19
1gram-NN-freq10	0.47	0.33	0.20	0.15	0.28	0.21
2gram-ALGUM-NN-freq01	0.05	0.04	0.44	0.32	0.09	0.07
2gram-ALGUM-NN-freq03	0.38	0.24	0.19	0.13	0.25	0.17
2gram-ALGUM-NN-freq05	0.50	0.29	0.13	0.09	0.21	0.13
2gram-ALGUM-NN-freq10	0.00	0.00	0.00	0.00	0.00	0.00
2gram-TODOS-NN-freq01	0.09	0.07	0.41	0.28	0.15	0.11
2gram-TODOS-NN-freq03	0.62	0.42	0.14	0.11	0.23	0.17
2gram-TODOS-NN-freq05	0.69	0.38	0.10	0.06	0.17	0.11
2gram-TODOS-NN-freq10	0.00	0.00	0.00	0.00	0.00	0.00
comb-1gram-2gram-ALGUM-NN-freq01	0.04	0.04	0.93	0.74	0.09	0.08
comb-1gram-2gram-ALGUM-NN-freq03	0.18	0.14	0.55	0.40	0.28	0.21
comb-1gram-2gram-ALGUM-NN-freq05	0.29	0.21	0.40	0.28	0.33	0.24
comb-1gram-2gram-ALGUM-NN-freq10	0.43	0.30	0.20	0.15	0.27	0.20
comb-1gram-2gram-TODOS-NN-freq01	0.06	0.05	0.93	0.70	0.11	0.09
comb-1gram-2gram-TODOS-NN-freq03	0.19	0.15	0.51	0.38	0.28	0.22
comb-1gram-2gram-TODOS-NN-freq05	0.30	0.21	0.37	0.26	0.33	0.23
comb-1gram-2gram-TODOS-NN-freq10	0.47	0.33	0.20	0.15	0.28	0.21
comb-1gram-freq05+2gram-todos-NN- freq03+1gram infreq com adj	0.15	0.11	0.51	0.36	0.23	0.17
ODDS-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-3--1dp	0.06	0.05	0.65	0.51	0.11	0.10
ODDS-1gram-freq05-2gram-avg-4-0dp	0.09	0.07	0.44	0.30	0.15	0.11
ODDS-1gram-freq05-2gram-avg-5-1dp	0.27	0.20	0.36	0.26	0.31	0.22
ODDS-1gram-freq05-2gram-avg-6-2dp	0.27	0.20	0.36	0.26	0.31	0.22
ODDS-1gram-freq05-2gram-avg-7-3dp	0.29	0.21	0.33	0.23	0.31	0.22
CHI2-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-3--1dp	0.06	0.05	0.65	0.51	0.11	0.10
CHI2-1gram-freq05-2gram-avg-4-0dp	0.10	0.07	0.44	0.30	0.16	0.11
CHI2-1gram-freq05-2gram-avg-5-1dp	0.13	0.09	0.41	0.26	0.20	0.13
CHI2-1gram-freq05-2gram-avg-6-2dp	0.29	0.18	0.30	0.19	0.30	0.19
CHI2-1gram-freq05-2gram-avg-7-3dp	0.29	0.18	0.30	0.19	0.30	0.19
PMI-1gram-freq05-2gram-avg-1--3dp	0.06	0.05	0.65	0.51	0.11	0.10
PMI-1gram-freq05-2gram-avg-2--2dp	0.06	0.05	0.65	0.51	0.12	0.10
PMI-1gram-freq05-2gram-avg-3--1dp	0.07	0.05	0.64	0.43	0.12	0.09
PMI-1gram-freq05-2gram-avg-4-0dp	0.08	0.05	0.54	0.28	0.15	0.09
PMI-1gram-freq05-2gram-avg-5-1dp	0.13	0.08	0.39	0.21	0.20	0.11
PMI-1gram-freq05-2gram-avg-6-2dp	0.29	0.18	0.30	0.19	0.30	0.19
PMI-1gram-freq05-2gram-avg-7-3dp	0.29	0.18	0.30	0.19	0.30	0.19
LDA-Tec1	0.11	0.07	0.50	0.30	0.17	0.11
LDA-Tec2	0.09	0.07	0.68	0.43	0.16	0.11

Dataset: D2-CAM2

MELHORES RESULTADOS

método	F1 Approximate Matching	F1 Exact Matching
1gram-NN-freq	0.30	0.21
2gram-some-NN-freq	0.25	0.17
2gram-all-NN-freq	0.23	0.17
Comb-1gram-2gram-some-NN	0.33	0.24
Comb-1gram-2gram-all-NN	0.33	0.23
Comb-1gram-2gram-all-NN-JJ	0.23	0.17
ODDS	0.31	0.22
CHI2	0.30	0.19
PMI	0.30	0.19
LDA-Tec1	0.17	0.11
LDA-Tec2	0.16	0.11

RESULTADOS – MINERAÇÃO DE OPINIÕES BASEADA EM ASPECTOS USANDO APRENDIZADO SEMISSUPERVISIONADO

Este apêndice contém os resultados obtidos na avaliação experimental que foi apresentada no Capítulo 3 para a coleção de revisões de Laptop para a tarefa de classificação de sentimentos dos aspectos. Na sequência são apresentados os resultados para as medidas de Acurácia, Erro, Desvio Padrão, Precisão-Micro, Cobertura-Micro, Precisão-Macro, Cobertura-Macro, Micro- F^1 e Macro- F^1 para cada algoritmo avaliado com as respectivas configurações de variações de parâmetros. Para cada algoritmo e configuração também são apresentados os resultados considerados para a quantidade de exemplos rotulados utilizados para treinar o modelo. Os resultados da coleção de revisões de restaurantes estão disponíveis em: <http://sites.labic.icmc.usp.br/ipm/tese/cap3>

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
GNetMine	1	$\alpha = 0.1$	48,21	51,79	3,99	0,48	0,48	0,44	0,40	0,48	0,42
		$\alpha = 0.3$	47,97	52,03	3,83	0,48	0,48	0,42	0,41	0,48	0,42
		$\alpha = 0.5$	47,57	52,43	3,89	0,48	0,48	0,42	0,42	0,48	0,42
		$\alpha = 0.7$	47,29	52,71	3,95	0,47	0,47	0,42	0,42	0,47	0,42
		$\alpha = 0.9$	47,11	52,89	3,90	0,47	0,47	0,42	0,42	0,47	0,42
	10	$\alpha = 0.1$	53,62	46,38	4,41	0,54	0,54	0,45	0,44	0,54	0,44
		$\alpha = 0.3$	57,89	42,11	3,21	0,58	0,58	0,55	0,48	0,58	0,52
		$\alpha = 0.5$	58,56	41,44	3,04	0,59	0,59	0,56	0,49	0,59	0,53
		$\alpha = 0.7$	58,79	41,21	2,88	0,59	0,59	0,56	0,50	0,59	0,53
		$\alpha = 0.9$	58,93	41,07	2,84	0,59	0,59	0,56	0,50	0,59	0,53
	20	$\alpha = 0.1$	55,08	44,92	3,65	0,55	0,55	0,48	0,45	0,55	0,47
		$\alpha = 0.3$	59,87	40,13	2,73	0,60	0,60	0,57	0,50	0,60	0,53
		$\alpha = 0.5$	60,88	39,12	2,63	0,61	0,61	0,61	0,51	0,61	0,56
		$\alpha = 0.7$	61,29	38,71	2,47	0,61	0,61	0,61	0,52	0,61	0,56
		$\alpha = 0.9$	61,62	38,38	2,39	0,62	0,62	0,61	0,52	0,62	0,56
	30	$\alpha = 0.1$	55,80	44,20	3,45	0,56	0,56	0,49	0,46	0,56	0,47
		$\alpha = 0.3$	60,68	39,32	2,53	0,61	0,61	0,56	0,50	0,61	0,53
		$\alpha = 0.5$	61,61	38,39	2,62	0,62	0,62	0,61	0,51	0,62	0,56
		$\alpha = 0.7$	62,07	37,93	2,51	0,62	0,62	0,63	0,52	0,62	0,57
		$\alpha = 0.9$	62,32	37,68	2,47	0,62	0,62	0,63	0,53	0,62	0,57
	40	$\alpha = 0.1$	56,00	44,00	3,37	0,56	0,56	0,50	0,46	0,56	0,48
		$\alpha = 0.3$	60,98	39,02	2,67	0,61	0,61	0,58	0,50	0,61	0,54
		$\alpha = 0.5$	61,74	38,26	2,45	0,62	0,62	0,61	0,51	0,62	0,56
		$\alpha = 0.7$	62,32	37,68	2,48	0,62	0,62	0,63	0,52	0,62	0,57
		$\alpha = 0.9$	62,66	37,34	2,57	0,63	0,63	0,63	0,53	0,63	0,57
	50	$\alpha = 0.1$	56,23	43,77	3,46	0,56	0,56	0,52	0,46	0,56	0,49
		$\alpha = 0.3$	61,17	38,83	2,64	0,61	0,61	0,59	0,51	0,61	0,54
		$\alpha = 0.5$	61,96	38,04	2,70	0,62	0,62	0,60	0,52	0,62	0,56
		$\alpha = 0.7$	62,50	37,50	2,61	0,63	0,63	0,62	0,52	0,63	0,57
		$\alpha = 0.9$	62,86	37,14	2,57	0,63	0,63	0,63	0,53	0,63	0,58
	60	$\alpha = 0.1$	56,51	43,49	3,30	0,57	0,57	0,54	0,46	0,57	0,50
		$\alpha = 0.3$	61,37	38,63	2,68	0,61	0,61	0,59	0,51	0,61	0,55
		$\alpha = 0.5$	62,12	37,88	2,57	0,62	0,62	0,61	0,52	0,62	0,56
		$\alpha = 0.7$	62,66	37,34	2,65	0,63	0,63	0,63	0,52	0,63	0,57
		$\alpha = 0.9$	63,07	36,93	2,58	0,63	0,63	0,65	0,53	0,63	0,58
	70	$\alpha = 0.1$	57,02	42,98	3,18	0,57	0,57	0,55	0,47	0,57	0,50
		$\alpha = 0.3$	61,46	38,54	2,60	0,61	0,61	0,60	0,51	0,61	0,55

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		$\alpha = 0.5$	62,11	37,89	2,43	0,62	0,62	0,62	0,52	0,62	0,56
		$\alpha = 0.7$	62,57	37,43	2,42	0,63	0,63	0,64	0,52	0,63	0,57
		$\alpha = 0.9$	63,04	36,96	2,42	0,63	0,63	0,65	0,53	0,63	0,58
LPBHN	1	nenhum	42,83	57,17	2,69	0,43	0,43	0,26	0,34	0,43	0,29
	10	nenhum	49,66	50,34	3,33	0,50	0,50	0,41	0,40	0,50	0,41
	20	nenhum	54,67	45,33	3,15	0,55	0,55	0,47	0,45	0,55	0,46
	30	nenhum	56,86	43,14	3,01	0,57	0,57	0,52	0,47	0,57	0,49
	40	nenhum	58,03	41,97	2,75	0,58	0,58	0,56	0,48	0,58	0,52
	50	nenhum	59,00	41,00	2,59	0,59	0,59	0,59	0,49	0,59	0,54
	60	nenhum	59,61	40,39	2,38	0,60	0,60	0,61	0,50	0,60	0,55
	70	nenhum	60,28	39,72	2,29	0,60	0,60	0,63	0,51	0,60	0,56
TagBased	1	B= 0.1 y= 0.1	41,68	58,32	3,91	0,42	0,42	0,14	0,33	0,42	0,20
		B= 0.1 y= 1	41,65	58,35	3,82	0,42	0,42	0,14	0,33	0,42	0,20
		B= 0.1 y= 10	41,62	58,38	3,57	0,42	0,42	0,15	0,33	0,42	0,20
		B= 0.1 y= 100	42,58	57,42	3,02	0,43	0,43	0,16	0,34	0,43	0,22
		B= 0.1 y= 1000	42,97	57,03	2,78	0,43	0,43	0,29	0,34	0,43	0,31
		B= 1 y= 0.1	42,37	57,63	3,10	0,42	0,42	0,16	0,33	0,42	0,22
		B= 1 y= 1	42,32	57,68	3,04	0,42	0,42	0,16	0,33	0,42	0,22
		B= 1 y= 10	42,80	57,20	2,71	0,43	0,43	0,21	0,34	0,43	0,26
		B= 1 y= 100	43,42	56,58	2,95	0,43	0,43	0,32	0,34	0,43	0,33
		B= 1 y= 1000	46,78	53,22	3,79	0,47	0,47	0,45	0,38	0,47	0,41
		B= 10 y= 0.1	42,21	57,79	3,66	0,42	0,42	0,26	0,34	0,42	0,29
		B= 10 y= 1	42,30	57,70	3,75	0,42	0,42	0,26	0,34	0,42	0,29
		B= 10 y= 10	42,60	57,40	3,48	0,43	0,43	0,33	0,34	0,43	0,34
		B= 10 y= 100	45,35	54,65	3,67	0,45	0,45	0,45	0,37	0,45	0,40
		B= 10 y= 1000	47,21	52,79	4,17	0,47	0,47	0,42	0,40	0,47	0,41
		B= 100 y= 0.1	42,27	57,73	4,11	0,42	0,42	0,29	0,34	0,42	0,31
		B= 100 y= 1	42,45	57,55	4,07	0,42	0,42	0,30	0,34	0,42	0,32
		B= 100 y= 10	42,98	57,02	3,84	0,43	0,43	0,38	0,35	0,43	0,36
		B= 100 y= 100	45,36	54,64	4,15	0,45	0,45	0,44	0,37	0,45	0,40
		B= 100 y= 1000	46,67	53,33	4,55	0,47	0,47	0,42	0,40	0,47	0,41
		B= 1000 y= 0.1	42,33	57,67	4,50	0,42	0,42	0,29	0,34	0,42	0,32
		B= 1000 y= 1	42,38	57,62	4,50	0,42	0,42	0,31	0,34	0,42	0,32
		B= 1000 y= 10	42,83	57,17	4,36	0,43	0,43	0,37	0,35	0,43	0,36
		B= 1000 y= 100	45,24	54,76	4,10	0,45	0,45	0,44	0,37	0,45	0,40
		B= 1000 y= 1000	46,62	53,38	4,53	0,47	0,47	0,42	0,40	0,47	0,41
TagBased	10	B= 0.1 y= 0.1	42,66	57,34	2,71	0,43	0,43	0,17	0,33	0,43	0,23

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		B= 0.1 y= 1	42,66	57,34	2,71	0,43	0,43	0,17	0,33	0,43	0,23
		B= 0.1 y= 10	42,66	57,34	2,71	0,43	0,43	0,17	0,33	0,43	0,23
		B= 0.1 y= 100	42,70	57,30	2,71	0,43	0,43	0,20	0,33	0,43	0,25
		B= 0.1 y= 1000	43,04	56,96	2,88	0,43	0,43	0,31	0,34	0,43	0,32
		B= 1 y= 0.1	43,15	56,85	2,84	0,43	0,43	0,36	0,34	0,43	0,35
		B= 1 y= 1	43,20	56,80	2,85	0,43	0,43	0,37	0,34	0,43	0,35
		B= 1 y= 10	44,06	55,94	3,15	0,44	0,44	0,42	0,35	0,44	0,38
		B= 1 y= 100	51,79	48,21	3,89	0,52	0,52	0,43	0,42	0,52	0,43
		B= 1 y= 1000	56,05	43,95	3,41	0,56	0,56	0,48	0,46	0,56	0,47
		B= 10 y= 0.1	50,77	49,23	5,44	0,51	0,51	0,49	0,41	0,51	0,45
		B= 10 y= 1	51,27	48,73	5,28	0,51	0,51	0,49	0,41	0,51	0,45
		B= 10 y= 10	55,05	44,95	4,46	0,55	0,55	0,53	0,45	0,55	0,49
		B= 10 y= 100	58,86	41,14	3,08	0,59	0,59	0,56	0,50	0,59	0,53
		B= 10 y= 1000	59,47	40,53	2,97	0,59	0,59	0,56	0,51	0,59	0,54
		B= 100 y= 0.1	53,98	46,02	5,33	0,54	0,54	0,53	0,44	0,54	0,48
		B= 100 y= 1	54,51	45,49	5,08	0,55	0,55	0,55	0,45	0,55	0,49
		B= 100 y= 10	57,13	42,87	3,92	0,57	0,57	0,56	0,48	0,57	0,52
		B= 100 y= 100	59,30	40,70	3,01	0,59	0,59	0,56	0,51	0,59	0,54
		B= 100 y= 1000	58,97	41,03	3,03	0,59	0,59	0,54	0,52	0,59	0,53
		B= 1000 y= 0.1	54,30	45,70	5,27	0,54	0,54	0,55	0,45	0,54	0,49
		B= 1000 y= 1	54,84	45,16	5,02	0,55	0,55	0,56	0,45	0,55	0,50
		B= 1000 y= 10	57,29	42,71	3,88	0,57	0,57	0,56	0,48	0,57	0,52
		B= 1000 y= 100	59,28	40,72	3,05	0,59	0,59	0,56	0,51	0,59	0,53
		B= 1000 y= 1000	59,00	41,00	3,06	0,59	0,59	0,55	0,52	0,59	0,53
TagBased	20	B= 0.1 y= 0.1	42,69	57,31	2,71	0,43	0,43	0,20	0,33	0,43	0,25
		B= 0.1 y= 1	42,69	57,31	2,71	0,43	0,43	0,20	0,33	0,43	0,25
		B= 0.1 y= 10	42,69	57,31	2,71	0,43	0,43	0,20	0,33	0,43	0,25
		B= 0.1 y= 100	42,77	57,23	2,72	0,43	0,43	0,25	0,34	0,43	0,28
		B= 0.1 y= 1000	42,87	57,13	2,81	0,43	0,43	0,30	0,34	0,43	0,32
		B= 1 y= 0.1	44,27	55,73	3,23	0,44	0,44	0,47	0,35	0,44	0,40
		B= 1 y= 1	44,53	55,47	3,28	0,45	0,45	0,48	0,35	0,45	0,41
		B= 1 y= 10	47,36	52,64	3,61	0,47	0,47	0,47	0,38	0,47	0,42
		B= 1 y= 100	56,11	43,89	3,25	0,56	0,56	0,50	0,46	0,56	0,48
		B= 1 y= 1000	58,24	41,76	2,85	0,58	0,58	0,52	0,48	0,58	0,50
		B= 10 y= 0.1	58,23	41,77	3,40	0,58	0,58	0,59	0,48	0,58	0,53
		B= 10 y= 1	58,69	41,31	3,30	0,59	0,59	0,59	0,48	0,59	0,53
		B= 10 y= 10	60,95	39,05	2,79	0,61	0,61	0,61	0,51	0,61	0,56

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		B= 10 y= 100	62,28	37,72	2,86	0,62	0,62	0,60	0,53	0,62	0,56
		B= 10 y= 1000	62,41	37,59	2,89	0,62	0,62	0,59	0,54	0,62	0,57
		B= 100 y= 0.1	60,71	39,29	2,91	0,61	0,61	0,61	0,51	0,61	0,56
		B= 100 y= 1	61,07	38,93	2,82	0,61	0,61	0,61	0,51	0,61	0,56
		B= 100 y= 10	62,09	37,91	2,82	0,62	0,62	0,61	0,53	0,62	0,57
		B= 100 y= 100	62,54	37,46	2,92	0,63	0,63	0,59	0,55	0,63	0,57
		B= 100 y= 1000	62,38	37,62	2,96	0,62	0,62	0,58	0,55	0,62	0,57
		B= 1000 y= 0.1	60,96	39,04	2,90	0,61	0,61	0,61	0,51	0,61	0,56
		B= 1000 y= 1	61,28	38,72	2,82	0,61	0,61	0,62	0,52	0,61	0,56
		B= 1000 y= 10	62,25	37,75	2,68	0,62	0,62	0,61	0,53	0,62	0,57
		B= 1000 y= 100	62,48	37,52	2,87	0,62	0,62	0,59	0,55	0,62	0,57
		B= 1000 y= 1000	62,30	37,70	2,89	0,62	0,62	0,58	0,55	0,62	0,57
TagBased	30	B= 0.1 y= 0.1	42,71	57,29	2,70	0,43	0,43	0,22	0,33	0,43	0,26
		B= 0.1 y= 1	42,71	57,29	2,70	0,43	0,43	0,22	0,33	0,43	0,26
		B= 0.1 y= 10	42,71	57,29	2,70	0,43	0,43	0,22	0,33	0,43	0,26
		B= 0.1 y= 100	42,78	57,22	2,71	0,43	0,43	0,26	0,34	0,43	0,29
		B= 0.1 y= 1000	42,79	57,21	2,74	0,43	0,43	0,27	0,34	0,43	0,30
		B= 1 y= 0.1	46,46	53,54	3,66	0,46	0,46	0,49	0,37	0,46	0,42
		B= 1 y= 1	47,18	52,82	3,79	0,47	0,47	0,49	0,38	0,47	0,43
		B= 1 y= 10	51,53	48,47	3,82	0,52	0,52	0,50	0,42	0,52	0,46
		B= 1 y= 100	58,08	41,92	3,11	0,58	0,58	0,53	0,48	0,58	0,50
		B= 1 y= 1000	59,19	40,81	3,01	0,59	0,59	0,53	0,49	0,59	0,51
		B= 10 y= 0.1	61,63	38,37	2,93	0,62	0,62	0,62	0,51	0,62	0,56
		B= 10 y= 1	61,87	38,13	2,88	0,62	0,62	0,63	0,52	0,62	0,57
		B= 10 y= 10	63,00	37,00	2,87	0,63	0,63	0,63	0,53	0,63	0,58
		B= 10 y= 100	64,07	35,93	2,82	0,64	0,64	0,62	0,55	0,64	0,59
		B= 10 y= 1000	64,09	35,91	2,85	0,64	0,64	0,62	0,56	0,64	0,59
		B= 100 y= 0.1	63,26	36,74	2,88	0,63	0,63	0,63	0,54	0,63	0,58
		B= 100 y= 1	63,41	36,59	2,87	0,63	0,63	0,63	0,54	0,63	0,58
		B= 100 y= 10	64,14	35,86	2,75	0,64	0,64	0,63	0,56	0,64	0,59
		B= 100 y= 100	64,33	35,67	2,72	0,64	0,64	0,61	0,57	0,64	0,59
		B= 100 y= 1000	64,16	35,84	2,74	0,64	0,64	0,61	0,57	0,64	0,59
		B= 1000 y= 0.1	63,47	36,53	2,85	0,63	0,63	0,63	0,54	0,63	0,58
		B= 1000 y= 1	63,56	36,44	2,87	0,64	0,64	0,62	0,55	0,64	0,58
		B= 1000 y= 10	64,14	35,86	2,68	0,64	0,64	0,62	0,56	0,64	0,59
		B= 1000 y= 100	64,26	35,74	2,78	0,64	0,64	0,61	0,57	0,64	0,59
		B= 1000 y= 1000	64,04	35,96	2,78	0,64	0,64	0,61	0,57	0,64	0,59

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
TagBased	40	B= 0.1 y= 0.1	42,70	57,30	2,67	0,43	0,43	0,22	0,33	0,43	0,27
		B= 0.1 y= 1	42,70	57,30	2,67	0,43	0,43	0,22	0,33	0,43	0,27
		B= 0.1 y= 10	42,70	57,30	2,67	0,43	0,43	0,22	0,33	0,43	0,27
		B= 0.1 y= 100	42,75	57,25	2,67	0,43	0,43	0,25	0,34	0,43	0,29
		B= 0.1 y= 1000	42,77	57,23	2,73	0,43	0,43	0,26	0,34	0,43	0,29
		B= 1 y= 0.1	49,71	50,29	3,89	0,50	0,50	0,52	0,40	0,50	0,45
		B= 1 y= 1	50,53	49,47	3,80	0,51	0,51	0,52	0,41	0,51	0,46
		B= 1 y= 10	54,74	45,26	3,37	0,55	0,55	0,53	0,45	0,55	0,49
		B= 1 y= 100	58,93	41,07	2,90	0,59	0,59	0,56	0,48	0,59	0,52
		B= 1 y= 1000	59,82	40,18	2,89	0,60	0,60	0,56	0,49	0,60	0,52
		B= 10 y= 0.1	62,96	37,04	2,72	0,63	0,63	0,64	0,53	0,63	0,58
		B= 10 y= 1	63,15	36,85	2,68	0,63	0,63	0,64	0,53	0,63	0,58
		B= 10 y= 10	64,13	35,87	2,77	0,64	0,64	0,65	0,54	0,64	0,59
		B= 10 y= 100	65,16	34,84	2,71	0,65	0,65	0,64	0,56	0,65	0,60
		B= 10 y= 1000	65,39	34,61	2,78	0,65	0,65	0,64	0,57	0,65	0,60
		B= 100 y= 0.1	64,55	35,45	2,75	0,65	0,65	0,64	0,56	0,65	0,59
		B= 100 y= 1	64,69	35,31	2,79	0,65	0,65	0,64	0,56	0,65	0,60
		B= 100 y= 10	65,48	34,52	2,68	0,65	0,65	0,64	0,57	0,65	0,61
		B= 100 y= 100	65,80	34,20	2,72	0,66	0,66	0,63	0,59	0,66	0,61
		B= 100 y= 1000	65,68	34,32	2,66	0,66	0,66	0,63	0,59	0,66	0,61
		B= 1000 y= 0.1	64,75	35,25	2,73	0,65	0,65	0,64	0,56	0,65	0,60
		B= 1000 y= 1	64,87	35,13	2,70	0,65	0,65	0,64	0,56	0,65	0,60
		B= 1000 y= 10	65,60	34,40	2,56	0,66	0,66	0,64	0,58	0,66	0,61
		B= 1000 y= 100	65,70	34,30	2,86	0,66	0,66	0,63	0,59	0,66	0,61
		B= 1000 y= 1000	65,62	34,38	2,63	0,66	0,66	0,63	0,59	0,66	0,61
TagBased	50	B= 0.1 y= 0.1	42,71	57,29	2,67	0,43	0,43	0,23	0,33	0,43	0,27
		B= 0.1 y= 1	42,71	57,29	2,67	0,43	0,43	0,23	0,33	0,43	0,27
		B= 0.1 y= 10	42,71	57,29	2,67	0,43	0,43	0,23	0,33	0,43	0,27
		B= 0.1 y= 100	42,75	57,25	2,68	0,43	0,43	0,25	0,34	0,43	0,29
		B= 0.1 y= 1000	42,77	57,23	2,69	0,43	0,43	0,26	0,34	0,43	0,29
		B= 1 y= 0.1	52,93	47,07	3,59	0,53	0,53	0,55	0,43	0,53	0,48
		B= 1 y= 1	53,66	46,34	3,60	0,54	0,54	0,55	0,44	0,54	0,49
		B= 1 y= 10	56,93	43,07	3,21	0,57	0,57	0,56	0,47	0,57	0,51
		B= 1 y= 100	59,63	40,37	3,01	0,60	0,60	0,59	0,49	0,60	0,53
		B= 1 y= 1000	60,13	39,87	3,01	0,60	0,60	0,59	0,50	0,60	0,54
		B= 10 y= 0.1	64,10	35,90	2,82	0,64	0,64	0,65	0,54	0,64	0,59
		B= 10 y= 1	64,31	35,69	2,86	0,64	0,64	0,66	0,54	0,64	0,60

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		B= 10 y= 10	65,13	34,87	2,83	0,65	0,65	0,66	0,56	0,65	0,60
		B= 10 y= 100	65,78	34,22	2,59	0,66	0,66	0,65	0,57	0,66	0,61
		B= 10 y= 1000	65,80	34,20	2,61	0,66	0,66	0,65	0,57	0,66	0,61
		B= 100 y= 0.1	65,57	34,43	2,61	0,66	0,66	0,65	0,57	0,66	0,61
		B= 100 y= 1	65,70	34,30	2,65	0,66	0,66	0,65	0,57	0,66	0,61
		B= 100 y= 10	66,12	33,88	2,55	0,66	0,66	0,64	0,58	0,66	0,61
		B= 100 y= 100	66,37	33,63	2,55	0,66	0,66	0,64	0,59	0,66	0,62
		B= 100 y= 1000	66,35	33,65	2,55	0,66	0,66	0,64	0,59	0,66	0,62
		B= 1000 y= 0.1	65,72	34,28	2,51	0,66	0,66	0,64	0,58	0,66	0,61
		B= 1000 y= 1	65,83	34,17	2,61	0,66	0,66	0,64	0,58	0,66	0,61
		B= 1000 y= 10	66,19	33,81	2,52	0,66	0,66	0,64	0,59	0,66	0,61
		B= 1000 y= 100	66,25	33,75	2,58	0,66	0,66	0,63	0,59	0,66	0,61
		B= 1000 y= 1000	66,27	33,73	2,65	0,66	0,66	0,63	0,60	0,66	0,61
TagBased	60	B= 0.1 y= 0.1	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 y= 1	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 y= 10	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 y= 100	42,74	57,26	2,66	0,43	0,43	0,25	0,34	0,43	0,29
		B= 0.1 y= 1000	42,75	57,25	2,66	0,43	0,43	0,26	0,34	0,43	0,29
		B= 1 y= 0.1	55,74	44,26	3,33	0,56	0,56	0,57	0,46	0,56	0,50
		B= 1 y= 1	56,26	43,74	3,29	0,56	0,56	0,57	0,46	0,56	0,51
		B= 1 y= 10	58,42	41,58	3,06	0,58	0,58	0,58	0,48	0,58	0,53
		B= 1 y= 100	60,24	39,76	2,99	0,60	0,60	0,60	0,50	0,60	0,54
		B= 1 y= 1000	60,52	39,48	2,88	0,61	0,61	0,60	0,50	0,61	0,54
		B= 10 y= 0.1	65,04	34,96	2,78	0,65	0,65	0,66	0,55	0,65	0,60
		B= 10 y= 1	65,19	34,81	2,80	0,65	0,65	0,67	0,56	0,65	0,61
		B= 10 y= 10	65,92	34,08	2,77	0,66	0,66	0,67	0,57	0,66	0,61
		B= 10 y= 100	66,32	33,68	2,73	0,66	0,66	0,66	0,58	0,66	0,62
		B= 10 y= 1000	66,51	33,49	2,80	0,67	0,67	0,66	0,58	0,67	0,62
		B= 100 y= 0.1	66,48	33,52	2,66	0,66	0,66	0,65	0,58	0,66	0,62
		B= 100 y= 1	66,58	33,42	2,69	0,67	0,67	0,65	0,59	0,67	0,62
		B= 100 y= 10	66,88	33,12	2,59	0,67	0,67	0,65	0,59	0,67	0,62
		B= 100 y= 100	67,11	32,89	2,68	0,67	0,67	0,65	0,60	0,67	0,62
		B= 100 y= 1000	67,13	32,87	2,69	0,67	0,67	0,64	0,60	0,67	0,62
		B= 1000 y= 0.1	66,54	33,46	2,63	0,67	0,67	0,65	0,59	0,67	0,62
		B= 1000 y= 1	66,71	33,29	2,62	0,67	0,67	0,65	0,59	0,67	0,62
		B= 1000 y= 10	66,91	33,09	2,58	0,67	0,67	0,65	0,60	0,67	0,62
		B= 1000 y= 100	67,18	32,82	2,74	0,67	0,67	0,65	0,61	0,67	0,62

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
TagBased	70	B= 1000 γ = 1000	67,15	32,85	2,68	0,67	0,67	0,64	0,61	0,67	0,62
		B= 0.1 γ = 0.1	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 γ = 1	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 γ = 10	42,71	57,29	2,66	0,43	0,43	0,24	0,33	0,43	0,28
		B= 0.1 γ = 100	42,73	57,27	2,67	0,43	0,43	0,25	0,34	0,43	0,29
		B= 0.1 γ = 1000	42,73	57,27	2,66	0,43	0,43	0,25	0,34	0,43	0,29
		B= 1 γ = 0.1	57,64	42,36	3,07	0,58	0,58	0,58	0,47	0,58	0,52
		B= 1 γ = 1	57,93	42,07	3,05	0,58	0,58	0,58	0,48	0,58	0,52
		B= 1 γ = 10	59,41	40,59	3,02	0,59	0,59	0,60	0,49	0,59	0,54
		B= 1 γ = 100	60,55	39,45	2,82	0,61	0,61	0,61	0,50	0,61	0,55
		B= 1 γ = 1000	60,76	39,24	2,85	0,61	0,61	0,61	0,50	0,61	0,55
		B= 10 γ = 0.1	65,63	34,37	2,58	0,66	0,66	0,66	0,56	0,66	0,61
		B= 10 γ = 1	65,85	34,15	2,54	0,66	0,66	0,66	0,56	0,66	0,61
		B= 10 γ = 10	66,33	33,67	2,53	0,66	0,66	0,67	0,57	0,66	0,62
		B= 10 γ = 100	66,80	33,20	2,56	0,67	0,67	0,67	0,58	0,67	0,62
		B= 10 γ = 1000	66,90	33,10	2,54	0,67	0,67	0,66	0,58	0,67	0,62
		B= 100 γ = 0.1	67,01	32,99	2,53	0,67	0,67	0,66	0,59	0,67	0,62
		B= 100 γ = 1	67,08	32,92	2,49	0,67	0,67	0,66	0,59	0,67	0,62
		B= 100 γ = 10	67,43	32,57	2,51	0,67	0,67	0,66	0,60	0,67	0,63
		B= 100 γ = 100	67,71	32,29	2,52	0,68	0,68	0,66	0,61	0,68	0,63
		B= 100 γ = 1000	67,75	32,25	2,54	0,68	0,68	0,65	0,61	0,68	0,63
		B= 1000 γ = 0.1	67,17	32,83	2,50	0,67	0,67	0,65	0,60	0,67	0,62
		B= 1000 γ = 1	67,19	32,81	2,51	0,67	0,67	0,65	0,60	0,67	0,62
		B= 1000 γ = 10	67,50	32,50	2,56	0,67	0,67	0,65	0,60	0,67	0,63
		B= 1000 γ = 100	67,80	32,20	2,55	0,68	0,68	0,65	0,61	0,68	0,63
		B= 1000 γ = 1000	67,75	32,25	2,56	0,68	0,68	0,65	0,61	0,68	0,63
TCBHN	1	η = 0.01 ε = 0.01	45,49	54,51	4,11	0,45	0,45	0,42	0,41	0,45	0,41
		η = 0.05 ε = 0.01	44,84	55,16	3,75	0,45	0,45	0,41	0,41	0,45	0,41
		η = 0.1 ε = 0.01	44,69	55,31	3,76	0,45	0,45	0,41	0,41	0,45	0,41
		η = 0.5 ε = 0.01	44,38	55,62	3,78	0,44	0,44	0,41	0,41	0,44	0,41
	10	η = 0.01 ε = 0.01	57,22	42,78	3,28	0,57	0,57	0,53	0,52	0,57	0,53
		η = 0.05 ε = 0.01	55,24	44,76	3,45	0,55	0,55	0,51	0,51	0,55	0,51
		η = 0.1 ε = 0.01	55,10	44,90	3,44	0,55	0,55	0,51	0,51	0,55	0,51
		η = 0.5 ε = 0.01	54,54	45,46	3,67	0,55	0,55	0,51	0,50	0,55	0,50
	20	η = 0.01 ε = 0.01	60,58	39,42	2,66	0,61	0,61	0,56	0,56	0,61	0,56
		η = 0.05 ε = 0.01	58,51	41,49	3,01	0,59	0,59	0,55	0,54	0,59	0,55
		η = 0.1 ε = 0.01	58,07	41,93	2,82	0,58	0,58	0,54	0,54	0,58	0,54

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
	30	$\eta = 0.5$ $\varepsilon = 0.01$	56,60	43,40	3,21	0,57	0,57	0,53	0,53	0,57	0,53
		$\eta = 0.01$ $\varepsilon = 0.01$	62,66	37,34	2,26	0,63	0,63	0,59	0,58	0,63	0,59
		$\eta = 0.05$ $\varepsilon = 0.01$	60,31	39,69	2,65	0,60	0,60	0,57	0,57	0,60	0,57
		$\eta = 0.1$ $\varepsilon = 0.01$	59,67	40,33	2,90	0,60	0,60	0,56	0,56	0,60	0,56
	40	$\eta = 0.5$ $\varepsilon = 0.01$	57,68	42,32	3,30	0,58	0,58	0,54	0,54	0,58	0,54
		$\eta = 0.01$ $\varepsilon = 0.01$	64,51	35,49	2,63	0,65	0,65	0,61	0,60	0,65	0,61
		$\eta = 0.05$ $\varepsilon = 0.01$	62,23	37,77	2,59	0,62	0,62	0,59	0,59	0,62	0,59
		$\eta = 0.1$ $\varepsilon = 0.01$	61,27	38,73	2,60	0,61	0,61	0,58	0,58	0,61	0,58
	50	$\eta = 0.5$ $\varepsilon = 0.01$	58,20	41,80	2,74	0,58	0,58	0,55	0,55	0,58	0,55
		$\eta = 0.01$ $\varepsilon = 0.01$	65,67	34,33	2,57	0,66	0,66	0,62	0,62	0,66	0,62
		$\eta = 0.05$ $\varepsilon = 0.01$	63,15	36,85	2,45	0,63	0,63	0,60	0,60	0,63	0,60
		$\eta = 0.1$ $\varepsilon = 0.01$	61,90	38,10	2,68	0,62	0,62	0,59	0,59	0,62	0,59
	60	$\eta = 0.5$ $\varepsilon = 0.01$	58,37	41,63	2,85	0,58	0,58	0,55	0,56	0,58	0,56
		$\eta = 0.01$ $\varepsilon = 0.01$	66,59	33,41	2,55	0,67	0,67	0,63	0,63	0,67	0,63
		$\eta = 0.05$ $\varepsilon = 0.01$	63,89	36,11	2,71	0,64	0,64	0,61	0,61	0,64	0,61
		$\eta = 0.1$ $\varepsilon = 0.01$	62,19	37,81	2,71	0,62	0,62	0,59	0,59	0,62	0,59
	70	$\eta = 0.5$ $\varepsilon = 0.01$	58,20	41,80	2,49	0,58	0,58	0,56	0,56	0,58	0,56
		$\eta = 0.01$ $\varepsilon = 0.01$	67,57	32,43	2,38	0,68	0,68	0,64	0,64	0,68	0,64
		$\eta = 0.05$ $\varepsilon = 0.01$	64,27	35,73	2,85	0,64	0,64	0,61	0,61	0,64	0,61
		$\eta = 0.1$ $\varepsilon = 0.01$	62,63	37,37	2,96	0,63	0,63	0,60	0,60	0,63	0,60
	EM	$\eta = 0.5$ $\varepsilon = 0.01$	58,04	41,96	2,82	0,58	0,58	0,56	0,56	0,58	0,56
		Weight Unl. Docs. = 0.1 # Components = 10	42,55	57,45	3,50	0,43	0,43	0,39	0,34	0,43	0,37
		Weight Unl. Docs. = 0.1 # Components = 1	42,59	57,41	3,53	0,43	0,43	0,39	0,34	0,43	0,36
		Weight Unl. Docs. = 0.1 # Components = 2	41,39	58,61	4,13	0,41	0,41	0,38	0,34	0,41	0,36
		Weight Unl. Docs. = 0.1 # Components = 5	42,16	57,84	3,79	0,42	0,42	0,39	0,34	0,42	0,36
		Weight Unl. Docs. = 0.3 # Components = 10	44,55	55,45	8,24	0,45	0,45	0,42	0,38	0,45	0,40
		Weight Unl. Docs. = 0.3 # Components = 1	52,69	47,31	5,44	0,53	0,53	0,46	0,44	0,53	0,45
		Weight Unl. Docs. = 0.3 # Components = 2	49,76	50,24	6,76	0,50	0,50	0,44	0,41	0,50	0,43
		Weight Unl. Docs. = 0.3 # Components = 5	47,87	52,13	6,52	0,48	0,48	0,43	0,40	0,48	0,41
		Weight Unl. Docs. = 0.5 # Components = 10	43,20	56,80	8,51	0,43	0,43	0,41	0,39	0,43	0,40
		Weight Unl. Docs. = 0.5 # Components = 1	52,16	47,84	4,49	0,52	0,52	0,46	0,46	0,52	0,46
		Weight Unl. Docs. = 0.5 # Components = 2	50,78	49,22	5,11	0,51	0,51	0,44	0,43	0,51	0,44
		Weight Unl. Docs. = 0.5 # Components = 5	48,87	51,13	5,86	0,49	0,49	0,42	0,41	0,49	0,42
		Weight Unl. Docs. = 0.7 # Components = 10	42,84	57,16	7,56	0,43	0,43	0,41	0,39	0,43	0,40
		Weight Unl. Docs. = 0.7 # Components = 1	49,95	50,05	5,12	0,50	0,50	0,45	0,45	0,50	0,45
		Weight Unl. Docs. = 0.7 # Components = 2	50,39	49,61	4,61	0,50	0,50	0,44	0,43	0,50	0,44
		Weight Unl. Docs. = 0.7 # Components = 5	48,27	51,73	5,73	0,48	0,48	0,42	0,41	0,48	0,42

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Weight Unl. Docs. = 0.9 # Components = 10	44,20	55,80	6,29	0,44	0,44	0,41	0,40	0,44	0,41
		Weight Unl. Docs. = 0.9 # Components = 1	48,69	51,31	5,17	0,49	0,49	0,45	0,45	0,49	0,45
		Weight Unl. Docs. = 0.9 # Components = 2	50,00	50,00	4,31	0,50	0,50	0,45	0,44	0,50	0,44
		Weight Unl. Docs. = 0.9 # Components = 5	47,72	52,28	5,75	0,48	0,48	0,42	0,41	0,48	0,42
	10	Weight Unl. Docs. = 0.1 # Components = 10	58,27	41,73	2,58	0,58	0,58	0,54	0,51	0,58	0,52
		Weight Unl. Docs. = 0.1 # Components = 1	58,28	41,72	2,56	0,58	0,58	0,54	0,51	0,58	0,52
		Weight Unl. Docs. = 0.1 # Components = 2	58,33	41,67	2,49	0,58	0,58	0,54	0,51	0,58	0,52
		Weight Unl. Docs. = 0.1 # Components = 5	58,26	41,74	2,55	0,58	0,58	0,54	0,51	0,58	0,52
		Weight Unl. Docs. = 0.3 # Components = 10	58,05	41,95	2,44	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.3 # Components = 1	58,34	41,66	2,35	0,58	0,58	0,54	0,51	0,58	0,52
		Weight Unl. Docs. = 0.3 # Components = 2	58,10	41,90	2,47	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.3 # Components = 5	57,99	42,01	2,38	0,58	0,58	0,53	0,50	0,58	0,52
		Weight Unl. Docs. = 0.5 # Components = 10	56,67	43,33	2,69	0,57	0,57	0,52	0,50	0,57	0,51
		Weight Unl. Docs. = 0.5 # Components = 1	58,10	41,90	2,21	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.5 # Components = 2	57,75	42,25	2,59	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.5 # Components = 5	57,01	42,99	2,69	0,57	0,57	0,53	0,50	0,57	0,51
		Weight Unl. Docs. = 0.7 # Components = 10	55,55	44,45	3,23	0,56	0,56	0,52	0,48	0,56	0,50
		Weight Unl. Docs. = 0.7 # Components = 1	57,63	42,37	2,40	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.7 # Components = 2	56,91	43,09	2,73	0,57	0,57	0,52	0,50	0,57	0,51
		Weight Unl. Docs. = 0.7 # Components = 5	56,16	43,84	3,16	0,56	0,56	0,52	0,49	0,56	0,50
		Weight Unl. Docs. = 0.9 # Components = 10	54,44	45,56	3,30	0,54	0,54	0,50	0,47	0,54	0,49
		Weight Unl. Docs. = 0.9 # Components = 1	57,18	42,82	2,39	0,57	0,57	0,52	0,51	0,57	0,52
		Weight Unl. Docs. = 0.9 # Components = 2	56,81	43,19	2,60	0,57	0,57	0,52	0,50	0,57	0,51
		Weight Unl. Docs. = 0.9 # Components = 5	55,69	44,31	2,95	0,56	0,56	0,52	0,48	0,56	0,50
	20	Weight Unl. Docs. = 0.1 # Components = 10	59,82	40,18	2,07	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.1 # Components = 1	59,84	40,16	2,07	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.1 # Components = 2	59,85	40,15	2,08	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.1 # Components = 5	59,82	40,18	2,06	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.3 # Components = 10	59,54	40,46	2,07	0,60	0,60	0,55	0,53	0,60	0,54
		Weight Unl. Docs. = 0.3 # Components = 1	59,52	40,48	2,01	0,60	0,60	0,55	0,53	0,60	0,54
		Weight Unl. Docs. = 0.3 # Components = 2	59,62	40,38	2,07	0,60	0,60	0,55	0,54	0,60	0,54
		Weight Unl. Docs. = 0.3 # Components = 5	59,54	40,46	2,14	0,60	0,60	0,55	0,53	0,60	0,54
		Weight Unl. Docs. = 0.5 # Components = 10	58,41	41,59	2,33	0,58	0,58	0,54	0,53	0,58	0,53
		Weight Unl. Docs. = 0.5 # Components = 1	59,18	40,82	2,04	0,59	0,59	0,55	0,53	0,59	0,54
		Weight Unl. Docs. = 0.5 # Components = 2	59,01	40,99	2,19	0,59	0,59	0,55	0,53	0,59	0,54
		Weight Unl. Docs. = 0.5 # Components = 5	58,60	41,40	2,60	0,59	0,59	0,54	0,53	0,59	0,53
		Weight Unl. Docs. = 0.7 # Components = 10	57,34	42,66	2,57	0,57	0,57	0,53	0,51	0,57	0,52

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Weight Unl. Docs. = 0.7 # Components = 1	58,85	41,15	2,08	0,59	0,59	0,54	0,53	0,59	0,54
		Weight Unl. Docs. = 0.7 # Components = 2	58,32	41,68	2,35	0,58	0,58	0,54	0,52	0,58	0,53
		Weight Unl. Docs. = 0.7 # Components = 5	57,53	42,47	2,42	0,58	0,58	0,53	0,51	0,58	0,52
		Weight Unl. Docs. = 0.9 # Components = 10	56,47	43,53	2,65	0,56	0,56	0,52	0,50	0,56	0,51
		Weight Unl. Docs. = 0.9 # Components = 1	58,46	41,54	2,12	0,58	0,58	0,54	0,53	0,58	0,54
		Weight Unl. Docs. = 0.9 # Components = 2	57,95	42,05	2,39	0,58	0,58	0,53	0,52	0,58	0,53
	30	Weight Unl. Docs. = 0.9 # Components = 5	57,21	42,79	2,55	0,57	0,57	0,53	0,51	0,57	0,52
		Weight Unl. Docs. = 0.1 # Components = 10	60,51	39,49	1,96	0,61	0,61	0,57	0,55	0,61	0,56
		Weight Unl. Docs. = 0.1 # Components = 1	60,50	39,50	1,97	0,60	0,60	0,57	0,55	0,60	0,56
		Weight Unl. Docs. = 0.1 # Components = 2	60,48	39,52	1,95	0,60	0,60	0,57	0,55	0,60	0,56
		Weight Unl. Docs. = 0.1 # Components = 5	60,49	39,51	1,92	0,60	0,60	0,57	0,55	0,60	0,56
		Weight Unl. Docs. = 0.3 # Components = 10	60,54	39,46	2,07	0,61	0,61	0,56	0,55	0,61	0,56
		Weight Unl. Docs. = 0.3 # Components = 1	60,48	39,52	2,08	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.3 # Components = 2	60,46	39,54	2,06	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.3 # Components = 5	60,45	39,55	2,12	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.5 # Components = 10	59,64	40,36	2,08	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.5 # Components = 1	59,93	40,07	2,04	0,60	0,60	0,56	0,55	0,60	0,55
		Weight Unl. Docs. = 0.5 # Components = 2	59,66	40,34	2,07	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.5 # Components = 5	59,71	40,29	2,16	0,60	0,60	0,56	0,54	0,60	0,55
		Weight Unl. Docs. = 0.7 # Components = 10	58,62	41,38	2,26	0,59	0,59	0,55	0,53	0,59	0,54
		Weight Unl. Docs. = 0.7 # Components = 1	59,51	40,49	1,86	0,60	0,60	0,55	0,55	0,60	0,55
		Weight Unl. Docs. = 0.7 # Components = 2	59,08	40,92	2,26	0,59	0,59	0,55	0,54	0,59	0,54
		Weight Unl. Docs. = 0.7 # Components = 5	58,38	41,62	2,16	0,58	0,58	0,54	0,53	0,58	0,54
		Weight Unl. Docs. = 0.9 # Components = 10	57,81	42,19	2,41	0,58	0,58	0,54	0,52	0,58	0,53
		Weight Unl. Docs. = 0.9 # Components = 1	59,13	40,87	1,77	0,59	0,59	0,55	0,54	0,59	0,55
		Weight Unl. Docs. = 0.9 # Components = 2	58,52	41,48	2,42	0,59	0,59	0,54	0,53	0,59	0,54
		Weight Unl. Docs. = 0.9 # Components = 5	57,97	42,03	2,60	0,58	0,58	0,54	0,52	0,58	0,53
	40	Weight Unl. Docs. = 0.1 # Components = 10	61,21	38,79	2,11	0,61	0,61	0,58	0,56	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 1	61,21	38,79	2,11	0,61	0,61	0,58	0,56	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 2	61,20	38,80	2,10	0,61	0,61	0,58	0,56	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 5	61,21	38,79	2,10	0,61	0,61	0,58	0,56	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 10	60,96	39,04	2,10	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 1	60,93	39,07	2,11	0,61	0,61	0,57	0,56	0,61	0,56
		Weight Unl. Docs. = 0.3 # Components = 2	60,98	39,02	2,08	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 5	61,02	38,98	2,19	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 10	60,47	39,53	2,06	0,60	0,60	0,57	0,55	0,60	0,56
		Weight Unl. Docs. = 0.5 # Components = 1	60,72	39,28	2,06	0,61	0,61	0,57	0,56	0,61	0,56

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Weight Unl. Docs. = 0.5 # Components = 2	60,58	39,42	2,06	0,61	0,61	0,57	0,55	0,61	0,56
		Weight Unl. Docs. = 0.5 # Components = 5	60,50	39,50	2,15	0,60	0,60	0,57	0,55	0,60	0,56
		Weight Unl. Docs. = 0.7 # Components = 10	59,64	40,36	2,23	0,60	0,60	0,56	0,55	0,60	0,55
		Weight Unl. Docs. = 0.7 # Components = 1	60,27	39,73	1,80	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.7 # Components = 2	60,03	39,97	2,31	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.7 # Components = 5	59,79	40,21	2,23	0,60	0,60	0,56	0,55	0,60	0,55
		Weight Unl. Docs. = 0.9 # Components = 10	58,84	41,16	2,18	0,59	0,59	0,55	0,54	0,59	0,54
		Weight Unl. Docs. = 0.9 # Components = 1	60,00	40,00	1,98	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.9 # Components = 2	59,45	40,55	2,23	0,59	0,59	0,55	0,55	0,59	0,55
		Weight Unl. Docs. = 0.9 # Components = 5	59,01	40,99	2,28	0,59	0,59	0,55	0,54	0,59	0,54
	50	Weight Unl. Docs. = 0.1 # Components = 10	61,47	38,53	2,29	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 1	61,44	38,56	2,30	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 2	61,44	38,56	2,29	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.1 # Components = 5	61,44	38,56	2,29	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 10	61,30	38,70	2,17	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 1	61,22	38,78	2,08	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 2	61,34	38,66	2,16	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.3 # Components = 5	61,29	38,71	2,15	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 10	61,23	38,77	2,26	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 1	61,14	38,86	2,02	0,61	0,61	0,57	0,57	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 2	61,05	38,95	2,16	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 5	61,21	38,79	2,28	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 10	60,41	39,59	2,13	0,60	0,60	0,57	0,56	0,60	0,56
		Weight Unl. Docs. = 0.7 # Components = 1	60,86	39,14	1,93	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 2	60,60	39,40	1,97	0,61	0,61	0,57	0,56	0,61	0,56
		Weight Unl. Docs. = 0.7 # Components = 5	60,59	39,41	1,95	0,61	0,61	0,57	0,56	0,61	0,56
		Weight Unl. Docs. = 0.9 # Components = 10	59,92	40,08	2,20	0,60	0,60	0,56	0,55	0,60	0,56
		Weight Unl. Docs. = 0.9 # Components = 1	60,46	39,54	1,99	0,60	0,60	0,57	0,56	0,60	0,56
		Weight Unl. Docs. = 0.9 # Components = 2	60,16	39,84	2,15	0,60	0,60	0,56	0,56	0,60	0,56
		Weight Unl. Docs. = 0.9 # Components = 5	59,97	40,03	2,11	0,60	0,60	0,56	0,55	0,60	0,56
	60	Weight Unl. Docs. = 0.1 # Components = 10	61,70	38,30	2,12	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 1	61,70	38,30	2,10	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 2	61,71	38,29	2,10	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 5	61,70	38,30	2,11	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 10	61,50	38,50	2,05	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.3 # Components = 1	61,51	38,49	2,08	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 2	61,53	38,47	2,07	0,62	0,62	0,58	0,57	0,62	0,58

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Weight Unl. Docs. = 0.3 # Components = 5	61,50	38,50	2,07	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.5 # Components = 10	61,39	38,61	2,05	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 1	61,34	38,66	2,06	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.5 # Components = 2	61,40	38,60	2,07	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.5 # Components = 5	61,32	38,68	2,05	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 10	61,09	38,91	2,01	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 1	61,34	38,66	1,89	0,61	0,61	0,58	0,57	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 2	61,06	38,94	1,90	0,61	0,61	0,57	0,57	0,61	0,57
		Weight Unl. Docs. = 0.7 # Components = 5	60,97	39,03	1,91	0,61	0,61	0,57	0,57	0,61	0,57
		Weight Unl. Docs. = 0.9 # Components = 10	60,62	39,38	1,99	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.9 # Components = 1	60,98	39,02	1,93	0,61	0,61	0,57	0,57	0,61	0,57
		Weight Unl. Docs. = 0.9 # Components = 2	60,62	39,38	1,93	0,61	0,61	0,57	0,56	0,61	0,57
		Weight Unl. Docs. = 0.9 # Components = 5	60,61	39,39	1,97	0,61	0,61	0,57	0,56	0,61	0,57
	70	Weight Unl. Docs. = 0.1 # Components = 10	61,84	38,16	2,34	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 1	61,83	38,17	2,34	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 2	61,84	38,16	2,34	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.1 # Components = 5	61,84	38,16	2,34	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 10	61,81	38,19	2,24	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 1	61,78	38,22	2,24	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 2	61,79	38,21	2,23	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.3 # Components = 5	61,83	38,17	2,24	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.5 # Components = 10	61,82	38,18	2,17	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.5 # Components = 1	61,78	38,22	2,12	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.5 # Components = 2	61,74	38,26	2,09	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.5 # Components = 5	61,84	38,16	2,11	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.7 # Components = 10	61,50	38,50	1,97	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.7 # Components = 1	61,65	38,35	2,01	0,62	0,62	0,58	0,58	0,62	0,58
		Weight Unl. Docs. = 0.7 # Components = 2	61,49	38,51	1,98	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.7 # Components = 5	61,59	38,41	2,04	0,62	0,62	0,58	0,57	0,62	0,58
		Weight Unl. Docs. = 0.9 # Components = 10	61,31	38,69	1,99	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.9 # Components = 1	61,45	38,55	1,94	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.9 # Components = 2	61,32	38,68	2,00	0,61	0,61	0,58	0,57	0,61	0,58
		Weight Unl. Docs. = 0.9 # Components = 5	61,27	38,73	1,88	0,61	0,61	0,58	0,57	0,61	0,58
MNB (Self Training)	1	# Ex. Inserted Training Set= 10	38,59	61,41	7,42	0,39	0,39	0,36	0,34	0,39	0,35
		# Ex. Inserted Training Set= 15	39,10	60,90	7,48	0,39	0,39	0,36	0,35	0,39	0,36
		# Ex. Inserted Training Set= 20	41,30	58,70	7,56	0,41	0,41	0,38	0,36	0,41	0,37
		# Ex. Inserted Training Set= 5	37,81	62,19	7,70	0,38	0,38	0,34	0,34	0,38	0,34

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
	10	# Ex. Inserted Training Set= 10	59,40	40,60	2,54	0,59	0,59	0,54	0,52	0,59	0,53
		# Ex. Inserted Training Set= 15	59,55	40,45	2,55	0,60	0,60	0,54	0,52	0,60	0,53
		# Ex. Inserted Training Set= 20	59,75	40,25	2,40	0,60	0,60	0,54	0,53	0,60	0,53
		# Ex. Inserted Training Set= 5	59,50	40,50	2,58	0,59	0,59	0,54	0,53	0,59	0,53
	20	# Ex. Inserted Training Set= 10	61,30	38,70	2,13	0,61	0,61	0,56	0,55	0,61	0,55
		# Ex. Inserted Training Set= 15	61,46	38,54	2,16	0,61	0,61	0,56	0,55	0,61	0,56
		# Ex. Inserted Training Set= 20	61,43	38,57	2,27	0,61	0,61	0,56	0,55	0,61	0,56
		# Ex. Inserted Training Set= 5	61,37	38,63	2,22	0,61	0,61	0,56	0,55	0,61	0,56
	30	# Ex. Inserted Training Set= 10	62,75	37,25	2,12	0,63	0,63	0,58	0,57	0,63	0,58
		# Ex. Inserted Training Set= 15	62,82	37,18	2,09	0,63	0,63	0,58	0,57	0,63	0,58
		# Ex. Inserted Training Set= 20	62,74	37,26	2,10	0,63	0,63	0,58	0,57	0,63	0,58
		# Ex. Inserted Training Set= 5	62,64	37,36	2,06	0,63	0,63	0,58	0,57	0,63	0,57
	40	# Ex. Inserted Training Set= 10	63,75	36,25	2,22	0,64	0,64	0,59	0,58	0,64	0,59
		# Ex. Inserted Training Set= 15	63,82	36,18	2,32	0,64	0,64	0,59	0,58	0,64	0,59
		# Ex. Inserted Training Set= 20	63,82	36,18	2,26	0,64	0,64	0,59	0,58	0,64	0,59
		# Ex. Inserted Training Set= 5	63,76	36,24	2,13	0,64	0,64	0,59	0,58	0,64	0,59
	50	# Ex. Inserted Training Set= 10	64,79	35,21	2,04	0,65	0,65	0,61	0,60	0,65	0,60
		# Ex. Inserted Training Set= 15	64,76	35,24	2,10	0,65	0,65	0,61	0,59	0,65	0,60
		# Ex. Inserted Training Set= 20	64,75	35,25	2,08	0,65	0,65	0,61	0,60	0,65	0,60
		# Ex. Inserted Training Set= 5	64,77	35,23	2,08	0,65	0,65	0,61	0,59	0,65	0,60
	60	# Ex. Inserted Training Set= 10	65,74	34,26	2,17	0,66	0,66	0,62	0,61	0,66	0,61
		# Ex. Inserted Training Set= 15	65,71	34,29	2,23	0,66	0,66	0,62	0,61	0,66	0,61
		# Ex. Inserted Training Set= 20	65,73	34,27	2,27	0,66	0,66	0,62	0,61	0,66	0,61
		# Ex. Inserted Training Set= 5	65,80	34,20	2,20	0,66	0,66	0,62	0,61	0,66	0,61
	70	# Ex. Inserted Training Set= 10	66,59	33,41	2,06	0,67	0,67	0,63	0,62	0,67	0,62
		# Ex. Inserted Training Set= 15	66,54	33,46	2,10	0,67	0,67	0,63	0,62	0,67	0,62
		# Ex. Inserted Training Set= 20	66,58	33,42	2,13	0,67	0,67	0,63	0,62	0,67	0,62
		# Ex. Inserted Training Set= 5	66,62	33,38	2,04	0,67	0,67	0,63	0,62	0,67	0,62
MNB	1		37,04	62,96	7,28	0,37	0,37	0,38	0,36	0,37	0,37
	10		42,31	57,69	3,71	0,42	0,42	0,42	0,42	0,42	0,42
	20		47,16	52,84	3,70	0,47	0,47	0,46	0,46	0,47	0,46
	30		49,35	50,65	3,55	0,49	0,49	0,48	0,48	0,49	0,48
	40		51,26	48,74	3,35	0,51	0,51	0,50	0,50	0,51	0,50
	50		52,75	47,25	3,44	0,53	0,53	0,51	0,51	0,53	0,51
	60		54,23	45,77	3,01	0,54	0,54	0,53	0,53	0,54	0,53
	70		55,06	44,94	3,36	0,55	0,55	0,53	0,54	0,55	0,54
KNN	1	k= 7	41,55	58,45	3,85	0,42	0,42	0,37	0,35	0,42	0,36

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		k= 17	41,55	58,45	3,85	0,42	0,42	0,37	0,35	0,42	0,36
		k= 37	41,55	58,45	3,85	0,42	0,42	0,37	0,35	0,42	0,36
		k= 57	41,55	58,45	3,85	0,42	0,42	0,37	0,35	0,42	0,36
	10	k= 7	43,02	56,98	3,50	0,43	0,43	0,42	0,42	0,43	0,42
		k= 17	43,05	56,95	3,49	0,43	0,43	0,42	0,42	0,43	0,42
		k= 37	43,05	56,95	3,49	0,43	0,43	0,42	0,42	0,43	0,42
		k= 57	43,05	56,95	3,49	0,43	0,43	0,42	0,42	0,43	0,42
	20	k= 7	46,82	53,18	3,72	0,47	0,47	0,45	0,45	0,47	0,45
		k= 17	46,91	53,09	3,57	0,47	0,47	0,46	0,45	0,47	0,45
		k= 37	46,88	53,12	3,56	0,47	0,47	0,45	0,45	0,47	0,45
		k= 57	46,88	53,12	3,56	0,47	0,47	0,45	0,45	0,47	0,45
	30	k= 7	48,73	51,27	3,99	0,49	0,49	0,47	0,47	0,49	0,47
		k= 17	49,09	50,91	4,23	0,49	0,49	0,48	0,48	0,49	0,48
		k= 37	49,11	50,89	4,24	0,49	0,49	0,48	0,48	0,49	0,48
		k= 57	49,13	50,87	4,25	0,49	0,49	0,48	0,48	0,49	0,48
	40	k= 7	50,15	49,85	3,87	0,50	0,50	0,48	0,48	0,50	0,48
		k= 17	50,88	49,12	3,94	0,51	0,51	0,50	0,49	0,51	0,50
		k= 37	50,89	49,11	3,78	0,51	0,51	0,50	0,50	0,51	0,50
		k= 57	50,96	49,04	3,74	0,51	0,51	0,50	0,50	0,51	0,50
	50	k= 7	50,83	49,17	3,66	0,51	0,51	0,49	0,49	0,51	0,49
		k= 17	51,93	48,07	3,73	0,52	0,52	0,50	0,50	0,52	0,50
		k= 37	51,97	48,03	3,79	0,52	0,52	0,51	0,51	0,52	0,51
		k= 57	51,91	48,09	3,73	0,52	0,52	0,51	0,51	0,52	0,51
	60	k= 7	51,91	48,09	3,26	0,52	0,52	0,50	0,50	0,52	0,50
		k= 17	52,97	47,03	3,44	0,53	0,53	0,51	0,52	0,53	0,51
		k= 37	53,10	46,90	3,29	0,53	0,53	0,52	0,52	0,53	0,52
		k= 57	53,12	46,88	3,34	0,53	0,53	0,52	0,52	0,53	0,52
	70	k= 7	52,57	47,43	3,35	0,53	0,53	0,51	0,51	0,53	0,51
		k= 17	53,91	46,09	3,18	0,54	0,54	0,52	0,52	0,54	0,52
		k= 37	54,31	45,69	3,33	0,54	0,54	0,53	0,53	0,54	0,53
		k= 57	54,28	45,72	3,37	0,54	0,54	0,53	0,53	0,54	0,53
SVM	1	Kernel=Linear C= 0.001	36,58	63,42	9,02	0,37	0,37	0,35	0,34	0,37	0,35
		Kernel=Linear C= 0.01	36,55	63,45	9,04	0,37	0,37	0,34	0,34	0,37	0,34
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	36,58	63,42	9,04	0,37	0,37	0,35	0,34	0,37	0,35
		Kernel=Linear C= 1	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Linear C= 0.0001	37,41	62,59	8,67	0,37	0,37	0,35	0,35	0,37	0,35

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 0.00001	36,80	63,20	8,87	0,37	0,37	0,35	0,35	0,37	0,35
		Kernel=Linear C= 10	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Linear C= 100	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Linear C= 1000	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Linear C= 10000	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Linear C= 100000	37,45	62,55	8,64	0,37	0,37	0,36	0,35	0,37	0,36
		Kernel=Poly C= 0.001	35,95	64,05	9,66	0,36	0,36	0,25	0,34	0,36	0,29
		Kernel=Poly C= 0.01	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 1	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 0.0001	35,95	64,05	9,66	0,36	0,36	0,25	0,34	0,36	0,29
		Kernel=Poly C= 0.00001	35,95	64,05	9,66	0,36	0,36	0,25	0,34	0,36	0,29
		Kernel=Poly C= 10	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 100	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 1000	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 10000	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=Poly C= 100000	36,02	63,98	9,54	0,36	0,36	0,26	0,34	0,36	0,30
		Kernel=RBF C= 0.001	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 0.01	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 1	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 0.0001	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 0.00001	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 10	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 100	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 1000	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 10000	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
		Kernel=RBF C= 100000	37,68	62,32	8,48	0,38	0,38	0,36	0,35	0,38	0,36
SVM	10	Kernel=Linear C= 0.001	38,53	61,47	8,08	0,39	0,39	0,37	0,37	0,39	0,37
		Kernel=Linear C= 0.01	39,09	60,91	8,86	0,39	0,39	0,34	0,37	0,39	0,35
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	40,25	59,75	7,11	0,40	0,40	0,44	0,39	0,40	0,42
		Kernel=Linear C= 1	41,20	58,80	6,34	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Linear C= 0.0001	40,25	59,75	6,99	0,40	0,40	0,45	0,39	0,40	0,42
		Kernel=Linear C= 0.00001	39,25	60,75	7,41	0,39	0,39	0,45	0,39	0,39	0,41

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 10	40,90	59,10	6,47	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Linear C= 100	40,90	59,10	6,47	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Linear C= 1000	40,90	59,10	6,47	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Linear C= 10000	40,90	59,10	6,47	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Linear C= 100000	40,90	59,10	6,47	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=Poly C= 0.001	36,15	63,85	9,30	0,36	0,36	0,29	0,34	0,36	0,32
		Kernel=Poly C= 0.01	37,57	62,43	9,54	0,38	0,38	0,47	0,35	0,38	0,40
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	38,35	61,65	9,39	0,38	0,38	0,47	0,37	0,38	0,41
		Kernel=Poly C= 1	34,77	65,23	9,56	0,35	0,35	0,46	0,36	0,35	0,41
		Kernel=Poly C= 0.0001	36,13	63,87	9,25	0,36	0,36	0,34	0,34	0,36	0,34
		Kernel=Poly C= 0.00001	35,47	64,53	9,87	0,35	0,35	0,38	0,35	0,35	0,36
		Kernel=Poly C= 10	34,56	65,44	9,59	0,35	0,35	0,46	0,36	0,35	0,41
		Kernel=Poly C= 100	34,56	65,44	9,59	0,35	0,35	0,46	0,36	0,35	0,41
		Kernel=Poly C= 1000	34,56	65,44	9,59	0,35	0,35	0,46	0,36	0,35	0,41
		Kernel=Poly C= 10000	34,56	65,44	9,60	0,35	0,35	0,47	0,36	0,35	0,41
		Kernel=Poly C= 100000	34,56	65,44	9,60	0,35	0,35	0,47	0,36	0,35	0,41
		Kernel=RBF C= 0.001	39,67	60,33	7,92	0,40	0,40	0,46	0,39	0,40	0,42
		Kernel=RBF C= 0.01	39,55	60,45	7,90	0,40	0,40	0,44	0,38	0,40	0,41
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	38,71	61,29	8,37	0,39	0,39	0,31	0,36	0,39	0,33
		Kernel=RBF C= 1	37,34	62,66	9,04	0,37	0,37	0,30	0,36	0,37	0,33
		Kernel=RBF C= 0.0001	39,67	60,33	7,92	0,40	0,40	0,46	0,39	0,40	0,42
		Kernel=RBF C= 0.00001	39,67	60,33	7,92	0,40	0,40	0,46	0,39	0,40	0,42
		Kernel=RBF C= 10	42,31	57,69	6,83	0,42	0,42	0,44	0,41	0,42	0,42
		Kernel=RBF C= 100	40,96	59,04	6,80	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=RBF C= 1000	40,92	59,08	6,77	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=RBF C= 10000	40,96	59,04	6,80	0,41	0,41	0,43	0,41	0,41	0,42
		Kernel=RBF C= 100000	40,96	59,04	6,80	0,41	0,41	0,43	0,41	0,41	0,42
SVM	20	Kernel=Linear C= 0.001	42,35	57,65	7,70	0,42	0,42	0,39	0,39	0,42	0,39
		Kernel=Linear C= 0.01	41,77	58,23	6,96	0,42	0,42	0,40	0,39	0,42	0,39
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	45,17	54,83	5,34	0,45	0,45	0,47	0,43	0,45	0,45
		Kernel=Linear C= 1	46,21	53,79	4,26	0,46	0,46	0,46	0,45	0,46	0,45
		Kernel=Linear C= 0.0001	44,07	55,93	7,07	0,44	0,44	0,49	0,42	0,44	0,45
		Kernel=Linear C= 0.00001	43,96	56,04	7,43	0,44	0,44	0,49	0,42	0,44	0,45
		Kernel=Linear C= 10	45,65	54,35	4,66	0,46	0,46	0,46	0,45	0,46	0,45

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 100	45,66	54,34	4,67	0,46	0,46	0,46	0,45	0,46	0,45
		Kernel=Linear C= 1000	45,66	54,34	4,68	0,46	0,46	0,46	0,45	0,46	0,45
		Kernel=Linear C= 10000	45,66	54,34	4,67	0,46	0,46	0,46	0,45	0,46	0,45
		Kernel=Linear C= 100000	45,65	54,35	4,66	0,46	0,46	0,46	0,45	0,46	0,45
		Kernel=Poly C= 0.001	37,62	62,38	8,81	0,38	0,38	0,35	0,34	0,38	0,35
		Kernel=Poly C= 0.01	39,39	60,61	8,23	0,39	0,39	0,54	0,36	0,39	0,43
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	45,11	54,89	6,09	0,45	0,45	0,50	0,40	0,45	0,44
		Kernel=Poly C= 1	39,48	60,52	9,53	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=Poly C= 0.0001	38,85	61,15	8,36	0,39	0,39	0,41	0,35	0,39	0,38
		Kernel=Poly C= 0.00001	38,06	61,94	8,04	0,38	0,38	0,40	0,35	0,38	0,37
		Kernel=Poly C= 10	39,20	60,80	9,65	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=Poly C= 100	39,20	60,80	9,65	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=Poly C= 1000	39,20	60,80	9,65	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=Poly C= 10000	39,21	60,79	9,65	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=Poly C= 100000	39,21	60,79	9,65	0,39	0,39	0,49	0,40	0,39	0,44
		Kernel=RBF C= 0.001	44,85	55,15	6,77	0,45	0,45	0,49	0,42	0,45	0,45
		Kernel=RBF C= 0.01	43,50	56,50	7,11	0,44	0,44	0,42	0,39	0,44	0,40
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	42,83	57,17	6,02	0,43	0,43	0,36	0,38	0,43	0,37
		Kernel=RBF C= 1	42,20	57,80	5,94	0,42	0,42	0,35	0,38	0,42	0,37
		Kernel=RBF C= 0.0001	44,85	55,15	6,77	0,45	0,45	0,49	0,42	0,45	0,45
		Kernel=RBF C= 0.00001	44,85	55,15	6,77	0,45	0,45	0,49	0,42	0,45	0,45
		Kernel=RBF C= 10	47,93	52,07	4,34	0,48	0,48	0,47	0,45	0,48	0,46
		Kernel=RBF C= 100	45,57	54,43	4,74	0,46	0,46	0,46	0,44	0,46	0,45
		Kernel=RBF C= 1000	45,51	54,49	4,84	0,46	0,46	0,46	0,44	0,46	0,45
		Kernel=RBF C= 10000	45,39	54,61	4,87	0,45	0,45	0,46	0,44	0,45	0,45
		Kernel=RBF C= 100000	45,45	54,55	4,80	0,45	0,45	0,46	0,44	0,45	0,45
SVM	30	Kernel=Linear C= 0.001	41,39	58,61	8,30	0,41	0,41	0,32	0,38	0,41	0,35
		Kernel=Linear C= 0.01	45,08	54,92	7,35	0,45	0,45	0,44	0,41	0,45	0,43
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	49,66	50,34	4,31	0,50	0,50	0,49	0,47	0,50	0,48
		Kernel=Linear C= 1	48,56	51,44	4,42	0,49	0,49	0,48	0,47	0,49	0,47
		Kernel=Linear C= 0.0001	46,95	53,05	6,88	0,47	0,47	0,51	0,44	0,47	0,47
		Kernel=Linear C= 0.00001	46,59	53,41	7,02	0,47	0,47	0,51	0,44	0,47	0,47
		Kernel=Linear C= 10	47,87	52,13	4,61	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=Linear C= 100	47,88	52,12	4,62	0,48	0,48	0,48	0,47	0,48	0,47

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 1000	47,88	52,12	4,62	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=Linear C= 10000	47,88	52,12	4,62	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=Linear C= 100000	47,88	52,12	4,62	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=Poly C= 0.001	37,78	62,22	8,23	0,38	0,38	0,29	0,34	0,38	0,31
		Kernel=Poly C= 0.01	42,09	57,91	7,62	0,42	0,42	0,57	0,38	0,42	0,46
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	48,28	51,72	5,35	0,48	0,48	0,51	0,43	0,48	0,47
		Kernel=Poly C= 1	44,17	55,83	7,99	0,44	0,44	0,50	0,43	0,44	0,46
		Kernel=Poly C= 0.0001	41,84	58,16	7,06	0,42	0,42	0,44	0,36	0,42	0,40
		Kernel=Poly C= 0.00001	38,44	61,56	8,90	0,38	0,38	0,41	0,35	0,38	0,38
		Kernel=Poly C= 10	43,10	56,90	8,21	0,43	0,43	0,50	0,43	0,43	0,46
		Kernel=Poly C= 100	43,06	56,94	8,20	0,43	0,43	0,50	0,42	0,43	0,46
		Kernel=Poly C= 1000	43,06	56,94	8,20	0,43	0,43	0,50	0,42	0,43	0,46
		Kernel=Poly C= 10000	43,06	56,94	8,20	0,43	0,43	0,50	0,42	0,43	0,46
		Kernel=Poly C= 100000	43,06	56,94	8,20	0,43	0,43	0,50	0,42	0,43	0,46
		Kernel=RBF C= 0.001	48,01	51,99	6,02	0,48	0,48	0,53	0,44	0,48	0,48
		Kernel=RBF C= 0.01	43,48	56,52	8,50	0,43	0,43	0,41	0,40	0,43	0,41
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	44,70	55,30	6,76	0,45	0,45	0,39	0,40	0,45	0,40
		Kernel=RBF C= 1	46,68	53,32	6,41	0,47	0,47	0,44	0,42	0,47	0,43
		Kernel=RBF C= 0.0001	48,01	51,99	6,02	0,48	0,48	0,53	0,44	0,48	0,48
		Kernel=RBF C= 0.00001	48,01	51,99	6,02	0,48	0,48	0,53	0,44	0,48	0,48
		Kernel=RBF C= 10	49,96	50,04	4,02	0,50	0,50	0,49	0,47	0,50	0,48
		Kernel=RBF C= 100	48,27	51,73	4,84	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=RBF C= 1000	48,04	51,96	4,96	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=RBF C= 10000	47,98	52,02	4,84	0,48	0,48	0,48	0,47	0,48	0,47
		Kernel=RBF C= 100000	47,98	52,02	4,84	0,48	0,48	0,48	0,47	0,48	0,47
SVM	40	Kernel=Linear C= 0.001	42,95	57,05	9,01	0,43	0,43	0,32	0,39	0,43	0,35
		Kernel=Linear C= 0.01	45,82	54,18	7,46	0,46	0,46	0,46	0,43	0,46	0,44
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	50,84	49,16	4,48	0,51	0,51	0,51	0,49	0,51	0,50
		Kernel=Linear C= 1	50,57	49,43	4,06	0,51	0,51	0,49	0,49	0,51	0,49
		Kernel=Linear C= 0.0001	47,79	52,21	6,47	0,48	0,48	0,51	0,45	0,48	0,48
		Kernel=Linear C= 0.00001	47,58	52,42	6,80	0,48	0,48	0,53	0,45	0,48	0,48
		Kernel=Linear C= 10	49,50	50,50	4,27	0,49	0,49	0,49	0,49	0,49	0,49
		Kernel=Linear C= 100	49,48	50,52	4,29	0,49	0,49	0,49	0,49	0,49	0,49
		Kernel=Linear C= 1000	49,48	50,52	4,29	0,49	0,49	0,49	0,49	0,49	0,49

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 10000	49,49	50,51	4,30	0,49	0,49	0,49	0,49	0,49	0,49
		Kernel=Linear C= 100000	49,48	50,52	4,29	0,49	0,49	0,49	0,49	0,49	0,49
		Kernel=Poly C= 0.001	37,85	62,15	8,27	0,38	0,38	0,30	0,35	0,38	0,32
		Kernel=Poly C= 0.01	43,30	56,70	7,46	0,43	0,43	0,56	0,39	0,43	0,46
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	50,64	49,36	4,79	0,51	0,51	0,52	0,46	0,51	0,48
		Kernel=Poly C= 1	47,11	52,89	6,29	0,47	0,47	0,50	0,45	0,47	0,47
		Kernel=Poly C= 0.0001	41,38	58,62	7,87	0,41	0,41	0,47	0,37	0,41	0,41
		Kernel=Poly C= 0.00001	39,82	60,18	9,27	0,40	0,40	0,46	0,36	0,40	0,40
		Kernel=Poly C= 10	44,57	55,43	7,59	0,45	0,45	0,50	0,44	0,45	0,47
		Kernel=Poly C= 100	44,33	55,67	7,82	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 1000	44,33	55,67	7,82	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 10000	44,34	55,66	7,81	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 100000	44,33	55,67	7,82	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=RBF C= 0.001	48,61	51,39	5,91	0,49	0,49	0,53	0,45	0,49	0,49
		Kernel=RBF C= 0.01	44,13	55,87	8,70	0,44	0,44	0,41	0,41	0,44	0,41
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	45,49	54,51	6,77	0,45	0,45	0,40	0,41	0,45	0,40
		Kernel=RBF C= 1	46,74	53,26	6,65	0,47	0,47	0,44	0,42	0,47	0,43
		Kernel=RBF C= 0.0001	48,61	51,39	5,91	0,49	0,49	0,53	0,45	0,49	0,49
		Kernel=RBF C= 0.00001	48,61	51,39	5,91	0,49	0,49	0,53	0,45	0,49	0,49
		Kernel=RBF C= 10	52,03	47,97	3,93	0,52	0,52	0,51	0,50	0,52	0,50
		Kernel=RBF C= 100	50,07	49,93	4,22	0,50	0,50	0,49	0,49	0,50	0,49
		Kernel=RBF C= 1000	49,44	50,56	4,36	0,49	0,49	0,49	0,48	0,49	0,49
		Kernel=RBF C= 10000	49,44	50,56	4,36	0,49	0,49	0,49	0,48	0,49	0,49
		Kernel=RBF C= 100000	49,44	50,56	4,36	0,49	0,49	0,49	0,48	0,49	0,49
SVM	50	Kernel=Linear C= 0.001	42,26	57,74	8,58	0,42	0,42	0,30	0,38	0,42	0,34
		Kernel=Linear C= 0.01	47,98	52,02	6,12	0,48	0,48	0,49	0,45	0,48	0,47
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	52,67	47,33	3,94	0,53	0,53	0,52	0,50	0,53	0,51
		Kernel=Linear C= 1	51,76	48,24	4,02	0,52	0,52	0,50	0,50	0,52	0,50
		Kernel=Linear C= 0.0001	47,50	52,50	6,79	0,48	0,48	0,52	0,45	0,48	0,48
		Kernel=Linear C= 0.00001	48,52	51,48	6,62	0,49	0,49	0,54	0,46	0,49	0,50
		Kernel=Linear C= 10	50,54	49,46	4,22	0,51	0,51	0,50	0,50	0,51	0,50
		Kernel=Linear C= 100	50,55	49,45	4,23	0,51	0,51	0,50	0,50	0,51	0,50
		Kernel=Linear C= 1000	50,54	49,46	4,22	0,51	0,51	0,50	0,50	0,51	0,50
		Kernel=Linear C= 10000	50,55	49,45	4,23	0,51	0,51	0,50	0,50	0,51	0,50

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Linear C= 100000	50,55	49,45	4,22	0,51	0,51	0,50	0,50	0,51	0,50
		Kernel=Poly C= 0.001	38,15	61,85	8,62	0,38	0,38	0,29	0,35	0,38	0,32
		Kernel=Poly C= 0.01	43,70	56,30	7,85	0,44	0,44	0,57	0,40	0,44	0,47
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	52,21	47,79	4,65	0,52	0,52	0,53	0,47	0,52	0,50
		Kernel=Poly C= 1	49,43	50,57	4,85	0,49	0,49	0,50	0,46	0,49	0,48
		Kernel=Poly C= 0.0001	40,33	59,67	9,24	0,40	0,40	0,51	0,38	0,40	0,44
		Kernel=Poly C= 0.00001	41,31	58,69	6,73	0,41	0,41	0,49	0,36	0,41	0,42
		Kernel=Poly C= 10	44,55	55,45	7,50	0,45	0,45	0,50	0,44	0,45	0,47
		Kernel=Poly C= 100	44,40	55,60	7,70	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 1000	44,40	55,60	7,70	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 10000	44,40	55,60	7,70	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=Poly C= 100000	44,41	55,59	7,70	0,44	0,44	0,50	0,44	0,44	0,47
		Kernel=RBF C= 0.001	50,14	49,86	5,43	0,50	0,50	0,55	0,46	0,50	0,50
		Kernel=RBF C= 0.01	45,14	54,86	9,36	0,45	0,45	0,42	0,41	0,45	0,42
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	46,31	53,69	8,56	0,46	0,46	0,46	0,43	0,46	0,44
		Kernel=RBF C= 1	47,30	52,70	6,75	0,47	0,47	0,49	0,44	0,47	0,46
		Kernel=RBF C= 0.0001	50,14	49,86	5,43	0,50	0,50	0,55	0,46	0,50	0,50
		Kernel=RBF C= 0.00001	50,14	49,86	5,43	0,50	0,50	0,55	0,46	0,50	0,50
		Kernel=RBF C= 10	53,64	46,36	3,61	0,54	0,54	0,52	0,51	0,54	0,52
		Kernel=RBF C= 100	51,03	48,97	4,05	0,51	0,51	0,50	0,50	0,51	0,50
		Kernel=RBF C= 1000	50,43	49,57	4,14	0,50	0,50	0,50	0,49	0,50	0,50
		Kernel=RBF C= 10000	50,44	49,56	4,15	0,50	0,50	0,50	0,49	0,50	0,50
		Kernel=RBF C= 100000	50,44	49,56	4,14	0,50	0,50	0,50	0,49	0,50	0,50
SVM	60	Kernel=Linear C= 0.001	42,37	57,63	8,75	0,42	0,42	0,30	0,39	0,42	0,34
		Kernel=Linear C= 0.01	47,35	52,65	7,52	0,47	0,47	0,47	0,45	0,47	0,46
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	54,12	45,88	4,09	0,54	0,54	0,53	0,52	0,54	0,52
		Kernel=Linear C= 1	53,11	46,89	3,60	0,53	0,53	0,51	0,52	0,53	0,51
		Kernel=Linear C= 0.0001	48,75	51,25	7,41	0,49	0,49	0,53	0,46	0,49	0,49
		Kernel=Linear C= 0.00001	49,22	50,78	7,08	0,49	0,49	0,54	0,47	0,49	0,50
		Kernel=Linear C= 10	51,42	48,58	3,70	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=Linear C= 100	51,40	48,60	3,69	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=Linear C= 1000	51,41	48,59	3,69	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=Linear C= 10000	51,40	48,60	3,69	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=Linear C= 100000	51,40	48,60	3,69	0,51	0,51	0,50	0,51	0,51	0,50

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Poly C= 0.001	40,29	59,71	6,44	0,40	0,40	0,35	0,35	0,40	0,35
		Kernel=Poly C= 0.01	44,44	55,56	7,16	0,44	0,44	0,57	0,41	0,44	0,48
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	53,92	46,08	4,02	0,54	0,54	0,54	0,49	0,54	0,51
		Kernel=Poly C= 1	51,11	48,89	3,89	0,51	0,51	0,51	0,48	0,51	0,50
		Kernel=Poly C= 0.0001	41,80	58,20	8,77	0,42	0,42	0,50	0,38	0,42	0,43
		Kernel=Poly C= 0.00001	40,73	59,27	7,59	0,41	0,41	0,48	0,36	0,41	0,41
		Kernel=Poly C= 10	45,65	54,35	7,06	0,46	0,46	0,51	0,46	0,46	0,48
		Kernel=Poly C= 100	45,46	54,54	7,25	0,45	0,45	0,51	0,46	0,45	0,48
		Kernel=Poly C= 1000	45,46	54,54	7,25	0,45	0,45	0,51	0,46	0,45	0,48
		Kernel=Poly C= 10000	45,46	54,54	7,24	0,45	0,45	0,51	0,46	0,45	0,48
		Kernel=Poly C= 100000	45,46	54,54	7,25	0,45	0,45	0,51	0,46	0,45	0,48
		Kernel=RBF C= 0.001	50,76	49,24	6,50	0,51	0,51	0,55	0,47	0,51	0,51
		Kernel=RBF C= 0.01	45,09	54,91	9,60	0,45	0,45	0,40	0,42	0,45	0,41
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	45,38	54,62	7,50	0,45	0,45	0,40	0,42	0,45	0,41
		Kernel=RBF C= 1	48,64	51,36	8,08	0,49	0,49	0,49	0,45	0,49	0,47
		Kernel=RBF C= 0.0001	50,76	49,24	6,50	0,51	0,51	0,55	0,47	0,51	0,51
		Kernel=RBF C= 0.00001	50,76	49,24	6,50	0,51	0,51	0,55	0,47	0,51	0,51
		Kernel=RBF C= 10	55,23	44,77	3,42	0,55	0,55	0,53	0,53	0,55	0,53
		Kernel=RBF C= 100	52,20	47,80	3,77	0,52	0,52	0,51	0,51	0,52	0,51
		Kernel=RBF C= 1000	51,43	48,57	3,81	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=RBF C= 10000	51,41	48,59	3,80	0,51	0,51	0,50	0,51	0,51	0,50
		Kernel=RBF C= 100000	51,43	48,57	3,80	0,51	0,51	0,50	0,51	0,51	0,50
SVM	70	Kernel=Linear C= 0.001	43,11	56,89	10,17	0,43	0,43	0,33	0,40	0,43	0,36
		Kernel=Linear C= 0.01	47,71	52,29	8,36	0,48	0,48	0,49	0,45	0,48	0,47
		Kernel=Linear C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Linear C= 0.1	55,52	44,48	3,89	0,56	0,56	0,54	0,53	0,56	0,54
		Kernel=Linear C= 1	54,45	45,55	3,77	0,54	0,54	0,53	0,53	0,54	0,53
		Kernel=Linear C= 0.0001	50,27	49,73	5,64	0,50	0,50	0,53	0,47	0,50	0,50
		Kernel=Linear C= 0.00001	50,75	49,25	5,89	0,51	0,51	0,56	0,48	0,51	0,52
		Kernel=Linear C= 10	52,71	47,29	3,76	0,53	0,53	0,51	0,52	0,53	0,51
		Kernel=Linear C= 100	52,68	47,32	3,78	0,53	0,53	0,51	0,52	0,53	0,52
		Kernel=Linear C= 1000	52,64	47,36	3,81	0,53	0,53	0,51	0,52	0,53	0,51
		Kernel=Linear C= 10000	52,65	47,35	3,81	0,53	0,53	0,51	0,52	0,53	0,51
		Kernel=Linear C= 100000	52,65	47,35	3,81	0,53	0,53	0,51	0,52	0,53	0,51
		Kernel=Poly C= 0.001	39,14	60,86	7,68	0,39	0,39	0,30	0,35	0,39	0,32

Algorithm	Number of Labeled Examples per Class (percent)	Parameters	Accuracy (%)	Error (%)	Standard Deviation (Accuracy)	Micro-Precision	Micro-Recall	Macro-Precision	Macro-Recall	Micro-F1	Macro-F1
		Kernel=Poly C= 0.01	46,73	53,27	6,20	0,47	0,47	0,58	0,42	0,47	0,48
		Kernel=Poly C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=Poly C= 0.1	54,82	45,18	3,97	0,55	0,55	0,55	0,50	0,55	0,52
		Kernel=Poly C= 1	52,28	47,72	3,76	0,52	0,52	0,52	0,49	0,52	0,50
		Kernel=Poly C= 0.0001	41,50	58,50	9,55	0,42	0,42	0,53	0,38	0,42	0,44
		Kernel=Poly C= 0.00001	38,53	61,47	8,94	0,39	0,39	0,41	0,36	0,39	0,38
		Kernel=Poly C= 10	46,42	53,58	6,96	0,46	0,46	0,51	0,46	0,46	0,49
		Kernel=Poly C= 100	45,82	54,18	7,26	0,46	0,46	0,51	0,46	0,46	0,48
		Kernel=Poly C= 1000	45,81	54,19	7,27	0,46	0,46	0,51	0,46	0,46	0,48
		Kernel=Poly C= 10000	45,82	54,18	7,26	0,46	0,46	0,51	0,46	0,46	0,48
		Kernel=Poly C= 100000	45,81	54,19	7,27	0,46	0,46	0,51	0,46	0,46	0,48
		Kernel=RBF C= 0.001	51,85	48,15	6,22	0,52	0,52	0,57	0,48	0,52	0,52
		Kernel=RBF C= 0.01	43,26	56,74	10,58	0,43	0,43	0,35	0,41	0,43	0,38
		Kernel=RBF C= 0	42,65	57,35	2,72	0,43	0,43	0,14	0,33	0,43	0,20
		Kernel=RBF C= 0.1	47,58	52,42	7,15	0,48	0,48	0,46	0,44	0,48	0,45
		Kernel=RBF C= 1	48,25	51,75	7,03	0,48	0,48	0,50	0,46	0,48	0,48
		Kernel=RBF C= 0.0001	51,85	48,15	6,22	0,52	0,52	0,57	0,48	0,52	0,52
		Kernel=RBF C= 0.00001	51,85	48,15	6,22	0,52	0,52	0,57	0,48	0,52	0,52
		Kernel=RBF C= 10	55,93	44,07	3,27	0,56	0,56	0,54	0,53	0,56	0,54
		Kernel=RBF C= 100	53,71	46,29	3,82	0,54	0,54	0,52	0,52	0,54	0,52
		Kernel=RBF C= 1000	52,77	47,23	4,03	0,53	0,53	0,52	0,52	0,53	0,52
		Kernel=RBF C= 10000	52,70	47,30	4,06	0,53	0,53	0,51	0,52	0,53	0,52
		Kernel=RBF C= 100000	52,71	47,29	4,06	0,53	0,53	0,51	0,52	0,53	0,52