Explorando Informação Relacional para Análise de Sentimentos em Redes Sociais

por

Juliano Cícero Bitu Rabelo

Tese de Doutorado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

Recife
2015

Universidade Federal de Pernambuco

Centro de Informática

Pós-graduação em Ciência da Computação

Juliano Cícero Bitu Rabelo

Explorando Informação Relacional para Análise de Sentimentos em Redes Sociais

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIA DA COMPUTAÇÃO.

Orientador: **Ricardo Bastos Cavalcanti Prudêncio**
Coorientadora: **Flávia de Almeida Barros**

Recife
2015

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

R114e Rabelo, Juliano Cícero Bitu
Explorando informação relacional para análise de sentimentos em redes sociais / Juliano Cícero Bitu Rabelo. – Recife: O Autor, 2015.
129 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da computação, 2015.
Inclui referências.

1. Inteligência artificial. 2. Aprendizado de máquina. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2015-169

Juliano Cícero Bitu Rabelo

Explorando Informação Relacional para Análise de Sentimento em Redes Sociais

Trabalho aprovado. Recife, 25/08/2015:

_____	_____
Ricardo B. Cavalcanti Prudencio	**Flávia de Almeida Barros**
Orientador	Coorientadora
_____	_____
Carlos Guimarães Ferraz	**Geber Lisboa Ramalho**
_____	_____
Patrícia Restelli Tedesco	**Renato Fernandes Corrêa**
_____	
Wagner Meira Júnior	

Recife
2015

Resumo

A web, inicialmente um mero repositório de informações estáticas, transformou-se numa enorme fonte de aplicações diversas, proporcionando ou fomentando entretenimento, negócios e relacionamentos. Com essa evolução, a web passou a conter uma enorme quantidade de informações valiosas sobre produtos e serviços, especialmente em sites de compra, sites específicos para avaliação de produtos e até mesmo em redes sociais. Com as ferramentas adequadas, é possível monitorar opiniões ou mensurar a aceitação de um objeto qualquer a partir de dados disponíveis online, ao invés de realizar pesquisas de opinião usuais, que são demoradas, trabalhosas, tem alto custo e alcançam um número bastante restrito de pessoas. Com o monitoramento online, todo o processo de consolidação de opiniões pode ser realizado de forma automática, oferecendo um feedback imediato e mais representativo sobre o objeto avaliado. O problema geral desta proposta de tese é a classificação dos usuários de acordo com suas opiniões sobre um objeto de interesse. Comumente, a classificação das opiniões emitidas por um dado usuário é feita através da classificação de sentimentos expressos em textos, postagens ou comentários. Se a classificação de opiniões, no entanto, for realizada em ambientes nos quais haja conexões entre seus usuários (como as redes sociais), uma nova dimensão de informação se apresenta: através da análise dos relacionamentos, é possível inferir as opiniões de usuários a partir da opinião de seus contatos. A abordagem proposta neste trabalho para realização de análise de sentimento em redes sociais é baseada no princípio da assortatividade, que estabelece que indivíduos tendem a se conectar a outros com os quais apresentam alto grau de semelhança. A partir desse conceito, são aplicadas técnicas de classificação coletiva sobre o grafo que representa a rede social. A intenção é explorar o fato de que a classificação coletiva não utiliza apenas as características locais dos nós no processo de inferência, mas também as características e classes dos nós relacionados. Além disso, a classificação é executada de forma simultânea sobre todas as instâncias, o que permite considerar as influências que cada instância exerce sobre outras às quais está relacionada. Para avaliação da viabilidade do método proposto, foi implementado um protótipo que usa um algoritmo de relaxation labeling para a classificação coletiva de opiniões de usuários, e foi desenvolvido um estudo de caso para predição de preferência política de usuários do Twitter, que alcançou resultados promissores.

Palavras-chave: Classificação Coletiva; Redes sociais; Processamento de linguagem natural.

Abstract

The web, which was initially a mere repository for static information, has turned into a huge source of different applications, containing not only information but also promoting entertainment, business and relationships. Thus, the web currently has plenty of valuable information on products and services, especially in shopping, product evaluation and social networks websites. With the proper tools, it is possible to monitor opinions or to measure acceptance of a given object from data available online, instead of running usual polls, which are time and labor consuming, expensive and have limited reach. With online monitoring, the opinion consolidation process may be done automatically, offering an immediate, representative feedback on the evaluated object. This thesis proposal general problem is the classification of users according to his/her opinions given a target object. Commonly, the user opinion classification is performed through the use of text classifiers over his/her texts, comments or posts. If this opinion classification process takes place in environments where there are connections among its users (like social networks), a new information dimension shows up: through analysis of users relationships, it is possible to infer users opinions by using his/her contacts opinions. The approach proposed here to social networks sentiment analysis is based on the homophily principle, which states that users are more likely to connect to similar others. Using that concept, we apply collective classification techniques on the graph that represents the social network. The intention is to leverage the fact that collective classification uses not only the local node features in the inference process, but also the features and classes of the neighborhood. Besides, the classification is executed simultaneously on all nodes, which allows considering the influences of each node on its neighbors. To evaluate the proposed method, we implemented a prototype which uses a relaxation labeling algorithm for the collective classification of users opinions, and developed a case study to predict the political preference of users in Twitter, achieving promising results.

Keywords: Collective classification; Social networks; Natural language processing.

Índice de Ilustrações

Figura 1: Arquitetura usual para sistemas de Análise de Sentimento 25

Figura 2: Exemplo de classificação coletiva modelado através de Campos Aleatórios de Markov 54

Figura 3: Cálculo dos ϕ cliques potenciais para o exemplo apresentado na Figura 2 54

Figura 4: Exemplo de classificação coletiva modelado através de Campos Aleatórios de Markov 58

Figura 5: Macroarquitetura da abordagem proposta.....72

Figura 6: Grafo construído a partir da coleta de *tweets* contendo *hashtags* específicas (restrições: nós autores com no mínimo 2 *tweets* e nós quaisquer com no mínimo 5 conexões). Os nós azuis são os democratas, os vermelhos são os republicanos e os cinzas são os nós de usuários que não se manifestaram. ..92

Figura 7: Grafo resultante da aplicação do algortimo de pruning sobre o grafo inicial apresentado na Figura 6. Os nós azuis são os democratas, os vermelhos são os republicanos e os cinzas são os nós de usuários que não se manifestaram. O grafo é direcionado e as setas representam a relação seguidor-seguido (as setas mais largas são devido a falha de renderização do aplicativo usado para visualização).....93

Figura 8: Representação gráfica dos dados da Tabela 2 (precisão do classificador textual). O eixo horizontal representa a variação do parâmetro θ_2. 98

Figura 9: Representação gráfica dos dados da Tabela 3 (precisão do classificador textual considerando apenas os nós que atendem aos limiares definidos). O eixo horizontal representa a variação do parâmetro θ_2.....100

Figura 10: Precisão e erro do classificador coletivo sobre os nós que não atenderam aos limiares definidos e por isso foram desconsiderados na

classificação textual (isto é, nós que não pertencem ao conjunto-semente). O eixo horizontal representa a variação do parâmetro θ_2. 102

Figura 11: Gráfico com os dados de precisão e erro do classificador coletivo, acrescido da linha que informa a quantidade de nós efetivamente classificada para cada combinação de limiares. O eixo horizontal representa a variação do parâmetro θ_2. 103

Figura 12: Precisão (C) e erro (E) obtidos pela combinação do classificador textual com o classificador coletivo em todos os usuários. O eixo horizontal representa a variação do parâmetro θ_2. 105

Figura 13: Precisão da classificação coletiva de acordo com o conjunto inicial de nós selecionados de três formas.....106

Índice de Tabelas

Tabela 1: Dados coletados para os experimentos.....89

Tabela 2: Distribuição das classes nos conjuntos-semente.....96

Tabela 3: Resultados obtidos pelo classificador textual (marcação inicial dos nós) 97

Tabela 4: Taxas de acerto (C) e erro (E) obtidas pelo classificador textual na etapa de marcação inicial 99

Tabela 5: Precisão (C) e erro (E) obtidos pelo classificador coletivo sobre os nós não marcados pelo classificador textual 101

Tabela 6: Precisão (C) e erro (E) obtidos pela combinação do classificador textual com o classificador coletivo em todos os usuários 104

Sumário

1 Introdução 13

1.1 Contexto da Tese 14

1.1.1 Problema Geral 14

1.1.2 Problema Específico 16

1.1.3 Objetivo 17

1.2 Estudo de Caso 17

1.3 Hipótese 18

1.4 Contribuições 18

1.5 Conteúdo da Tese 19

2 Análise de Sentimento - Revisão da Literatura 21

2.1 Conceitos e Definições 21

2.1.1 Objeto 21

2.1.2 Trecho Opinativo sobre uma Característica 22

2.1.3 Explicitude de uma Característica 22

2.1.4 Autor 22

2.1.5 Opinião 22

2.1.6 Orientação da Opinião 22

2.1.7 Modelo de Objeto 22

2.1.8 Modelo de Documento Opinativo 23

2.1.9 Emoção 23

2.1.10 Subjetividade de sentenças 24

2.1.11 Explicitude de Opiniões 24

2.1.12 Sentença Opinativa 24

2.2 Etapas da Análise de Sentimento 25

2.2.1 Captura, Segmentação e Apresentação 25

2.2.2 Identificação de Atributos 27

2.2.3 Classificação 29

2.3 Abordagens para Classificação de Opiniões 31
2.3.1 Não supervisionadas 31
2.3.2 Semi-supervisionadas 32
2.3.3 Supervisionadas 33
2.4 Avaliação 33
2.5 Complicadores 34
2.5.1 Construções gramaticais complexas 34
2.5.2 Figuras de Linguagem 35
2.5.3 Resolução de co-referência 37
2.5.4 Tratamento de negações 38
2.5.5 Sentenças comparativas 38
2.5.6 Erros gramaticais ou de ortografia 38
2.5.7 Uso de gírias 39
2.5.8 Resolução de contexto 39
2.6 Classificação de Viés Opinativo de Usuário 40
2.7 Considerações Finais 41
3 Classificação Coletiva – Revisão da Literatura 43
3.1 Caracterização do Problema 44
3.2 Métodos de Inferência Coletiva 46
3.2.1 Algoritmos Baseados em Classificadores Locais 46
3.2.2 Algoritmos Baseados em Modelagens Globais 52
3.3 Partida a Frio 60
3.3.1 Classificação do Conjunto-Semente 61
3.3.2 Seleção do Conjunto-Semente 62
3.4 Trabalhos Relacionados 63
3.4.1 Uso de Informação Relacional entre Documentos ou Sentenças 64
3.4.2 Uso de Informação Relacional entre Características 64
3.4.3 Uso de Informação Relacional entre Pessoas 65
3.4.4 Técnicas de Mineração de Links 65

3.5 Considerações Finais 69
4 Classificação Coletiva Aplicada à Análise de Sentimentos em Redes Sociais 71
4.1 Crawler 74
4.1.1 *Pruning* do Grafo 75
4.2 Seleção do Conjunto-Semente 76
4.2.1 Seleção Baseada em Informações Estruturais do Grafo 77
4.2.2 Seleção Baseada em Classificação Textual 77
4.3 Classificação do Conjunto-Semente 78
4.3.1 Classificadores Disponíveis Online 79
4.3.2 Classificadores da Plataforma Weka 80
4.4 Classificação Coletiva 81
4.4.1 Análise de Custo Computacional 83
4.4.2 Escalabilidade do Método Proposto 84
4.5 Considerações Finais 85
5 Estudo de Caso 86
5.1 Base de Dados Usada nos Experimentos 86
5.1.1 Procedimento de Coleta de Dados 86
5.1.2 Utilização dos Dados em Experimentos 88
5.2 Experimento I – Seleção do Conjunto-Semente a partir de Classificação Textual das Postagens 94
5.3 Experimento II – Seleção do Conjunto-Semente através de Análise Estrutural do Grafo 105
5.4 Limitações 108
5.5 Considerações Finais 108
6 Conclusões e Trabalhos Futuros 110
6.1 Principais Contribuições 111
6.2 Aplicabilidade 112
6.3 Trabalhos Futuros 113

6.4 Divulgação de Resultados.....115
7 Bibliografia.....117

1 Introdução

A World Wide Web, inicialmente um mero repositório de informações estáticas, transformou-se numa enorme fonte de aplicações diversas, contendo não apenas informações, mas também proporcionando ou fomentando entretenimento, negócios, relacionamentos, entre outros. Essa variedade de aplicações provocou uma popularização e consequente explosão de uso da web, destacando-se entre os mais demandados os sites de comércio eletrônico (como eBay, Amazon, Submarino) e, mais recentemente, as redes sociais (como os já descontinuados Orkut e MySpace, além de Facebook, Google+ e Twitter).

Neste cenário, a quantidade de informação online disponível em relação a produtos e serviços é imensa: ao mesmo tempo em que tem acesso fácil a sites de compras, os usuários também podem apresentar *feedback* imediato, sob forma de formulários projetados e disponibilizados pelos próprios fornecedores. Isso é feito mais comumente, de forma espontânea, em sites de avaliação de produtos, que podem ser independentes (como o Epinion.com) ou integrados aos sistemas de compras, fóruns especializados ou sites de discussão. Frequentemente, até mesmo as redes sociais são usadas para emitir opiniões sobre produtos, serviços, personalidades, eventos e outros temas, uma vez que esses sites estão se tornando o ponto central de comunicação entre os usuários da web.

Assim, a web (ou, mais especificamente, os sites mencionados acima) contem informação altamente valiosa para empresas, entidades públicas, políticos, personalidades ou quem quer que tenha interesse em monitorar opiniões ou mensurar a aceitação de um objeto: ao invés de pesquisas de opinião usuais, que são demoradas, trabalhosas, tem alto custo e alcançam um número bastante restrito de pessoas, existe a possibilidade de coletar opiniões a partir das informações disponíveis on-line. Com o uso de ferramentas apropriadas, todo o processo de consolidação de opiniões pode ser realizado de forma automática, oferecendo um *feedback* imediato e mais representativo sobre o objeto avaliado. À área de pesquisa que estuda esses problemas se dá o nome de Análise de Sentimento (Pang & Lee, 2008). De forma resumida, Análise de Sentimento (ou Mineração de Opiniões) tem por objetivo classificar a opinião de uma pessoa acerca de um determinado tema, normalmente através

da aplicação de técnicas de processamento de linguagem natural sobre o texto escrito por aquela pessoa. Assim, Análise de Sentimento pode ser entendida como uma especialização da área de Classificação.

1.1 Contexto da Tese

1.1.1 Problema Geral

Dentro deste contexto, o problema geral desta tese é a classificação dos usuários de acordo com suas opiniões dada uma entidade ou objeto de interesse. Comumente, a classificação das opiniões emitidas por um dado usuário é feita através da classificação de sentimentos expressos em textos, postagens ou comentários. Isto é, são utilizados classificadores de texto para identificar a opinião dos usuários a partir dos textos ou comentários explícitos. Se a classificação de opiniões, no entanto, for realizada em ambientes nos quais haja conexões entre seus usuários (como as redes sociais), uma nova dimensão de informação se apresenta: através da análise dos relacionamentos e afinidades entre os usuários, é possível inferir as opiniões de usuários a partir da opinião de seus contatos (Bollen et al., 2011), (Yuan & Gay, 2006), (McPherson, Smith-Lovin, & Cook, 2001). Nesse caso, a classificação pode ser feita até mesmo para aqueles usuários que nem se manifestaram acerca de determinado assunto (Mustafaraj et al., 2011), o que aumenta de forma significativa o alcance da análise de sentimento. De forma semelhante, também é possível usar o resultado da análise de sentimento para os usuários que se manifestaram acerca de determinado tema combinado com a análise dos relacionamentos da rede social, de forma a modificar a orientação ou reforçar a confiança da classificação de opiniões.

Abordagens clássicas baseadas em classificadores de texto para o problema acima são bastante difundidas na literatura (Abbasi, Chen, & Salem, 2008), (Wilson, Wiebe, & Hwa, 2004), (Hu & Liu, 2004), (Liu, 2006), (Popescu & Etzioni, 2005), (Su et al., 2008), (Titov & Mcdonald, 2008), (Ding, Liu, & Yu, 2008), (Morinaga et al., 2002), (Pak & Paroubek, 2010), porém a observação do problema sob o prisma aqui apresentado é pouco frequente. Como esses trabalhos se baseiam exclusivamente na análise textual do conteúdo a ser classificado, estão sujeitos a desafios difíceis de contornar (como ironia, sarcasmo, resolução de contexto, entre outros), aos quais o

método aqui proposto é insensível. Um trabalho com abordagem semelhante, que procura usar a informação relacional disponível em redes sociais (Guerra, 2013) usa os relacionamentos entre os usuários para a criação de um classificador textual a partir das postagens dos usuários identificados. Porém, o autor tem a classificação coletiva[1] como passo intermediário do processo, e assim não tem seu foco nessa etapa. Em outro trabalho (Conover et al., 2011), são usadas algumas técnicas baseadas em conteúdo e também uma técnica que faz uso da informação relacional entre os usuários da rede. Há ainda pesquisas (Wang et al., 2011) que usam classificação coletiva para inferir o sentimento associado a *hashtags* de usuários do Twitter. Aqui, cada nó do grafo corresponde a uma *hashtag* associada a um dado tópico e cada aresta representa uma relação de coocorrência.

A abordagem proposta neste trabalho para realização de análise de sentimento em redes sociais é baseada no princípio da homofilia (Macskassy & Provost, 2007), que estabelece que indivíduos tendem a se conectar a outros com os quais apresentam alto grau de semelhança. A partir desse conceito, são aplicadas técnicas de classificação coletiva (Rattingan, Maier, & Jensen, 2007), (Sen et al., 2008) sobre o grafo que representa a rede social. A intenção é explorar o fato de que a classificação coletiva não utiliza individualmente as características locais dos nós no processo de inferência, mas também as características dos nós relacionados. Além disso, a classificação é executada de forma simultânea sobre todas as instâncias, o que permite considerar as influências que cada instância exerce sobre outras às quais está relacionada.

Para que seja possível a aplicação de classificação coletiva, o grafo de usuários precisa ser parcialmente classificado, isto é, precisa dispor de um conjunto-semente de usuários cujas classificações são conhecidas. Esse conjunto-semente pode ser construído de várias formas, inclusive manualmente. A definição de um conjunto-semente pode ser entendida como um tipo de partida a frio (*cold start*), problema bastante comum em outros contextos, como nos sistemas de recomendação (Lam et al., 2008), (Park et al., 2006). O

[1] Processo de classificação que usa informações (atributos e classes) dos nós e de suas vizinhanças para determinar simultaneamente as classes de um conjunto de nós interconectados (Sen et al., 2008).

problema da partida a frio ocorre quando um sistema depende de um volume suficiente de informação coletada a priori (normalmente via intervenção manual). No contexto de classificação coletiva de opiniões, esse problema se reflete na necessidade de conhecimento prévio das opiniões de parte dos usuários da rede a respeito do tópico sob monitoramento.

1.1.2 Problema Específico

Dado o contexto acima, o problema específico abordado nesta tese é a classificação de usuários (de acordo com suas opiniões) através de classificação coletiva, lidando com a dificuldade da geração do conjunto inicial de usuários etiquetados (partida a frio). Para tratar esse problema, esta proposta apresenta duas abordagens que combinam classificadores textuais e inferência ativa (Rattingan, Maier, & Jensen, 2007), de forma a minimizar a quantidade necessária de nós iniciais que constituem o conjunto-semente de usuários etiquetados. Na primeira, classificadores textuais são usados para produzir de forma automática um conjunto inicial de nós classificados a partir do conteúdo das mensagens postadas nas redes sociais acerca de um determinado tema sob monitoramento. Na segunda, técnicas de inferência ativa são propostas para identificar os nós mais relevantes a serem etiquetados usando medidas de centralidade ou baseadas em funções de utilidade dos nós. Destacamos que nenhum outro trabalho na literatura foi encontrado usando inferência ativa no contexto de classificação de opiniões.

Como mostram os resultados já divulgados (Rabelo, Prudêncio, & Barros, 2012a), (Rabelo, Prudêncio, & Barros, 2012b), (Rabelo, Prudêncio, & Barros, 2012c), esta abordagem permite realizar a análise de sentimento em redes sociais com maior precisão do que as técnicas que dependem exclusivamente da análise textual, pois o método proposto não é sensível às dificuldades inerentes à abordagem de classificação textual, apontadas anteriormente, uma vez que não depende de processamento do texto emitido pelos usuários. Além disso, pela análise de relacionamento entre os nós de uma rede, é possível classificar a opinião de usuários que nem se manifestaram explicitamente sobre um determinado tema (o que amplia a cobertura da análise, pois há uma parcela significativa de usuários que consomem informação mas não produzem conteúdo em volume suficiente para a aplicação de análise textual (Mustafaraj et al., 2011)). A análise dos relacionamentos

entre os usuários permite ainda reforçar a confiança de classificações realizadas previamente (por exemplo, através de classificadores textuais) ou até modificar a classificação de usuários que postaram opiniões em redes sociais. De forma semelhante, seguindo a ideia de outras pesquisas na área (Guerra, 2013), podem ser construídos classificadores baseados em conteúdo para melhorar a classificação textual.

1.1.3 Objetivo

O objetivo desta tese é a proposição de um método de classificação de usuários em redes sociais através de uma abordagem cujo principal foco é a aplicação de algoritmos de classificação coletiva sobre um conjunto-semente de nós com classes previamente conhecidas e obtidas através de qualquer método (por exemplo, processamento textual das postagens ou marcação manual).

1.2 Estudo de Caso

Para avaliação da viabilidade do método proposto, foi implementado um protótipo que usa um algoritmo de *relaxation labeling* (Chakrabarti et al., 1998) para a classificação coletiva de opiniões de usuários em redes sociais, e foi desenvolvido um estudo de caso para predição de preferência política de usuários do Twitter, que alcançou resultados promissores. O experimento consistiu na coleta de 9098 postagens contendo *hashtags* relacionadas a política americana de 4719 usuários diferentes e respectivas informações relacionais (relacionamentos de seguidor e seguido) dos usuários, o que gerou um grafo de cerca de 97 mil nós e 1 milhão de arestas (posteriormente podado para um grafo de 1.244 nós e 78.272 arestas). As citações foram submetidas a um classificador textual, sendo a classe de cada usuário determinada com base num sistema simples de votação.

Sobre esse corpus, foram realizados dois experimentos: no primeiro, através da aplicação de conceitos de inferência ativa (Rattingan, Maier, & Jensen, 2007), (Bilgic & Getoor, 2010), foram determinados diferentes limiares para dois parâmetros: o número mínimo de postagens dos usuários e o score de confiança do classificador textual. Variando esses limiares foram gerados diferentes conjuntos-semente de nós iniciais, a partir dos quais foi avaliada a precisão geral do sistema formado pelo classificador textual e o classificador coletivo. No segundo experimento, foram aplicados limiares sobre

características estruturais do grafo (grau dos nós, centralidade e *betweeness*) para gerar os conjuntos-semente, e medida a precisão da classificação coletiva. Uma das possíveis extensões deste trabalho é a combinação das abordagens aqui avaliadas.

1.3 Hipótese

À luz do que foi exposto nesta seção, a hipótese desta tese é que o uso adequado da informação relacional (frequentemente indisponível ou negligenciada no âmbito de análise de sentimentos) presente em redes sociais permite a classificação dos usuários de acordo com suas opiniões com uma maior precisão do que abordagens tradicionais baseadas apenas em processamento textual. Esta hipótese foi testada em experimentos executados sobre um domínio específico (isto é, opiniões sobre política americana postadas por usuários do Twitter). Os resultados obtidos nos experimentos corroboram com a hipótese de tese, o que se pode considerar para cenários que apresentem condições semelhantes àquele aqui avaliado. A principal condição de semelhança para que a hipótese se verifique em outros cenários diz respeito à assortatividade do grafo que representa a rede de usuários: aqueles que apresentam coeficientes de assortatividade próximos a 1 possuem características adequadas para a aplicação do método aqui proposto (nos experimentos realizados, o grafo possui coeficiente de 0,72 – consulte a seção 5.1.2).

1.4 Contribuições

As principais contribuições deste trabalho são:

- Método para aplicação de algoritmos de classificação textual, classificação coletiva e de princípios de inferência ativa para o problema de análise de sentimento em redes sociais;
- Framework de análise de sentimento para redes sociais (validado em experimentos de política, mas que pode ser aplicado a diferentes domínios), cujo principal componente de classificação não está sujeito aos problemas intrínsecos à análise textual;

- Algoritmo de descarte seletivo de nós em grafos, com potencial para aplicação em qualquer cenário que envolva manipulação de grafos muito grandes.

1.5 Conteúdo da Tese

O restante do documento está organizado como se segue.

- Capítulo 2 - Análise de Sentimento - Revisão da Literatura: neste capítulo, é apresentada uma revisão da literatura da área de análise de sentimentos. Foi realizada uma varredura sobre métodos tradicionais de análise de sentimento, que aplicam algoritmos de classificação de texto para inferência sobre sua orientação de opinião. O Capítulo 2 traz ainda uma definição do problema, os principais complicadores e uma arquitetura genérica para sistemas que aplicam o método de análise de sentimento baseado em análise de texto;

- Capítulo 3 - Classificação Coletiva – Revisão da Literatura: este capítulo faz uma revisão da literatura da área de Classificação Coletiva, caracterizando e definindo formalmente o problema e apresentando em detalhes os principais algoritmos desenvolvidos. Como os algoritmos de Classificação Coletiva necessitam de um conjunto inicial de nós com classificação conhecida, o capítulo 3 apresenta também uma avaliação sobre o problema da Partida a Frio (*cold start*), que acontece exatamente em cenários em que alguma informação *a priori* é necessária para a execução do processo como um todo. Por fim, este capítulo traz uma revisão mais resumida da área de Mineração de Links, super-área na qual a Classificação Coletiva está enquadrada e que estuda problemas e técnicas similares, possivelmente aplicáveis ao problema abordado nesta tese;

- Capítulo 4 - Classificação Coletiva Aplicada à Análise de Sentimentos em Redes Sociais: este capítulo detalha a abordagem desenvolvida neste trabalho, que combina a aplicação de algoritmos de Classificação Coletiva aplicada ao problema de análise de sentimentos com alternativas para o tratamento da Partida a Frio, considerando inclusive técnicas de Inferência Ativa para seleção do conjunto-semente;

- Capítulo 5 - Estudo de Caso: neste capítulo, são apresentados detalhes sobre um estudo de caso e os experimentos desenvolvidos durante este trabalho, desde a montagem da base até a análise dos resultados alcançados;

- Capítulo 6 - Conclusões e Trabalhos Futuros: este capítulo traz as conclusões acerca do trabalho desenvolvido, destaca as principais contribuições, apresenta resultados alcançados durante o desenvolvimento e lista possibilidades de extensões futuras.

2 Análise de Sentimento - Revisão da Literatura

2.1 Conceitos e Definições

De forma resumida, análise de sentimentos é o tratamento computacional de opiniões, sentimentos e emoções expressas em linguagem natural (Abbasi, Chen, & Salem, 2008). É uma área de pesquisa relativamente nova, que está intimamente ligada a áreas mais tradicionais como processamento de linguagem natural (Jurafsky & Martin, 2008), (Indurkhya & Damerau, 2010) e mineração de textos (Feldman & Sanger, 2006), (Kao & Poteet, 2007), aplicando ferramentas e técnicas desenvolvidas por essas áreas, ou adequando e criando novas técnicas (Balahur, Mihalcea, & Montoyo, 2014). Em relação à nomenclatura, vê-se na literatura uma variedade de termos usados para se referir à análise de sentimentos: análise de subjetividade (*subjectivity analysis*), mineração de opiniões (*opinion mining*) ou extração de apreciações (*appraisal extraction*).

No restante da seção 2.1 serão apresentados conceitos e definições relacionados à área de análise de sentimento (Liu, 2010) que serão usados nas seções e capítulos seguintes.

2.1.1 Objeto

Um objeto *o* é uma entidade que pode ser um produto, pessoa, evento, empresa ou tópico. Cada objeto está associado a um par, *o: (T, A)*, onde *T* é uma hierarquia de componentes (ou partes) e subcomponentes aninhados, e *A* é um conjunto de atributos de *o*. Cada componente possui seu próprio conjunto de subcomponentes e atributos.

Um objeto pode ser representado como uma árvore de características, componentes ou atributos, mas também através de uma estrutura plana. Nesse caso, o termo "característica" é usado para representar componentes e atributos. Com essa simplificação, o objeto pode ser visto como uma característica especial, e uma opinião sobre ele é chamada de "opinião geral" sobre o objeto (por exemplo, "Eu gosto da CameraX"). Opiniões sobre características específicas são chamadas de "opiniões específicas" (por exemplo, "O display da CameraX é ótimo").

2.1.2 Trecho Opinativo sobre uma Característica

Seja um documento opinativo *d*, consistindo em uma sequência de sentenças $d = <s_1, s_2, ..., s_m>$. Um trecho opinativo sobre uma característica *c* de um objeto *o* avaliado em *d* é um grupo de sentenças em *d* que expressam opiniões sobre *c*.

2.1.3 Explicitude de uma Característica

Se uma característica *c* for citada textualmente numa sentença *s*, *c* é chamada de característica explícita em *s*. Caso não seja citada, mas esteja implícita, *c* é chamada de característica implícita em *s*. Um exemplo de característica implícita é "tamanho" na sentença "Essa câmera é muito grande", em que "grande" (que não é sinônimo de "tamanho") é um indicador de característica.

2.1.4 Autor

O autor é a entidade que expressa a opinião. Em websites de notícias, por exemplo, a empresa pode ser considerada a autora. Para blogs pessoais, redes sociais ou nas seções para comentários/avaliações em websites, o autor é a pessoa que escreveu o texto.

2.1.5 Opinião

Uma opinião sobre uma característica *c* é uma visão, atitude, emoção ou avaliação positiva ou negativa sobre *c* emitida pelo autor da opinião.

2.1.6 Orientação da Opinião

A orientação (também chamada de polaridade) de uma opinião acerca de uma característica *c* indica se a opinião é negativa, positiva ou neutra. Há trabalhos que aplicam uma escala mais detalhada (Wilson, Wiebe, & Hwa, Just how mad are you? Finding strong and weak opinion clauses, 2004).

2.1.7 Modelo de Objeto

Um objeto *o* é representado por um conjunto finito de características, $C = \{c_1, c_2, ..., c_m\}$, que inclui o próprio objeto como uma característica especial. Cada característica $c_i \in C$ pode ser expressa com qualquer elemento do conjunto finito de palavras ou frases $W_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$, que são sinônimos da característica, ou indicada por qualquer elemento do conjunto finito de indicadores de características $I_i = \{i_{i1}, i_{i2}, ..., i_{iq}\}$.

2.1.8 Modelo de Documento Opinativo

Um documento opinativo d contém opiniões sobre um conjunto de objetos $\{o_1, o_2, ..., o_q\}$ de um conjunto de autores $\{a_1, a_2, ..., a_p\}$. As opiniões sobre cada objeto o_j são expressas num subconjunto C_j de características de o_j. Uma opinião direta é uma quíntupla $\{o_j, c_{jk}, oo_{ijkl}, a_i, t_l\}$, onde o_j é um objeto, c_{jk} é uma característica do objeto o_j, oo_{ijkl} é a orientação da opinião da característica c_{jk} do objeto o_j, a_i é o autor da opinião e t_l é horário em que a opinião foi emitida por a_i. Para uma opinião sobre a característica c_{jk}, a_i usa uma palavra ou frase do conjunto de sinônimos W_{jk} ou de indicadores I_{jk}, e expressa uma opinião positiva, negativa ou neutra sobre c_{jk}.

Antes de introduzir o próximo conceito, consideremos a intensidade de uma opinião. Há opiniões com forte intensidade (como em "Essa câmera é horrível") e outras com intensidade mais baixa (como em "Essa câmera é ruim"). Dessa forma, a intensidade das opiniões pode ser interpretada e posta em escala (Martin & White, 2005). Por exemplo, opiniões positivas podem ser classificadas como boas, muito boas, excelentes e fantásticas[2]. Essa discussão se confunde com a definição de "emoção", apresentada a seguir.

2.1.9 Emoção

É um pensamento ou sentimento subjetivo. As emoções são estudadas em diferentes áreas, como psicologia, sociologia, filosofia, etc. Ainda assim, não se chegou a um consenso entre os pesquisadores quanto a um conjunto de emoções. Uma proposta bem aceita (Parrott, 2000) elenca seis tipos de emoções primárias (amor, alegria, surpresa, raiva, tristeza e medo), que podem ser subdivididas em diversas emoções secundárias.

É importante distinguir dois conceitos diferentes: o estado mental (ou sentimento) de uma pessoa, e a verbalização desse sentimento através de expressões linguísticas. Embora haja apenas seis diferentes tipos de emoções, há um grande número de formas de expressá-las através da linguagem. Essencialmente, o objetivo da análise de sentimento é inferir os sentimentos das pessoas com base nas expressões linguísticas. Além disso, no modelo aqui apresentado, é necessário também encontrar outras informações, que são

[2] Pode-se determinar a granularidade da escala de intensidades de acordo com a necessidade.

importantes para aplicações práticas, como data da opinião, autor, etc (consulte a seção 2.2.2).

2.1.10 Subjetividade de sentenças

Uma sentença objetiva expressa alguma informação factual, enquanto uma sentença subjetiva expressa algum sentimento ou crença pessoal. Sentenças subjetivas podem não conter uma opinião, como em: "Eu gostaria de uma câmera com alta resolução". Neste exemplo, embora não haja opinião explícita, implicitamente é bem provável que a opinião seja positiva acerca da câmera citada. Analogamente, também há sentenças objetivas que carregam opinião implícita. Por exemplo: "O fone de ouvido quebrou depois de dois dias!". Embora essa sentença estabeleça um fato concreto, ela implicitamente carrega uma opinião negativa acerca do objeto "fone de ouvido".

2.1.11 Explicitude de Opiniões

Uma opinião explícita da característica *c* é aquela explicitamente expressa numa sentença subjetiva. Uma opinião implícita de *c* é aquela que está implícita numa sentença objetiva[3].

2.1.12 Sentença Opinativa

É uma sentença (subjetiva ou objetiva) que expressa opiniões, seja explícita ou implicitamente.

À luz do que foi exposto até aqui, pode-se definir o objetivo da análise de sentimento da seguinte forma: dado um documento opinativo *d*, é necessário inicialmente descobrir todas as quíntuplas (o_j, c_{jk}, oo_{ijkl}, a_i, t_l) em *d*, e também identificar todos os sinônimos W_{jk} e indicadores de características I_{jk} de cada característica c_{jk} em *d*.

A partir desse modelo, vê-se que a tarefa da análise de sentimento é bastante complexa: além de identificar as informações no texto, é necessário ainda que elas sejam relacionadas de forma correta. A seguir, são apresentadas as tarefas envolvidas na análise de sentimento.

[3] Em geral, sentenças objetivas que contêm opiniões implícitas descrevem as razões porque os autores têm aquelas opiniões.

2.2 Etapas da Análise de Sentimento

Esta seção descreve as principais técnicas e métodos atualmente usados na área de análise de sentimento. A fim de guiar a apresentação, é mostrada, na Figura 1, a arquitetura usual para sistemas de análise de sentimento, identificada a partir da análise realizada sobre a literatura da área. A arquitetura proposta usa o modelo descrito na seção 2.1.

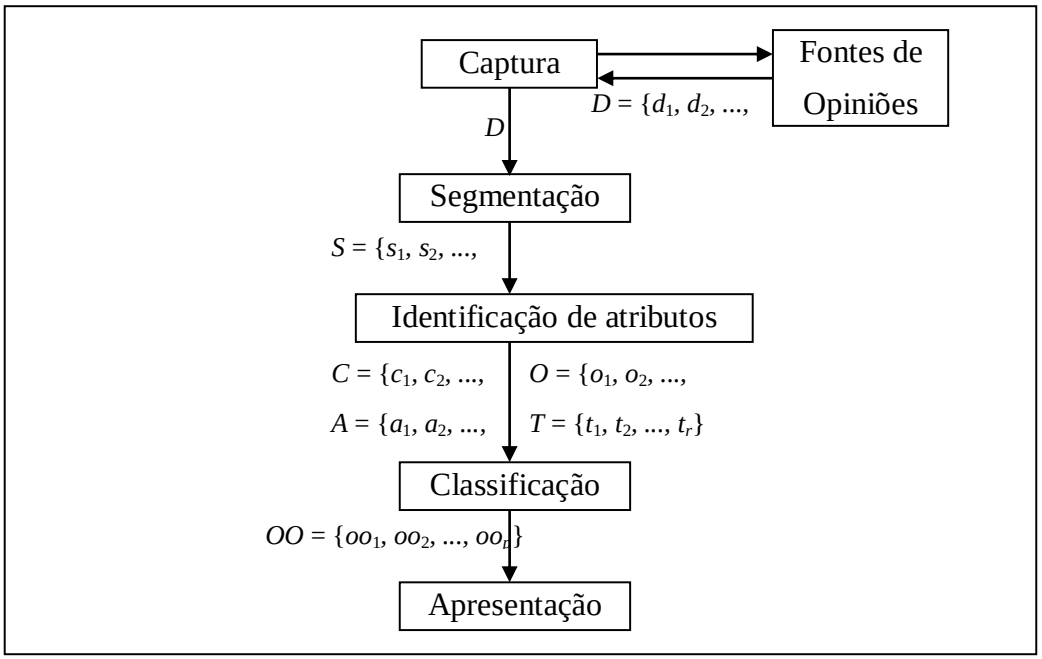

Figura 1: Arquitetura usual para sistemas de Análise de Sentimento

As etapas de captura de dados e segmentação serão apresentadas brevemente neste trabalho (seção 2.2.1), uma vez que não apresentam complexidade relevante na área. A etapa de apresentação dos resultados, embora possa envolver técnicas complexas para visualização, está mais relacionada à área de interface homem-máquina do que com a área de inteligência artificial, e por este motivo também será descrita superficialmente na seção 2.2.1. Aqui, será dedicada particular atenção às etapas de identificação de atributos (seção 2.2.2) e classificação (seção 2.2.3).

2.2.1 Captura, Segmentação e Apresentação

A captura de dados consiste na obtenção dos documentos que serão processados pelas fases subsequentes. Esta etapa pode ser tão simples quanto a consulta em

um banco de dados ou a leitura de um sistema de arquivos. Normalmente, entretanto, o *crawler* (Baeza-Yates & Ribeiro-Neto, 1999), componente responsável por executar esta fase, coleta dados das fontes de informação (por exemplo: redes sociais, sites de notícia, fóruns, etc), armazenando-os para processamento posterior. Isto pode ser feito mediante a aplicação de uma API[4] de acesso à base de informações (quando disponível), mas de forma mais genérica pode ser necessário implementar todo o mecanismo de acesso à base. Isto é comum para processamento, por exemplo, de sites de notícia e blogs. Neste cenário, ferramentas como o Apache Nutch (Laliwala & Shaikh, 2013) podem ser aplicadas para simplificar a implementação.

O objetivo da segmentação é quebrar o documento em unidades menores, que serão processadas pelas fases subsequentes. Em muitos casos, entretanto, não é necessário realizar este procedimento; assume-se que o documento fala sobre uma única característica ou objeto e que contem opiniões de um único autor, o que obviamente em muitos casos não acontece. Ou seja, não é realizada qualquer segmentação, e o documento inteiro é usado como sendo o texto que será avaliado. Esta abordagem pode ser adequada para casos específicos como, por exemplo, a classificação de sentimentos no Twitter, rede social que possui uma limitação de 140 caracteres por postagem, e que por isso tende a obedecer às restrições deste modelo.

Normalmente, entretanto, a segmentação envolve um pré-processamento simples do documento através da sua subdivisão em sentenças, e cada sentença é posteriormente classificada de acordo com o sentimento nela expresso. Antes da classificação de sentimento, entretanto, normalmente se faz uma análise para eliminar sentenças não opinativas (Hatzivassiloglou & Wiebe, 2000), (Wiebe et al., 2004), (Wiebe & Mihalcea, 2006), (Wilson, Wiebe, & Hwa, 2004), (Wilson, Wiebe, & Hoffmann, 2005). Na segmentação por sentença, assume-se que cada sentença é expressa por um mesmo autor (o que ocorre na maioria das vezes) e contem opinião sobre um único objeto ou característica (o que nem sempre ocorre). Ao fim, normalmente o sentimento geral do documento é

[4] As redes sociais normalmente disponibilizam APIs para acesso ao seu conteúdo de forma programática (e controlada). Alguns exemplos são o Twitter (https://dev.twitter.com/overview/api) e o Facebook (https://developers.facebook.com/).

definido através de alguma forma de agregação dos resultados individuais das sentenças (por exemplo, votação) (Wilson, Wiebe, & Hoffmann, 2005), (Meena & Prabhakar, 2007), (Pang & Lee, 2004).

Após a coleta e segmentação de documentos, são realizadas as etapas de identificação de atributos e classificação, que serão apresentadas em mais detalhes nas seções 2.2.2 e 2.2.3. A última etapa é a apresentação dos resultados ao usuário final. Em aplicações que envolvem mineração de opiniões, normalmente é importante estudar uma coleção de opiniões, e não opiniões isoladas (pois essas representam a visão de uma pessoa e geralmente não são relevantes para tomadas de decisão). Diferentemente do problema de recuperação de informação convencional, no qual a coleta de um documento com a informação procurada satisfaz a necessidade do usuário, na análise de sentimento o ideal é que se colete e avalie o maior número possível de opiniões acerca do objeto de interesse (Liu, 2010). Isso indica que algum tipo de sumário dos resultados minerados precisa ser criado, já que seria de pouca utilidade para o usuário receber todas as quíntuplas previamente identificadas. Normalmente, a forma mais eficaz de fazer isso é através de gráficos que estão ligados às sentenças e documentos que foram minerados para gerar as quíntuplas, de forma que o usuário recebe inicialmente o resumo, mas pode também verificar os resultados individualmente.

2.2.2 Identificação de Atributos

O objetivo dessa fase é a identificação de informações como autor, horário e objeto (ou característica) da opinião. Esse tipo de tarefa[5] é conhecido como *Named Entity Recognition* e já foi largamente estudado pela área de Extração de Informação (Sarawagi, 2008). As pesquisas atuais nessa etapa são altamente dependentes do formato do texto. Abaixo, serão descritas técnicas para a extração de informação a partir do formato mais genérico possível: texto livre inserido por usuários em sites de avaliação de produtos (exemplo 3).

[5] Às vezes, não é necessário processar o documento para identificar todas essas informações. Por exemplo, em websites de comentários sobre produtos, normalmente a data/hora e o autor de cada comentário são exibidos, e a identificação dessas informações passa a ser um problema de extração estruturada (Liu, 2006).

Excelente câmera! (em 5/10/2009)

Por: jbpires

Eu pesquisei bastante antes de comprar essa câmera... Vai ser difícil abandonar minha Nikon 35mm SLR, mas como vou viajar à Europa, preciso de uma câmera menor (e digital). As fotos dessa câmera são impressionantes. A funcionalidade 'auto' produz ótimas fotos na maioria das vezes. E, sendo digital, você não gasta filme se não gostar das fotos. (...)

Exemplo 1: Comentário fictício em texto livre sobre uma câmera fotográfica.

Para esse problema, as técnicas mais usadas são aquelas baseadas em aprendizagem não-supervisionada. Uma das propostas (Hu & Liu, 2004) é um método para a extração de características explícitas que consiste em dois passos:

(1) Encontrar substantivos e sintagmas nominais (que normalmente são usados para expressar as características dos produtos) através de um POS-tagger[6] e manter apenas os mais frequentes[7] (pois o vocabulário sobre produtos tende a convergir, e assim as entradas raras podem ser descartadas sem prejuízo ao desempenho do algoritmo);

(2) Encontrar características infrequentes através de palavras opinativas (normalmente, adjetivos ou advérbios que expressam opinião). A ideia aqui é que a mesma palavra opinativa pode ser usada para descrever características diferentes, e assim ela pode ser usada para identificar características infrequentes. Por exemplo, "zoom" é identificado como uma característica frequente no passo (1).

[6] Informações de classes morfológicas, providas por POS-taggers, são normalmente usadas em análise de sentimento porque são consideradas uma forma pouco elaborada de desambiguação de significado de palavras (Wilks & Stevenson, 1998).

[7] Um limiar pode ser definido experimentalmente.

> Ao processar a sentença “O zoom é ótimo”, e sabendo que “ótimo” é uma palavra opinativa positiva, pode-se identificar que “usabilidade” é outra característica da câmera a partir da sentença “A usabilidade é ótima”.

Também há propostas (Popescu & Etzioni, 2005) para aumentar a precisão do passo (1) do algoritmo anterior, através da remoção de sintagmas nominais que não são características de produto. O algoritmo consiste em calcular a pontuação de informação mútua ponto-a-ponto (*pointwise mutual information* - PMI) de cada sintagma nominal com discriminantes de meronímia[8]. A intenção é identificar, através de metabuscas na web, se um dado sintagma nominal ocorre simultaneamente com os discriminantes em documentos com frequência. Se isso não acontecer, aquele sintagma será descartado. Há também trabalhos que aplicam *clustering* para identificar características (Su et al., 2008), (Titov & Mcdonald, 2008).

2.2.3 Classificação

Esta etapa do processamento é responsável por definir a orientação das opiniões sobre a característica expressa numa sentença. A seguir, serão apresentadas as principais abordagens para realizar a tarefa de classificação de opiniões.

2.2.3.1 Características Usadas para Classificação

Sintáticas

Consistem no uso de regras de construção de sentenças em linguagem natural. As principais características sintáticas utilizadas são:

Classe Morfológica (Part of speech)

O uso da classe morfológica é bastante comum na área de análise de sentimento. Um refinamento comum do uso de classes morfológicas é o uso de padrões morfológicos, que são regras consistindo de sequências de classes morfológicas (por exemplo, substantivos seguidos de adjetivo). Diversos trabalhos na literatura adotam esse tipo de informação (Wilson, Wiebe, &

[8] Relações do tipo “parte de”.

Hoffmann, 2005), (Whitelaw, Garg, & Argamon, 2005), (Pak & Paroubek, 2010) e (Mullen & Collier, 2004).

Pontuação

Informações de pontuação são usadas basicamente para fazer a segmentação de texto, porém às vezes são usadas para auxiliar no processo de classificação (principalmente sinais específicos como exclamações). Um dos trabalhos (Mcdonald et al., 2007) inclui pontos de exclamação e de interrogação entre as características consideradas para classificação de orientação de opiniões.

Modificadores

Uso de modificadores como "muito", "não", "especialmente", que servem basicamente para ênfase ou inversão de orientação (Hogenboom et al., 2011).

Semânticas

Esse tipo de abordagem consiste em considerar o significado semântico das palavras, o que pode ser feito de forma manual, automática ou semi-automática. Normalmente fazem uso de ferramentas como SentiWordnet (Esuli & Sebastiani, 2006).

Baseadas em relacionamentos

Neste caso, aplica-se análise de links e citações para determinar o sentimento de documentos. Alguns trabalhos (Efron, 2004) mostram que páginas da web com opiniões tendem a referenciar outras páginas que possuem sentimentos similares.

Estilísticas

Consiste no uso de características estilísticas no processo de classificação de sentimento, como distribuição de comprimento de palavras, medidas sobre a riqueza/diversidade do vocabulário e frequência de uso de caracteres especiais (Abbasi, Chen, & Salem, 2008).

2.3 Abordagens para Classificação de Opiniões

2.3.1 Não supervisionadas

Esta etapa do processamento é responsável por definir a orientação das opiniões sobre a característica expressa numa sentença. Grande parte dos trabalhos aqui se baseia em léxicos (Ding, Liu, & Yu, 2008), (Hu & Liu, 2004).

A abordagem não supervisionada consiste em usar palavras e frases opinativas para determinar a orientação da opinião, além de tratar negações e cláusulas adversativas nas sentenças. Para ilustrar seu funcionamento, considere a sentença: "Esse carro não é tão confortável, mas tem grande autonomia". Um algoritmo baseado numa abordagem não supervisionada executa os seguintes passos:

1) **Identificação de palavras e frases opinativas**: de acordo com o léxico previamente montado, são identificados os termos opinativos e é atribuído o *score* -1 para termos negativos, +1 para termos positivos e 0 para termos dependentes de contexto. Assim, no exemplo acima teríamos "Esse carro não é tão *confortável*[+1], mas tem *grande*[0] autonomia".

2) **Tratamento de negações:** partículas negativas são usadas para inverter os *scores* atribuídos no passo (1). Assim, o resultado ficaria: "Esse carro não é tão *confortável*[-1], mas tem *grande*[0] autonomia".

3) **Cláusulas adversativas:** uma sentença iniciada por conjunção adversativa tem orientação inversa à anterior. Dessa forma, o resultado seria: "Esse carro não é tão *confortável*[-1], mas tem *grande*[+1] autonomia", identificando que "grande", nesse contexto, é uma qualidade.

4) **Agregação de opiniões:** por fim, uma função soma os *scores* das opiniões e calcula a orientação final da opinião sobre cada característica na sentença.

Há outras abordagens não supervisionadas menos comuns, como um exemplo (Zhang et al., 2013) que propõe um modelo hierárquico baseado em naive bayes para classificação em nível de sentença e de documento de forma simultânea.

2.3.2 Semi-supervisionadas

Conforme se pode perceber, o algoritmo apresentado na seção anterior é basicamente uma aplicação de regras conceituais suportadas por um léxico. Muitas outras regras são necessárias para cobrir uma porção de fato significativa do universo de opiniões num dado domínio, e ajustes seriam necessários para aplicação em domínios diferentes. Por esse motivo, há abordagens semiautomáticas para a obtenção de regras (Popescu & Etzioni, 2005), (Morinaga et al., 2002).

Há trabalhos (Sindhwani & Melville, 2008) cuja abordagem combina o uso de léxicos com polaridade previamente marcada com exemplos não etiquetados através de uma representação dos dados através de grafos bipartites. Outro trabalho (Goldberg & Zhu, 2006) aplica uma abordagem semi-supervisionada baseada em grafos para o problema de análise de sentimento (com foco em opiniões sobre filmes). Aqui, o grafo é criado sobre exemplos etiquetados (numa escala de 1 a 5) e não etiquetados, e é resolvido um problema de otimização que produz uma função de aproximação sobre o grafo completo para novos exemplos. Os resultados apresentados mostram que o desempenho do sistema é consideravelmente melhor que outras abordagens que não usam os exemplos não etiquetados, especialmente quando há poucos exemplos etiquetados disponíveis.

Há ainda esforços (Rao & Ravichandran, 2009) para determinação da polaridade de palavras de forma semi-automática, problema que é modelado através de propagação de labels num grafo em que cada nó representa uma palavra cuja polaridade deve ser determinada. As arestas do grafo tem pesos associados e representam relações (sinonímia e hipernímia) entre as duas palavras conectadas. Usando um conjunto-semente, o algoritmo consegue propagar os labels (positivo ou negativo) pelo grafo. Este modelo possui

semelhanças com o modelo de classificação proposto neste trabalho (consulte o capítulo 4), porém modela o texto de um documento como um grafo, e não os autores.

2.3.3 Supervisionadas

As abordagens supervisionadas assumem que há exemplos de treinamento disponíveis para cada classe (normalmente, negativo, positivo e neutro, mas também pode haver uma escala discreta de valores). A partir dos dados de treinamento, o sistema aprende um modelo de classificação a partir da aplicação de algoritmos como SVM, Naive Bayes, KNN, Regressão Logística, etc.

Um exemplo de aplicação de método supervisionado para classificação de textos curtos (Hu et al., 2013) realiza uma modelagem de otimização matemática que incorpora as teorias de consistência de sentimento e contágio emocional no processo de aprendizagem.

Em (Balahura & Turchib, 2013), é apresentada uma abordagem para classificação supervisionada em idiomas diferentes de inglês (os experimentos foram desenvolvidos sobre alemão, francês e espanhol) através do uso de tradução automática de sistemas como Google Translate, Bing Translator e Moses.

2.4 Avaliação

A avaliação de desempenho de sistemas de análise de sentimento é normalmente feita através de uma medida do quanto o seu resultado coincide com o de análises manuais, usando os conceitos de precisão e cobertura emprestados da área de Recuperação de Informações (Baeza-Yates & Ribeiro-Neto, 1999). Entretanto, alguns trabalhos mostram que análises manuais feitas por pessoas diferentes coincidem em apenas 79% dos casos (Ogneva, 2010), o que demonstra a dificuldade de realizar de forma satisfatória uma tarefa com tal grau de subjetividade associada, mesmo quando desempenhada por operadores humanos. Dessa forma, um sistema ideal que esteja correto em 100% dos casos, quando comparado com análises manuais, apresentaria uma taxa de erro calculada em aproximadamente 21%. Para sistemas de análise de sentimento que retornam uma escala de valores ao invés de um resultado binário, medir o

desempenho através de correlação pode ser mais indicado porque dessa forma se leva em consideração quão próximo do valor esperado está o resultado.

2.5 Complicadores

Não se pode questionar o valor que a análise automática de sentimentos pode trazer a empresas e outras entidades interessadas em monitorar tópicos de interesse em tempo real. Porém, algumas limitações da tecnologia atual requerem atenção especial quando da avaliação do resultado de análises realizadas automaticamente.

A classificação automática de sentimento de um texto em relação a um determinado objeto de interesse é uma tarefa relativamente simples em casos controlados. Por exemplo, se o texto em questão tiver uma estrutura semelhante a: “A CameraX é excelente!”, um algoritmo pouco sofisticado pode produzir os resultados esperados. Entretanto, esse não é o caso em aplicações reais; há algumas dificuldades inerentes ao processo de análise automática de texto expresso em linguagem natural que podem limitar o uso e a aplicabilidade de métodos de análise de sentimento. As principais dificuldades são brevemente apresentadas abaixo.

2.5.1 Construções gramaticais complexas

Uma das principais dificuldades da análise automática de sentimentos se dá quando há ocorrência de construções gramaticais complexas. Nesses casos, não só a classificação da orientação do sentimento propriamente dita se torna mais difícil, como também os demais passos (como identificação de atributos (Mullen & Malouf, 2006)). Considere, por exemplo, a seguinte opinião fictícia sobre a CameraX:

"Ontem, fui ao shopping no centro da cidade decidido a comprar uma câmera fotográfica. Já havia realizado várias pesquisas na internet e lido diversos comparativos, então sabia exatamente o que queria: uma CameraX. O problema é que nenhuma loja tinha disponibilidade de estoque! Fui a todas as lojas de eletrônicos do shopping e não a encontrei. Vários vendedores me ofereceram outros modelos: segundo eles, a CameraY tem uma resolução bastante superior, além de dispor de uma bateria com maior duração. A CameraZ foi outra opção bastante recomendada devido à qualidade das fotos produzidas (pelo menos de acordo com o vendedor da LojaA). No fim, acabei comprando a Camera1, pois era a que apresentava o maior número de semelhanças com a câmera que eu pretendia comprar, além de também ser da MarcaX. Só que agora me sinto um pouco frustrado! Ainda que a Camera1 seja semelhante, acho que só ficaria de fato satisfeito se estivesse com minha CameraX neste momento!"

Exemplo 2: Relato fictício sobre busca por produto e decisão de compra, ilustrando algumas dificuldades do processamento automático de texto para análise de sentimento.

Esse tipo de comentário traz uma série de dificuldades que são de difícil resolução. O autor não faz críticas específicas a nenhuma câmera, nem apresenta razões objetivas para sua preferência pela CameraX. Na verdade, ele elogia algumas características de outras câmeras. Ao ler o texto, qualquer pessoa poderia facilmente inferir qual a preferência do autor, bem como quais os aspectos positivos das câmeras citadas. Entretanto, essa tarefa seria de difícil tratamento por uma ferramenta de análise automática de texto.

2.5.2 Figuras de Linguagem

O uso de figuras de linguagem é outro aspecto que dificulta a tarefa de analisar automaticamente a opinião expressa num texto. De maneira geral, qualquer figura de linguagem acrescenta complexidade à tarefa de análise de sentimento, porém destacam-se as figuras de linguagem apresentadas nas subseções a seguir.

2.5.2.1 Ironia/sarcasmo

Quando se diz o contrário daquilo que se pensa, deixando entender uma distância intencional entre o que se diz e aquilo que realmente se pretende dizer. Por exemplo:
"Uma beleza, minha CameraX! Depois de 2h de uso, o display simplesmente apagou! Acho que vou comprar outra amanhã! Recomendo a todos que querem ter dor de cabeça."

Há trabalhos na área de análise de sentimentos focados especificamente no tratamento de ironias, dada sua alta frequência de uso, principalmente em redes sociais (Carvalho et al., 2009). A integração desses algoritmos de classificação específicos num sistema genérico de análise de sentimento pode requerer um pré-processamento para determinar se o texto deve ser tratado como irônico ou não, o que por si só já é uma tarefa complexa.

2.5.2.2 Anfibologia ou Ambiguidade

É a duplicidade de sentido em uma construção sintática. Pode ser proposital, inconsciente, ou acontecer por mero descuido do autor ao escrever seu texto. Um enunciado é ambíguo ou anfibológico quando permite mais de uma interpretação. Uma vez que a anfibologia está diretamente ligada à sintaxe (posição e organização das palavras dentro de um enunciado), à relação das palavras entre si e, de um modo geral, à construção das frases, a sua ocorrência assumirá diferentes formas de acordo com a língua de que se trate, pois cada idioma possui sua própria estrutura e sua sintaxe. Dessa forma, o tratamento automático de ambiguidades, além de acrescentar complexidade à tarefa de classificação de sentimento, é ainda mais difícil de tratar em sistemas que lidam com vários idiomas. São apresentados abaixo exemplos de construções ambíguas:

Sujeito posposto a verbo transitivo direto
Exemplo: "Nas comparações que realizei entre as duas, venceu a CameraX a CameraZ".
Qual câmera foi a vencedora?

Uso de determinadas comparações

Exemplo: “Naquela época, a MarcaX não tinha credibilidade, como acontecia com a MarcaY”.

O autor quis dizer que MarcaY também não tem credibilidade ou que, ao contrário da MarcaX, a MarcaY tem credibilidade?

Alguns termos inerentemente ambíguos

Exemplo: “Após a negociação, a MarcaX ficou com a MarcaY”.

A MarcaX foi vendida para a MarcaY ou a MarcaX comprou a MarcaY, e assim a incorporou?

2.5.3 Resolução de co-referência

Este é um problema tratado pela área de análise de discurso (Pardo & Nunes, 2008). De forma resumida, consiste em conectar pronomes e outras expressões de referência aos objetos corretos, de forma que o texto em questão possa ser adequadamente interpretado. Por exemplo:

> *“Na semana passada eu comprei uma CameraX e meu irmão comprou uma CameraY. Ao chegar em casa, fomos testar nossos novos “brinquedos”. Achei que minhas fotos tinham maior resolução, mas as dele tinham cores mais vivas. Meu irmão gostou bastante da câmera dele. Como eu queria uma câmera com boa qualidade de imagem, minha escolha não foi boa. Acabei devolvendo a câmera ontem.”*

Exemplo 3: Relato fictício sobre busca por produto e decisão de compra, ilustrando o problema da resolução de co-referência

Os objetos que devem ser identificados nesse exemplo são “CameraX” e “CameraY”, e fazer isso de forma adequada já é uma tarefa complexa. Entender o que significa “nossos novos brinquedos”, “minhas fotos”, “câmera dele” e “minha escolha” é ainda mais difícil. A sentença “eu queria uma câmera com boa qualidade de imagem” expressa uma opinião positiva sobre uma câmera, mas não sobre o objeto do comentário, e sim sobre um objeto ideal, que seria o objetivo do comprador. Na última sentença, é difícil saber que “a câmera” se refere à CameraX, e que o fato de devolvê-la carrega implicitamente uma opinião negativa. O autor das opiniões é o autor do comentário, exceto para a

sentença “meu irmão gostou bastante da câmera dele”, cujo autor da opinião é o irmão do autor do texto.

2.5.4 Tratamento de negações

A presença de palavras ou expressões com conotação negativa pode inverter a polaridade do sentimento expresso no texto. Dessa forma, existe a necessidade de tratar especificamente esses casos e há vários trabalhos cujo foco é o tratamento adequado de negações (Wiegand et al., 2010). Por exemplo:
“A CameraX demonstra bom desempenho em ambientes com pouca iluminação”
A conotação da frase acima seria o inverso se palavras ou expressões negativas fossem usadas:
“A CameraX não demonstra bom desempenho em ambientes com pouca iluminação”
“A CameraX nunca demonstra bom desempenho em ambientes com pouca iluminação”
“Nunca achei que a CameraX demonstrasse bom desempenho em ambientes com pouca iluminação”

Note que há uma grande variedade de construções possíveis de uso de expressões negativas.

2.5.5 Sentenças comparativas

Quando a opinião é expressa de forma comparativa, deve-se fazer um processamento específico, pois ao mesmo tempo em que se elogia uma entidade se está criticando outra (Ganapathibhotla & Liu, 2008). Por exemplo: “A CameraX é muito melhor que a CameraY”

2.5.6 Erros gramaticais ou de ortografia

Erros de escrita são bastante comuns em espaços destinados a comentários de usuários em websites e também em redes sociais. Às vezes, esses erros são propositais e podem ser vistos como uma forma “super-coloquial”, porém aceitável, de comunicação. Mas na maioria dos casos os erros são causados por falta de conhecimento das normas cultas da língua, e podem prejudicar o

processo de classificação, principalmente quando são erros ortográficos em palavras opinativas (normalmente usadas para construção de léxicos que servem como indicadores de polaridade). Por exemplo: “A CameraX é esselente”

Em sistemas baseados em léxicos para determinação da polaridade, a palavra que exprime o sentimento na sentença não seria encontrada e o processamento desse texto, a princípio simples, seria prejudicado. Alternativas para esse problema incluem algum tipo de comparação *fuzzy* entre as palavras ou um pré-processamento fonético sobre o texto (Ermakov & Ermakova, 2013).

2.5.7 Uso de gírias

Gírias ou expressões locais podem ser de difícil tratamento. Enquanto algumas gírias são de conhecimento geral, outras podem ser bastante particulares de regiões ou até de grupos de pessoas, alterando completamente o sentido do texto. Há trabalhos dedicados apenas ao tratamento de gírias no contexto de análise de sentimento (Soliman et al., 2013). Um tipo específico de gírias, são as contrações frequentemente usadas na internet, como kd (cadê), pq (porque), qd (quando), lol (laughing out loud), brb (be right back), etc. Esses casos são mais fáceis de tratar, até mesmo através de um dicionário estático, uma vez que não há as variações semânticas frequentes.

2.5.8 Resolução de contexto

Este é um problema complexo, visto com muita frequência em redes sociais. Refere-se à análise de um texto que, para seu completo entendimento, requer informações que não estão disponíveis naquele ambiente ou naquele texto. Por exemplo, durante um jogo de futebol é comum que torcedores postem mensagens comentando a partida, sem necessariamente se referir explicitamente a ela; dado o contexto, todos (ou boa parte das pessoas) que lerem as mensagens conseguirão identificar sobre o que se está falando. Outro caso típico acontece em redes sociais com limitação de caracteres por mensagem (por exemplo, Twitter), em que é comum a ocorrência de usuários “conversando” através de mensagens isoladas: nesse caso, para entender o sentido de uma determinada mensagem, pode ser necessário ler toda a troca de

mensagens que ocorreu anteriormente[9]. Mais um exemplo se vê em mensagens do tipo: "Acredito que a situação econômica da Europa influenciará a América Latina", que, por si só, pode representar uma opinião favorável ou desfavorável à situação econômica da América Latina. A orientação de opinião é esclarecida após a avaliação do contexto de crise atualmente verificado na Europa. Há trabalhos que procuram tratar o problema da resolução de contexto (Agarwal et al., 2015), porém de forma ainda superficial e dependente de bases pré-concebidas.

Todas essas particularidades demonstram que um sistema que trabalhe com dados reais para realizar análise de sentimento requer o tratamento de uma série de casos especiais, que tornam o problema de análise de sentimento ainda mais complexo.

2.6 Classificação de Viés Opinativo de Usuário

Em muitas ocasiões, ao invés de saber qual a orientação da opinião de uma mensagem ou documento isolado, é útil saber a opinião geral do autor que escreveu tal mensagem ou documento. Uma forma imediata de fazer isso é através de alguma forma de agregação das mensagens postadas (por exemplo, através de votação, com a orientação global sendo definida de acordo com a orientação prevalente, ou neutra em caso de equilíbrio nas orientações das postagens).

Entretanto, esse cenário abre a possibilidade para aplicação de outras técnicas. Há trabalhos (Conover et al., 2011) que apresentam algumas formas de classificar a orientação política de usuários do Twitter (baseadas tanto em conteúdo quanto na estrutura de relacionamentos da rede). A primeira abordagem usa um SVM sobre *hashtags* e outra sobre o conteúdo das mensagens. Os experimentos mostram que o SVM que usa apenas as *hashtags* apresentou um resultado melhor, com cerca de 91% de precisão. Outra

[9] O Twitter atualmente possui uma funcionalidade que permite o acesso a todas as mensagens trocadas dentro de um mesmo contexto pelos seus usuários, mas para isso requer que os usuários necessariamente indiquem que estão respondendo uma mensagem anterior (algo que pode não acontecer em todos os casos). Ainda assim, a resolução de contexto não é uma tarefa simples.

abordagem baseada em conteúdo usa análise de semântica latente para identificar estruturas associadas às preferências políticas dos usuários. Porém, a melhor abordagem usa a estrutura de comunidades definidas através de redes de difusão, que atinge precisão de 95%.

Outros trabalhos (Guerra, 2013) usam uma abordagem baseada em transferência de conhecimento (*transfer learning*) para a classificação de opiniões. Em primeiro lugar, é usado um conjunto inicial de usuários com opinião conhecida. Esse conjunto é expandido através de três formas de relacionamento entre usuários: conexão entre usuários das redes sociais, comunicação entre usuários através de troca de mensagens, e endosso de opiniões através de funções como "curtir" do Facebook ou "retweet" no Twitter. O sistema armazena o conjunto de usuários com orientação conhecida e as suas postagens são monitoradas. Os termos comumente usados por usuários de cada lado são usados para classificar novas mensagens.

Há também trabalhos (Hu et al., 2013) que aplicam uma abordagem supervisionada para classificação de textos curtos, como aqueles vistos no Twitter, através de uma modelagem de otimização matemática que aplica as teorias de consistência de sentimento e contágio emocional no processo de aprendizagem.

2.7 Considerações Finais

Esta seção apresentou uma revisão da área de Análise de Sentimentos, focando principalmente na abordagem tradicional para o problema, que é o processamento textual das opiniões emitidas. Foram apresentadas, ainda, dificuldades inerentes à classificação baseada no tratamento textual.

A proposta deste trabalho é fazer uso de uma nova dimensão de informação (o relacionamento entre usuários) existente em redes sociais, que é um ambiente frequentemente utilizado para emissão de opiniões sobre os mais variados assuntos. Conforme estabelece a hipótese desta tese, o uso dessa informação adicional permitirá uma classificação mais precisa dos usuários de acordo com suas opiniões, além de possibilitar ainda a inferência de sobre a

opinião de pessoas que sequer se manifestaram acerca de um determinado assunto.

3 Classificação Coletiva – Revisão da Literatura

A ocorrência de links (ou relacionamentos, de forma mais genérica) entre entidades é algo comum em qualquer repositório de informações e até mesmo entre pessoas. Esses links normalmente contem informações que podem indicar características das entidades, como a importância ou a categoria do objeto. Em alguns casos, nem todos os possíveis links entre os objetos são observados. Nessas situações, pode ser interessante inferir a existência de tais links. Em outros casos, os relacionamentos entre objetos podem ser úteis para se inferir propriedades dos próprios objetos.

Através de uma análise de links, padrões complexos podem emergir, como identificação de subestruturas a exemplo de comunidades, grupos ou subgrafos comuns. Algoritmos tradicionais de mineração de dados como mineração por regras associativas e análise de clusters comumente tem por objetivo encontrar padrões num conjunto de dados caracterizados por uma coleção de instâncias independentes, o que é consistente com o problema clássico de inferência estatística de identificar um modelo dada uma amostra independente e identicamente distribuída. Pode-se entender esse processo como um modelo de aprendizado para os atributos dos nós de um grafo homogêneo, ignorando os links entre os nós desse grafo.

Entretanto, essa modelagem é insuficiente para abordar problemas de mineração de dados em conjunto de dados heterogêneos e profundamente estruturados, que são mais bem representados como redes ou grafos. Os domínios normalmente consistem de vários tipos de objetos, que podem ser interligados de várias formas. Assim, o grafo pode conter diferentes tipos de nós e arestas. Aplicar diretamente rotinas tradicionais de inferência estatística, que assumem que as instâncias são independentes, pode levar a conclusões imprecisas (Jensen D. , 1999). As potenciais correlações entre os objetos interligados devem ser tratadas de forma adequada; aliás, essas interligações podem conter informações que devem ser exploradas. Essas informações podem ser usadas para melhorar a precisão preditiva dos modelos gerados por aprendizado[10], já que os atributos de objetos interligados são normalmente

[10] Até mesmo a a própria estrutura de grafo poder ser usada como um elemento importante a ser incluído no modelo.

correlacionados. Além disso, é mais provável que existam links entre objetos que apresentam algo em comum.

A abordagem proposta neste trabalho para a análise de sentimento é baseada no princípio da homofilia (Macskassy & Provost, 2007), que estabelece que indivíduos tendem a se conectar a outros com os quais apresentam alto grau de semelhança. A partir desse conceito, são aplicadas técnicas de Classificação Coletiva (Rattingan, Maier, & Jensen, 2007), (Sen et al., 2008) sobre o grafo que representa as instâncias que serão classificadas e suas interconexões. A intenção é explorar o fato de que a classificação coletiva não utiliza individualmente as características locais dos nós no processo de inferência, mas também as características dos nós relacionados. Além disso, a classificação é executada de forma simultânea sobre todas as instâncias, o que permite considerar as influências que cada instância exerce sobre outras às quais está relacionada.

Sendo a Classificação Coletiva parte importante do trabalho, é realizada neste capítulo uma revisão do estado-da-arte da área e os seus principais algoritmos.

3.1 Caracterização do Problema

Técnicas convencionais de aprendizagem de máquina focam na classificação de dados que consistem de objetos com estruturas idênticas, presumidamente independentes e identicamente distribuídas. Muitos cenários reais, entretanto, não apresentam essa homogeneidade estrutural. Quando as instâncias possuem relacionamentos entre si, as classes dos objetos relacionados tendem a apresentar alguma correlação. O desafio é desenvolver algoritmos de classificação que explorem tais correlações e, de forma conjunta, inferem as classes associadas aos objetos do grafo.

Classificação Coletiva é um método de classificar instâncias em conjunto (de forma simultânea). Para tanto, são usados algoritmos de inferência coletiva que exploram as dependências entre as instâncias, possibilitando que a Classificação Coletiva geralmente alcance melhores resultados do que métodos tradicionais de classificação em cenários em que existam relacionamentos entre as instâncias. A Classificação Coletiva pode ser entendida como um problema

de otimização combinatória, cujo objetivo é produzir a melhor classificação conjunta para as instâncias.

Obviamente, por necessitar que as instâncias sejam interdependentes, este método não é aplicável a problemas em que não exista relacionamento entre as instâncias. Esta restrição, entretanto, não impede sua utilização em problemas reais, afinal muitos problemas são mais bem modelados através de uma representação que guarde relacionamentos entre as instâncias. Isto acontece, por exemplo, em redes sociais, cocitacões em publicações, redes de e-mail, páginas web com referências entre si, e outros problemas em que existe relacionamento entre os objetos. Esses relacionamentos normalmente carregam informações implícitas, porém úteis, ao processo de classificação automática, comumente negligenciadas por algoritmos tradicionais de inferência. A seguir, será apresentada uma definição do problema junto a um exemplo que ajudará a ilustrar os princípios e o funcionamento do método.

Uma definição formal de classificação coletiva é dada a seguir (Macskassy & Provost, 2007): dado um grafo G=(V,E,X,Y,L), onde:

- V é um conjunto de nós;
- E é um conjunto de arestas;
- Cada $x_i \in X$ é um vetor de atributos para o nó $v_i \in V$ que representa uma classe conhecida para o nó v_i;
- Cada $Y_i \in Y$ é uma variável para o label do nó v_i, e representa que a classe do nó v_i é desconhecida;
- L é o conjunto de possíveis labels.

Além disso, é dado um conjunto de labels Y^K conhecidos para os nós $V^K \subset V$, tais que $Y^K = \{y_i | v_i \in V^K\}$. O objetivo da Classificação Coletiva é inferir Y^U, os valores de Y_i para os nós cujos labels ainda são desconhecidos ($V^U = V - V^K$), ou a distribuição de probabilidade sobre esses valores.

Como exemplo, considere a tarefa de classificar se uma página web na rede de uma universidade pertence a um professor ou a um aluno. Métodos tradicionais de aprendizagem supervisionada ignorariam as informações de relacionamento (neste caso, os hiperlinks) entre as páginas, e realizariam o processo de classificação considerando apenas atributos extraídos do conteúdo, como o texto da página, seu título ou sua URL. Um método de classificação

relacional usaria os hiperlinks entre as páginas para acrescentar atributos ao processo de classificação (por exemplo, o texto das páginas referenciadas), o que normalmente melhora a precisão da classificação. Porém, ganhos ainda maiores de precisão podem ser obtidos ao usar não atributos derivados dos objetos referenciados, mas as classes (labels) desses objetos, quando conhecidos (Jensen, Neville, & Gallagher, 2004). Para tanto, é necessário estimar e refinar iterativamente as classes dos nós desconhecidos. A este processo de inferência conjunta das classes de instâncias representadas por um grafo se dá o nome de Classificação Coletiva.

3.2 Métodos de Inferência Coletiva

Alguns métodos de inferência tradicional podem ser usados para classificar instâncias coletivamente, como árvores de junção (Huang & Darwiche, 1996) e eliminação de variáveis (Zhang & Poole, 1996). Apesar de esses algoritmos produzirem resultados exatos, eles tem custo computacional elevado e por esse motivo sua utilização em aplicações reais é inviável até mesmo em grafos pequenos. Por este motivo, a maior parte do esforço da comunidade de pesquisa de Inferência Coletiva se concentra no desenvolvimento de métodos de classificação por aproximação. Esta seção traz uma revisão dos principais algoritmos desenvolvidos, que podem ser divididos em duas super classes: aqueles que fazem uso de classificadores locais (descritos na seção 3.2.1) e os que modelam o problema de classificação coletiva como uma função global a ser otimizada (seção 3.2.2).

3.2.1 Algoritmos Baseados em Classificadores Locais

Nesta seção, serão descritos métodos de classificação coletiva que usam abordagens iterativas e informação da vizinhança local para gerar características que são usadas para aprender classificadores locais.

Estes algoritmos constroem vetores de características para os nós a partir da informação dos próprios nós e da sua vizinhança imediata. Esses vetores de características são usados juntamente com o conjunto de nós com classe conhecida Y^K para construir algum tipo de classificador local. Os principais algoritmos baseados neste princípio são detalhados nas subseções seguintes.

3.2.1.1 Algoritmo de Votação Ponderada da Vizinhança

Chamado inicialmente de Classificador Probabilístico Relacional da Vizinhança (Macskassy & Provost, 2003), é um algoritmo relacional que estima as probabilidades das labels desconhecidas baseado exclusivamente na vizinhança imediata cujos nós tem labels conhecidas. O algoritmo pressupõe duas condições básicas para operação:

1) Algumas das labels do grafo devem ser conhecidas;
2) O grafo apresenta a propriedade da homofilia (nós conectados representam entidades similares, e assim tendem a pertencer às mesmas classes).

Dessa forma, o algoritmo provavelmente não apresentará bons resultados ao ser aplicado em grafos com nós isolados ou com nenhum ou poucos nós com classe conhecida. Ele estima $P(y_i|N_i)$, a probabilidade de um nó v_i ser classificado com a label y_i dada a sua vizinhança N_i, como sendo a votação ponderada das classes dos nós em N_i que pertencem à classe y_i. Mais formalmente,

$$P(y_i|N_i) = \frac{1}{\sum_{v_j \in N_i} w(v_i, v_j)} \sum_{v_j \in N_i | label(v_j) = y_i} w(v_i, v_j)$$

onde $w(v_i, v_j)$ é o peso da aresta que liga os nós v_i e v_j. Os nós em N_i que não são do mesmo tipo de v_i são ignorados (no caso de grafos que representam mais de um tipo de nó). Se N_i for vazia ou não contiver nós com classe conhecida, o algoritmo fará a estimativa baseado nas probabilidades *a priori* (isto é, a distribuição de probabilidade das classes conhecidas em todo o grafo).

Usando o exemplo citado na seção 3.1, considere um grafo que representa páginas web interligadas, com cada página pertencendo a uma classe (página de aluno ou página de professor). Num grafo não direcionado, em que duas páginas são relacionadas se houver hiperlink entre elas, qualquer que seja a direção, o algoritmo classificaria uma página candidata *p* como sendo de aluno se a maioria das páginas ligadas a *p* (a despeito da direção) forem páginas de aluno.

Perceba, entretanto, que haverá um problema se muitas páginas na vizinhança imediata de *p* não tiverem classe conhecida. Isto porque o algoritmo

não propaga a informação ao longo do grafo. Isto é feito em algoritmos mais sofisticados, como os que serão apresentados a seguir.

3.2.1.2 Algoritmo de Classificação Iterativa

O Algoritmo de Classificação Iterativa (Lu & Getoor, 2003) se baseia num princípio de funcionamento bastante simples: para classificar nós com classe desconhecida, ele aplica iterativamente um algoritmo de classificação local que usa apenas as informações da vizinhança imediata de um nó para sua classificação. Mais formalmente: considere um nó $Y_i \in Y$, cuja classe é desconhecida, e suponha que as classes de todos os nós da sua vizinhança N_i são conhecidas. O algoritmo aplica um classificador local M_L que recebe as classes de N_i como parâmetros de entrada e retorna a melhor classe $y_i \in L$ para o nó Y_i (sendo L o conjunto de possíveis labels ou classes).

Numa definição mais genérica, o classificador local pode retornar uma distribuição de probabilidade sobre as possíveis classes em L, ao invés da classe em si. Nesse caso, o algoritmo seleciona a classe que corresponde ao valor com máxima probabilidade. Ou seja, ao invés de usar o valor retornado por M_L, será usado o valor $argmax_{l \in L} M_L$. Como se pode concluir da definição acima, o algoritmo de classificação local é bastante flexível e, dessa forma, diversos algoritmos de classificação (redes neurais, árvores de decisão, SVM, etc) podem ser usados como classificador local M_L.

Entretanto, como dificilmente serão conhecidas as classes de todos os nós da vizinhança de um determinado nó, é necessário repetir o processo de classificação iterativamente, sendo que em cada iteração será feita a classificação de um nó Y_i com a melhor estimativa considerando o resultado de M_L sobre a vizinhança N_i, num processo contínuo até que a convergência seja atingida ou que seja atingido um limite de iterações. Como se pode concluir da descrição acima, o Algoritmo de Classificação Iterativa não estabelece incerteza no processo de classificação. Isto é, ou os nós não recebem classificações (o que acontece quando toda a vizinhança imediata tem classe desconhecida na primeira iteração, ou se todos os nós tiverem classe desconhecida), ou recebem uma classificação definitiva.

Conforme mencionado anteriormente, diversos algoritmos locais podem ser usados para classificação a partir da vizinhança, mas como exemplo

considere um algoritmo simples de votação: o nó receberá a classe prevalente da sua vizinhança imediata em cada iteração, e para após um número pré-determinado de iterações ou se não houver alteração relevante na classificação do grafo. Esta definição seria semelhante ao algoritmo apresentado na seção 3.2.1.1, o que mostra que o Algoritmo de Classificação Iterativa pode ser entendido como um framework para classificação em grafos (Aggarwal, 2011). Abaixo é apresentado o pseudocódigo do algoritmo:

```
For each Y_i ∈ Y^U (etapa de bootstrap)
        K = Y^K ∩ N_i
        y_i = M_L(K)
End for
Repeat (etapa de classificação iterativa)
        Gerar O, uma ordenação aleatória dos nós em Y^U
        For each Y_i ∈ O
                y_i = M_L(N_i)
        End for
Until o limite de iterações seja alcançado ou que não haja
alteração relevante nas classificações
```

Algoritmo 1: Pseudocódigo para o algoritmo de Classificação Iterativa

3.2.1.3 Amostragem de Gibbs

O algoritmo de Amostragem de Gibbs (Gibbs Sampling ou Gibbs Sampler) (Geman & Geman, 1984) é um método que aplica estatística bayesiana e, assim como outros algoritmos de Cadeia de Markov de Monte Carlo, produz uma sequência de instâncias a partir de uma distribuição de probabilidade, sendo que cada instância está correlacionada às instâncias da vizinhança. Esta sequência pode ser usada para aproximar uma distribuição conjunta, como no caso da classificação coletiva (que pode ser entendida como uma otimização global).

A Amostragem de Gibbs foi proposta inicialmente para a área de visão computacional aplicada ao problema de restauração de imagens, e é um dos algoritmos mais precisos de inferência aproximada. Entretanto, apresenta custo computacional relativamente alto (McDowell, Gupta, & Aha, 2007) e é difícil

determinar quando sua convergência foi atingida[11]. Devido a essas dificuldades, normalmente o algoritmo de Amostragem de Gibbs é usado pelos pesquisadores da área de classificação coletiva numa versão simplificada, semelhante ao Algoritmo de Classificação Iterativa, em que é aplicado um classificador local M_L que estima a distribuição de probabilidade condicional para a classe do nó Y_i a partir de todas as classes dos nós na vizinhança imediata N_i.

A principal diferença dessa aproximação para o método original é que não há garantia de que essa distribuição de probabilidade condicional é a distribuição correta a partir da qual se deve fazer a amostragem. No melhor caso, é possível apenas assumir que a distribuição de probabilidade condicional dada pelo classificador local é uma aproximação da distribuição correta (Neville & Jensen, 2007). O pseudocódigo do algoritmo de Amostragem de Gibbs é apresentado abaixo.

[11] Os testes de convergência são computacionalmente caros ou de complexa implementação (Sen et al., 2008).

For each $Y_i \in Y^U$ (etapa de *bootstrap*)

$K = Y^K \cap N_i$

$y_i = M_L(K)$

End for

For n=1 until B (etapa de *burn-in*)

Gerar O, uma ordenação aleatória dos nós em Y^U

For each $Y_i \in O$

$y_i = M_L(N_i)$

End for

End for

For each $Y_i \in Y^U$ (etapa de inicialização das contagens amostrais)

For each $l \in L$

$count(i, l) = 0$

End for

End for

For n=1 until S (etapa de geração das amostras)

Gerar O, uma ordenação aleatória dos nós em Y^U

For each $Y_i \in O$

$y_i = M_L(N_i)$

$count(i, y_i) = count(i, y_i) + 1$

End for

For each $Y_i \in O$ (etapa de cálculo das classes finais)

$y_i = argmax_{l \in L} count(i, l)$

End for

* Normalmente B << S (ex: B=200 e S=2000)

Algoritmo 2: Pseudocódigo para o algoritmo de Amostragem de Gibbs

Ou seja, durante o *burn-in*, não se registra qualquer estatística. Na etapa de geração das amostras, registram-se quantas vezes cada classe $l \in L$ foi atribuída a um nó Y_i. O vetor normalizado das contagens representa a distribuição final de probabilidade para cada nó. Como acima foi ilustrada a seleção da maior probabilidade, não é necessário computar a distribuição de probabilidade para determinação da classificação, apenas identificar a maior contagem.

3.2.2 Algoritmos Baseados em Modelagens Globais

Como destacado no início deste capítulo, existe outra classe de algoritmos de classificação coletiva que modelam o problema como uma função global que deve ser otimizada. Entre esses algoritmos, os mais usados são o de Propagação Cíclica de Conhecimento e o de Etiquetagem Relaxada, que serão detalhados nesta seção. Entretanto, antes de apresentar os detalhes desses métodos, é necessário trazer alguns conceitos sobre Campos Aleatórios de Markov, que são usados para a modelagem do problema.

Um Campo Aleatório de Markov é definido por $\{G, \Psi\}$, onde G é um grafo conforme a definição apresentada na seção 3.1 e Ψ é um conjunto de cliques potenciais, formado por três tipos de funções:

- Para cada $Y_i \in \mathrm{Y}$, $\psi_i \in \Psi$ é um mapeamento $\psi_i: L \rightarrow \mathbb{R}_{\geq 0}$, onde $\mathbb{R}_{\geq 0}$ é um conjunto de números reais não negativos;

- Para cada aresta que liga um nó com classe desconhecida a um nó com classe conhecida $(Y_i, X_j) \in \mathrm{E}$, $\psi_{ij} \in \Psi$ é um mapeamento $\psi_{ij}: L \rightarrow \mathbb{R}_{\geq 0}$;

- Para cada aresta que liga dois nós com classes desconhecidas $(Y_i, Y_j) \in \mathrm{E}$, $\psi_{ij} \in \Psi$ é um mapeamento $\psi_{ij}: L \times L \rightarrow \mathbb{R}_{\geq 0}$.

Antes de apresentar uma definição para um clique potencial, considere que x representa os valores atribuídos a todas as variáveis observadas em G e x_i representa o valor atribuído ao nó X_i. Analogamente, considere que y representa os valores atribuídos a todas as variáveis não observadas em G e y_i representa o valor atribuído ao nó Y_i. Um clique potencial $\phi_i(y_i)$ é definido conforme a equação a seguir:

$$\phi_i(y_i) = \psi_i(y_i) \prod_{Y_i, Y_j \in E} \psi_{ij}(y_i)$$

Por simplicidade, o clique potencial $\phi_i(y_i)$ definido acima será representado adiante por ϕ_i. Dados o grafo *G* e o conjunto Ψ de cliques potenciais conforme definição acima, o Campo Aleatório de Markov está associado com a distribuição de probabilidade definida abaixo:

$$P(y|x) = \frac{1}{\mathrm{Z(x)}} \prod_{Y_i \in Y} \phi_i(y_i) \prod_{(Y_i, Y_j) \in E} \psi_{ij}(y_i, y_j)$$

onde x denota os valores observados de *X*, e Z(x) é uma função conforme definição abaixo:

$$\mathrm{Z(x)} = \sum_{y'} \prod_{Y_i \in Y} \phi_i(y'_i) \prod_{(Y_i, Y_j) \in E} \psi_{ij}(y'_i, y'_j)$$

Para ilustrar as definições acima, considere o exemplo usado anteriormente de classificação de páginas web nas classes "aluno" ou "professor", com o incremento de cliques potenciais e respectivos valores de mapeamento, conforme a Figura 2. No exemplo[12], ψ_1 e ψ_2 representam dois cliques potenciais definidos sobre as variáveis não observadas Y_1 e Y_2 do grafo. Analogamente, há uma função ψ definida para cada aresta em que pelo menos um dos vértices tem classe desconhecida (ou seja, uma aresta em que pelo menos um dos vértices é uma variável não observada). Por exemplo, ψ_{13} representa um mapeamento de *L* para $\mathbb{R}_{\geq 0}$, sendo *L* o conjunto de possíveis labels, que neste caso assume o valor *L*={aluno, professor}. Há apenas uma aresta entre duas variáveis não observadas no grafo, que está associada ao clique potencial ψ_{12}. O exemplo também mostra o cálculo dos ϕ cliques potenciais.

[12] Simplificação de um exemplo semelhante em (Sen et al., 2008).

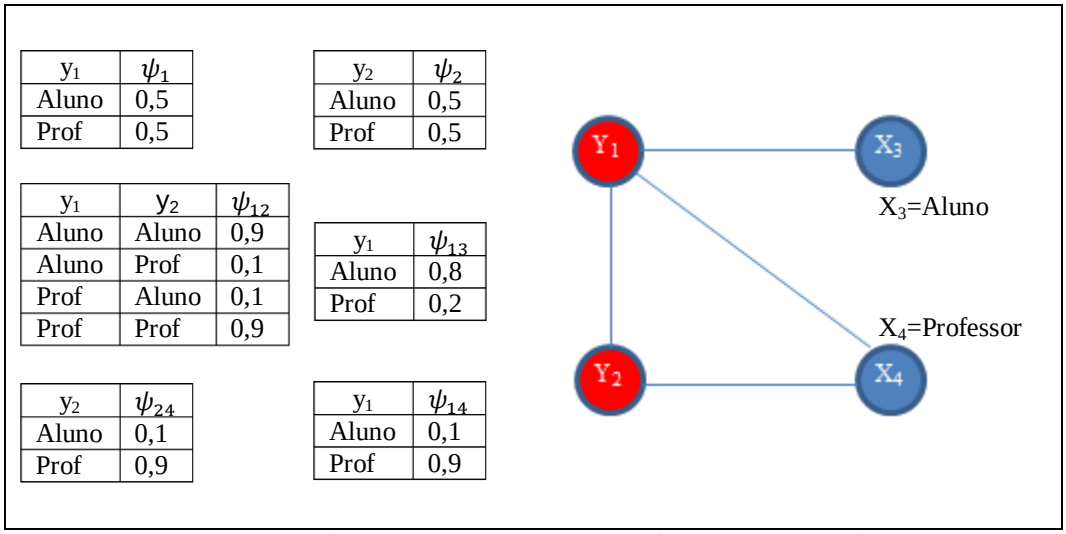

Figura 2: Exemplo de classificação coletiva modelado através de Campos Aleatórios de Markov

Em resumo, dada uma variável não observada Y_i, o algoritmo identifica todas as arestas que a conectam a uma variável observada no grafo e multiplica os cliques potenciais correspondentes pelo clique potencial definido pela própria Y_i. Tomando o exemplo apresentado na Figura 2, temos o seguinte cálculo para ϕ_1 e ϕ_2:

$\phi_1 = \psi_1 * \psi_{14} * \psi_{13} =$

y1	ϕ1
Aluno	0,04
Prof	0,09

$\phi_2 = \psi_2 * \psi_{24} =$

y2	ϕ_2
Aluno	0,05
Prof	0,45

Figura 3: Cálculo dos ϕ cliques potenciais para o exemplo apresentado na Figura 2

Como se pode ver, a teoria por trás do algoritmo é simples, e dessa forma é conceitualmente trivial classificar os nós não observados no grafo. Por exemplo, um possível critério seria assumir que a melhor classificação para Y_i é simplesmente o que corresponde à maior probabilidade marginal obtida pela soma se todas as outras variáveis da distribuição de probabilidade associada com o Campo Aleatório de Markov. Entretanto, o custo computacional para a execução do algoritmo em grafos reais é proibitivo, pois requer a soma sobre um número exponencial de termos, e é exatamente por este motivo que são necessários métodos de inferência por aproximação. Os dois principais são descritos a seguir.

3.2.2.1 Algoritmo de Propagação Cíclica de Conhecimento

É uma abordagem de programação dinâmica[13] baseada em envio de mensagens para inferência em grafos. O algoritmo é derivado de uma versão desenvolvida inicialmente exclusivamente para árvores[14] (ou seja, um tipo específico de grafo), chamada apenas de Propagação de Conhecimento (*Belief Propagation*). Sua ideia geral de funcionamento é simples: a classe de um nó será determinada através de mensagens enviadas pela rede, sendo que os nós com classes conhecidas (nós observados) propagam o conhecimento pelo grafo através dessas mensagens, que são enviadas por um nó informando aos nós adjacentes como atualizar seu conhecimento de acordo com probabilidades *a priori*, probabilidades condicionais e evidências (os nós observados). O envio de mensagens se mantem até que seja atingido um estado de convergência na rede (Yedidia, Freeman, & Weiss, 2005).

O ponto-chave do algoritmo é o cômputo de $m_{i\to j}(y_j)$, uma mensagem enviada de Y_i a Y_j, e de $b_i(y_i)$, o conhecimento sobre o nó Y_i que representa a probabilidade marginal de atribuir a classe y_i ao nó Y_i. O cálculo das mensagens é feito conforme a equação abaixo:

$$m_{i\to j}(y_j) = \alpha \sum_{y_i \in L} \psi_{ij}(y_i, y_j)\, \phi_i(y_i) \prod_{Y_k \in N_i \cap Y \setminus Y_j} m_{k\to i}(y_i)\,, \forall y_j \in L$$

onde α é uma constante de normalização que garante que a soma das mensagens enviadas do Y_i ao nó Y_j seja 1. Isto é:

$$\sum_{y_j} m_{i\to j}(y_j) = 1$$

A definição para o cálculo dos valores $b_i(y_i)$, é dada pela equação abaixo:

[13] Método de resolução de problemas que consiste em subdividir um problema complexo em problemas mais simples (Sniedovich, 2010).

[14] Para árvores, o algoritmo produz probabilidades marginais exatas, enquanto a versão cíclica produz uma aproximação das probabilidades. Esta aproximação, entretanto, é próxima dos valores exatos que seriam produzidos pelo algoritmo de Árvore de Junção (Murphy, Weiss, & Jordan, 1999).

$$b_i(y_i) = \propto \phi_i(y_i) \prod_{Y_j \in N_i \cap Y} m_{j \to i}(y_i), \forall y_i \in L$$

onde α é uma constante de normalização que garante que a soma das probabilidades marginais seja 1. Isto é:

$$\sum_{y_i} b_i(y_i) = 1$$

Em resumo, a execução do algoritmo se dá da seguinte forma: de início, o nó $Y_i \in Y$ envia mensagens aos seus vizinhos que também não tem classe conhecida, isto é, os nós em $N_i \cap Y$, até que as mensagens estabilizem. Após a estabilização ser atingida, é calculada a probabilidade marginal $b_i(y_i)$, conforme equação apresentada anteriormente. O pseudocódigo do algoritmo é listado a seguir:

For each $(Y_i, Y_j) \in E(G)$ (etapa de inicialização)

    For each $y_j \in L$

        $m_{i \to j}(y_j) = 1$

    End for

End for

Repeat

    For each $(Y_i, Y_j) \in E(G)$

        For each $y_j \in L$

        $m_{i \to j}(y_j) = \alpha \sum_{y_i} \psi_{ij}(y_i, y_j)\, \phi_i(y_i) \prod_{Y_k \in N_i \cap Y \setminus Y_j} m_{k \to i}(y_i)$

        End for

    End for

Until as mensagens $m_{i \to j}(y_j)$ não apresentem alterações (etapa de troca de mensagens)

For each $Y_i \in Y$ (etapa de cálculo das probabilidades)

    $b_i(y_i) = \propto \phi_i(y_i) \prod_{Y_j \in N_i \cap Y} m_{j \to i}(y_i)$

End for

Algoritmo 3: Pseudocódigo para o algoritmo de Propagação Cíclica de Conhecimento

Para ilustrar o funcionamento do algoritmo, considere o exemplo de classificação de páginas web nas classes “Aluno” e “Professor”, apresentado novamente a seguir para facilitar a referência:

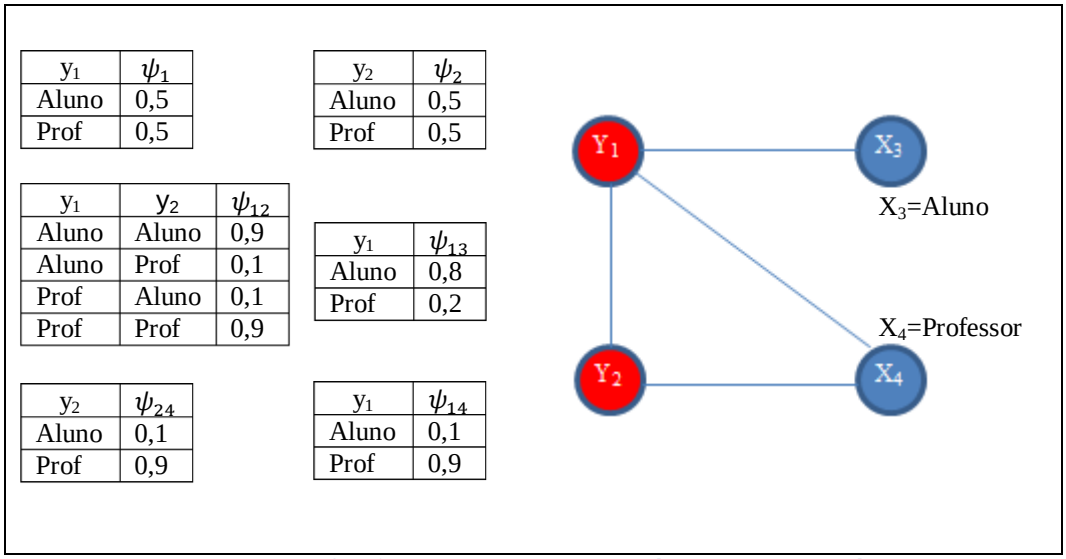

Figura 4: Exemplo de classificação coletiva modelado através de Campos Aleatórios de Markov

No exemplo ilustrado acima, como $Y = \{Y_1, Y_2\}$, temos que $N_i \cap Y \backslash Y_j = \emptyset, \forall ij$. Dessa forma o produtório da equação se anula e o cálculo das mensagens trocadas entre os nós Y_1 e Y_2 seria feito conforme as equações abaixo:

$$m_{1\to2}(y_2) = \sum_{y_1} \psi_{12}(y_1, y_2)\, \phi_1(y_1)$$

$$m_{2\to1}(y_1) = \sum_{y_2} \psi_{12}(y_1, y_2)\, \phi_2(y_2)$$

3.2.2.2 Etiquetagem Relaxada

Técnicas de inferência baseadas em etiquetagem relaxada (*relaxation labeling*) (Chakrabarti et al., 1998), (Rosenfeld, Hummel, & Zucker, 1976) são comumente usadas na área de visão computacional para tratamento de ambiguidade e ruído, porém são aplicáveis a vários outros cenários (Hummel & Zucker, 1983). Os elementos básicos do método são as características dos objetos e um conjunto de labels sob os quais esses objetos podem ser classificados. Assim como no algoritmo de Propagação Cíclica de Conhecimento, são usados esquemas probabilísticos de atribuição de labels, de forma que, para cada característica, pesos (ou probabilidades) são atribuídos a cada label no conjunto de objetos, representando uma estimativa da probabilidade de que um determinado label seja o correto para aquele objeto.

As abordagens probabilísticas são usadas para maximizar (ou minimizar) os pesos através de ajustes iterativos nos pesos, levando em

consideração as probabilidades associadas às características da vizinhança. O algoritmo de etiquetagem relaxada, dessa forma, não garante convergência, e assim é possível que não se obtenha uma solução final com um único label atribuído por objeto contendo peso 1 para cada característica.

Para detalhar o processo de classificação, considere as definições apresentadas na seção 3.1 e também que $b_j(y_j)$ é a probabilidade marginal de atribuir a label y_j ao nó $Y_j \in Y$.

Os axiomas usuais de probabilidade se aplicam, isto é:

- Cada probabilidade satisfaz à condição $0 \leq b_j(y_j) \leq 1$, e se $b_j(y_j) = 0$ significa que a label y_j é impossível para o nó Y_j, e $b_j(y_j) = 1$ significa que há certeza sobre a label (y_j).
- Os labels do conjunto L são mutuamente excludentes e exaustivos. Assim, para cada objeto, $\sum_L P_i(l_k) = 1$, o que significa que cada objeto é corretamente descrito por exatamente uma label do conjunto L.

O processo de atribuição de labels começa com uma atribuição inicial (possivelmente aleatória) de probabilidades para cada objeto. O algoritmo iterativamente ajusta essas probabilidades considerando as probabilidades da vizinhança de cada objeto, até que o processo atinja convergência (o que ocorre quando nenhuma ou pouca mudança ocorre entre iterações sucessivas). O pseudocódigo do algoritmo é apresentado a seguir:

For each $Y_i \in Y$ (etapa de inicialização)

    For each $y_i \in L$

        $b_i(y_i) = 1$

    End for

End for

Repeat

    For each $Y_j \in Y$

        For each $y_j \in L$

            $b_j(y_j) = \alpha\phi_j(y_j)\prod_{Y_i \in N_j \cap Y, y_i \in L} \psi_{ij}^{b_i(y_i)}(y_i, y_j)$

        End for

    End for

Until as mensagens $b_j(y_j)$ não apresentem alterações (etapa de troca de mensagens)

Algoritmo 4: Pseudocódigo para o algoritmo de Relaxation Labeling

Esse modelo de representação de objetos, características e vizinhança entre objetos é bastante natural para imagens: os objetos são pixels e as labels dependem do tipo de processamento (por exemplo, em algoritmos de binarização, as labels indicariam se um pixel deve ser considerado informação ou ruído), daí porque relaxation labeling é comumente usado para problemas de processamento automático de imagens (Kittler & Illingworth, 1985). Como este trabalho propõe considerar a influência da vizinhança dos nós representando pessoas conectadas sobre as suas opiniões, o modelo é também aderente ao problema aqui tratado.

3.3 Partida a Frio

Resgatando a definição do problema de Classificação Coletiva apresentada na seção 3.1, temos que: "é dado um conjunto de labels Y^K conhecidos para os nós $V^K \in V$, tais que $Y^K = \{y_i | v_i \in V^K\}$. O objetivo da Classificação Coletiva é inferir Y^U, os valores de Y_i para os nós cujos labels ainda são desconhecidos ($V^U = V - V^K$), ou a distribuição de probabilidade sobre esses valores".

Isto é, de forma genérica, o objetivo é calcular $P(Y^U|Y^K)$. Desta forma, o método requer que parte do grafo tenha classificação conhecida para que o conhecimento seja propagado através dos relacionamentos entre os nós (ou seja, o grafo precisa dispor de um conjunto-semente cujas classificações são conhecidas). Ao problema que ocorre quando um sistema depende de um volume suficiente de informação coletada *a priori* (normalmente via intervenção manual), dá-se o nome de partida a frio (*cold start*), que é comum em sistemas de recomendação (Lam et al., 2008), (Park et al., 2006). No contexto de classificação coletiva de opiniões, esse problema se reflete na necessidade de conhecimento prévio das opiniões de parte dos usuários da rede a respeito do tópico sob monitoramento.

Dois problemas relacionados à Partida a Frio são relevantes para este trabalho:

1) Estabelecer as classes dos nós pertencentes ao conjunto-semente;
2) Escolher, de acordo com critérios objetivos, quais nós devem compor o conjunto-semente de forma a maximizar a precisão do processo de Classificação Coletiva. Isto inclui a seleção dos nós propriamente dita, bem como a definição de quantos nós são suficientes e necessários para que a Classificação Coletiva tenha bom desempenho.

   A seguir são detalhados os dois problemas e alternativas de abordagens.

3.3.1 Classificação do Conjunto-Semente

Para obter as classes dos nós que formam o conjunto-semente, uma alternativa imediata seria aplicar algum método de inferência automática. Obviamente, para a classificação do conjunto-semente, não são aplicáveis métodos de classificação coletiva. As melhores opções são os classificadores que usam apenas informações locais dos nós, como os algoritmos tradicionais de análise de sentimento baseados em processamento de texto, descritos no capítulo 2.

Independente do método de classificação escolhido, é imprescindível que a precisão da classificação seja alta, uma vez que é o conhecimento do conjunto-semente que será propagado para toda a rede. Uma classificação inicial mal feita comprometerá todo o processo de Classificação Coletiva que virá a seguir, pois conhecimento incorreto será propagado. Como apresentado

na seção 2.5, os algoritmos de classificação de sentimento baseados em processamento de texto estão sujeitos a uma série de complicadores que tornam difícil a obtenção de altas taxas de precisão, especialmente em ambientes sem controle sobre os textos postados como as redes sociais.

Uma alternativa natural seria a classificação manual dos textos postados, que permitiria a análise do conteúdo por operadores potencialmente especializados e produziria resultados com boa precisão. Esta abordagem apresenta problemas inerentes ao processamento manual, como sujeição a erros por cansaço dos operadores, falta de uniformidade na análise dos textos (já que a interpretação fica a cargo de cada operador), entre outros. Entretanto, o grande problema da abordagem de classificação manual é que ela apresentaria custos proibitivos em aplicações de larga escala. Este custo pode ser minorado se for feita uma seleção criteriosa de quais nós devem ser analisados, conforme detalha a seção a seguir.

3.3.2 Seleção do Conjunto-Semente

Observe que, quanto menor o número de nós necessários para uma boa classificação coletiva, mais opções se tornam disponíveis para a realização da classificação inicial (por exemplo, a marcação manual passa a ser uma opção viável se poucos nós forem necessários no conjunto-semente). Se for possível identificar quais nós do grafo, uma vez que tenham sua classe conhecida, produzem o melhor resultado de classificação coletiva, a classificação desses nós específicos pode ser feita sob demanda, economizando recursos sem prejudicar a qualidade do resultado produzido.

O problema acima descrito é semelhante àquele tratado pela área de Aprendizagem Ativa (Bilgic, Mihalkova, & Getoor, 2010): num cenário em que os exemplos de treinamento são abundantes, porém sua marcação é custosa, escolher adequadamente quais instâncias devem ser usadas para treinamento do modelo resulta em classificadores muitas vezes com qualidade superior, necessitando de menos exemplos de treinamento. Entretanto, no nosso caso, o modelo aqui usado já existe e não será retreinado conforme os exemplos sejam selecionados. Porém, de forma análoga ao que acontece em Aprendizagem Ativa, a escolha adequada de quais nós devem ser inicialmente classificados impacta no resultado que a classificação coletiva produzirá. Esse problema é

estudado pela área de Inferência Ativa (Rattingan, Maier, & Jensen, 2007), (Bilgic & Getoor, 2010).

No domínio de Classificação Coletiva, o princípio da Inferência Ativa pode ser aplicado respondendo à seguinte pergunta: que características apresentadas pelos nós do grafo são bons discriminantes para se determinar aqueles nós que devem ser incluídos no conjunto-semente? Isto é, o objetivo aqui é escolher quais nós, formando o conjunto inicial, produzirão o melhor resultado para a classificação coletiva, uma vez que a seleção informada de nós iniciais produz melhores resultados do que uma seleção aleatória (Rattingan, Maier, & Jensen, 2007).

A seleção dos nós do conjunto-semente deve obedecer a dois critérios fundamentais:

(1) Os nós devem possuir caminhos curtos a muitos outros nós da rede;
(2) Os nós devem ser bem distribuídos na rede.

Isto é, idealmente a seleção inicial deve conter nós que são localmente influentes e globalmente dispersos.

Na seção 5.3, em que são detalhados os experimentos realizados para a seleção de nós iniciais seguindo estes métodos, a heurística usada para seleção dos nós foi baseada no grau, explorando o fato de que nós com muitas conexões são localmente influentes (Kempe, Kleinberg, & Tardos, 2003). Os resultados obtidos foram comparados com uma seleção aleatória, e confirmaram a expectativa de que a seleção informada produz melhores taxas de acerto.

3.4 Trabalhos Relacionados

A ideia de usar informação relacional no processo de classificação da análise de sentimento não é nova, porém anteriormente o foco esteve no processo de classificação de sentimento tradicional (baseado em processamento de texto), aplicando técnicas que aproveitam a informação relacional entre documentos, sentenças ou características dos objetos, e não entre as pessoas que emitiram as opiniões, como é a proposta deste trabalho. A seguir serão apresentadas mais informações sobre essas abordagens.

3.4.1 Uso de Informação Relacional entre Documentos ou Sentenças

Documentos longos (formados por vários parágrafos ou sentenças) normalmente possuem opiniões sobre diferentes características do objeto de interesse. Para tratar esses casos, foram desenvolvidas técnicas que modelam um documento como uma série de subdocumentos, usando explicitamente os relacionamentos entre esses subdocumentos para alcançar melhores taxas de classificação. Um exemplo de algoritmo desenvolvido dessa forma (Pang & Lee, 2004) executa um procedimento dividido em duas etapas para classificação de críticas sobre filmes: primeiro, o algoritmo divide o documento em sentenças e as classifica como objetivas ou subjetivas. As sentenças objetivas são descartadas, sob a prerrogativa de que não possuem informação opinativa[15]. No segundo passo, as sentenças classificadas como subjetivas passam por um processo de classificação de orientação. A informação relacional entre sentenças é usada no processo de classificação de subjetividade da seguinte forma: assumindo a premissa de que normalmente um documento não varia abruptamente entre sentenças objetivas e subjetivas, o algoritmo considera que sentenças próximas tem maior probabilidade de receber a mesma classificação quanto à subjetividade. Essa informação é considerada no processo, que é executado por um algoritmo de classificação coletiva.

3.4.2 Uso de Informação Relacional entre Características

Outra aplicação de uso de informação relacional ocorre em trabalhos que verificam o relacionamento entre características dos objetos. Um trabalho que empregou esta abordagem com sucesso (Snyder & Barzilay, 2007) tem por objetivo classificar avaliações sobre cada característica de um determinado objeto (por exemplo, acomodações e localização de um hotel) a partir do texto escrito pelos usuários. É construído um classificador linear para predizer se todas as características de um produto recebem a mesma avaliação, e o resultado desse processo é combinado com os classificadores individuais de cada característica para minimizar uma função de perda. Num exemplo fornecido no trabalho, vê-se que os classificadores independentes não tem boa precisão se empregados isoladamente, porém, ao incluir o processo de

[15] Apesar de, como mencionado na seção 2.1.10, ser possível que sentenças objetivas contenham opiniões implícitas.

classificação por concordância entre as características, a precisão fica próxima ao *gold standard*.

3.4.3 Uso de Informação Relacional entre Pessoas

Além do relacionamento entre sentenças, documentos e características, há algoritmos que usam informação relacional entre pessoas, porém num cenário diferente do tratado neste trabalho. Por exemplo, num estudo realizado em redes de newsgroups (Agrawal et al., 2003) sobre três tópicos diferentes (imigração, aborto e controle de armas), observou-se que a relação entre dois usuários na rede que mapeia as respostas a uma determinada postagem tende a ser antagônica: em média, 74% das respostas eram contrárias à opinião emitida no tópico original, enquanto apenas 7% eram coincidentes. Assumindo que os links dessa rede que mapeia as respostas implicam normalmente desacordo de opiniões, foi possível classificar os usuários através de algoritmos de particionamento de grafos com uma precisão superior àquela obtida por algoritmos que aplicavam apenas técnicas de processamento de texto.

3.4.4 Técnicas de Mineração de Links

A Classificação Coletiva é uma área de pesquisa que foca exatamente no ponto-chave deste trabalho, pois se baseia no uso dos relacionamentos entre instâncias para criar modelos de classificação para nós com classe desconhecida num grafo. Dessa forma, ela pode ser entendida como parte de Mineração de Links (Getoor & Diehl, 2005), uma área de pesquisa recente que estuda técnicas para diferentes problemas que envolvem classificação e predição em cenários com instâncias interrelacionadas. A seguir, será apresentada uma breve descrição de cada uma das subáreas que compõem o campo de Mineração de Links, que podem ser agrupadas em três classes: (1) técnicas orientadas a objeto; (2) técnicas orientadas a link e (3) técnicas orientadas a grafo.

3.4.4.1 Técnicas Orientadas a Objeto

Agrupam as técnicas da área de Mineração de Links cujo foco está nos objetos, isto é, nos nós que representam o grafo. Além da classificação de objetos baseada em links, onde se enquadram as técnicas de classificação coletiva, pertencem a esta categoria:

- Algoritmos de Ordenamento de Objetos Baseado em Links, cujo objetivo é explorar a estrutura de links de um grafo para ordenar ou priorizar o conjunto de objetos representado pelo grafo. A maior parte dos trabalhos nessa área se concentra em grafos com um único tipo de objeto e de link, e as principais técnicas desta subárea incluem os difundidos algoritmos de PageRank e HITS. O PageRank (Page et al., 1999) modela a navegação na web como uma visitação aleatória (*random walk*) em que o usuário seleciona e segue links aleatoriamente, e eventualmente reinicia o processo a partir de uma nova página. O rank de uma dada página é a fração de tempo que o usuário passaria na página se o processo continuasse indefinidamente. O HITS (Kleinberg, 1999) modela a web de uma forma mais complexa, considerando que há dois tipos de páginas: *hubs* e *authorities*. Hubs são páginas com links para muitas páginas do tipo *authorities*, e *authorities* são páginas apontadas por muitos hubs. A cada página da web se atribui uma pontuação para *hub* e para *authority*, que são calculadas por um algoritmo iterativo que atualiza as pontuações de uma página de acordo com as pontuações das páginas na sua vizinhança imediata. No domínio de análise de redes sociais, o ordenamento por links é uma tarefa central, cujo objetivo é ordenar indivíduos numa dada rede de acordo com sua importância (centralidade). Medidas de centralidade são objetos de estudo na área de análise de redes sociais há muito tempo (Wasserman & Faust, 1994). Essas medidas caracterizam algum aspecto da estrutura local ou global de acordo com a posição de um determinado indivíduo na rede, variando em termos de complexidade entre medidas locais, como grau de centralidade, até medidas globais como autovetores.

- Algoritmos de Detecção de Grupos, cujo objetivo é agrupar os nós do grafo em grupos que possuem características em comum. Várias técnicas foram propostas com esse foco, mas ultimamente o maior desafio é desenvolver métodos escaláveis que possam explorar grafos complexos, como aqueles usados para representar redes sociais. No domínio de análise voltada para redes sociais, há uma técnica chamada de *blockmodeling* (Ferligoj, Doreian, & Batagelj, 2005), que envolve o particionamento da rede em conjuntos de indivíduos que exibem

conjuntos similares de links com outros indivíduos. Uma medida de similaridade é definida entre esses conjuntos de links, e um algoritmo de agrupamento é usado para identificar posições. Métodos de partição espectral de grafos abordam o problema de detecção de grupos através da identificação de um conjunto aproximadamente mínimo de links que devem ser removidos do grafo para que seja obtido um determinado número de grupos (Newman, 2004).

- Algoritmos de Resolução de Entidades, cujo objetivo é determinar quais citações nos dados se referem aos mesmos objetos reais. Exemplos deste problema estão em bancos de dados (Ananthakrishna, Chaudhuri, & Ganti, 2002), (Culotta & McCallum, 2005) (remoção de duplicidade e integração de dados), processamento de linguagem natural (Li, Morie, & Roth, 2005) (resolução de correferência, consolidação de objetos), gerenciamento de informações pessoais e em vários outros campos. Normalmente, a resolução de entidades é vista como um problema de resolução par a par, em que cada par de referências é independentemente resolvido como sendo correferente ou não, dependendo da similaridade entre seus atributos (Kalashnikov, Mehrotra, & Chen, 2005). Recentemente, tem crescido o interesse sobre o uso de links para melhorar o resultado da resolução de entidades (Bhattacharya & Getoor, 2006). A ideia central é considerar, além dos atributos das referências que serão resolvidas, as outras referências às quais elas são ligadas. Alguns exemplos desses links são entre os coautores em referências bibliográficas (Pasula et al., 2003).

3.4.4.2 Técnicas Orientadas a Link

Esta categoria inclui apenas uma técnica: a Predição de Links, que é o problema de inferir a existência de um link entre dois objetos, de acordo com os dados observados (atributos dos objetos e links existentes). Entre os exemplos de aplicação, podem-se citar a predição de links entre os usuários de redes sociais (como inferência de amizade), predição de participação de atores em eventos (O'Madadhain, Hutchins, & Smyth., 2005), ou a predição de relações semânticas como “orientador de” baseada em links e conteúdo de páginas web (Craven et al., 2000). Normalmente, o problema se configura de uma das seguintes formas:

1) Alguns links são observados e se deseja inferir outros links;
2) É considerado o conjunto de links no tempo *t* e se deseja predizer que links existirão no tempo *t* +1.

A Predição de Links é normalmente vista como uma classificação binária simples: para cada dois objetos o_i e o_j, o objetivo é determinar se existe um link l_{ij} que conecta os objetos. Uma possível abordagem é fazer a predição baseada apenas nas propriedades estruturais da rede (Liben-Nowell & Kleinberg, 2003), entretanto também existem abordagens que usam outras informações, como um modelo de regressão logística que usa características relacionais (Popescul & Ungar, 2003).

3.4.4.3 Técnicas Orientadas a Grafo

É o conjunto de técnicas de Mineração de Links que focam no grafo como um todo. Nesta categoria, as duas principais técnicas são:

- Descoberta de Subgrafos, cujo objetivo é encontrar subgrafos de interesse ou comuns num conjunto de grafos. A descoberta desses padrões pode ser o objetivo final de uma aplicação, ou pode ser usada como entrada para classificação de grafos (apresentada logo a seguir). Em relação aos trabalhos que tentam encontrar os subgrafos mais frequentes, muitas das abordagens usam a propriedade Apriori[16] (Agrawal & Srikant, 1994). Outros trabalhos focam na aplicação de heurísticas, seja para busca de subgrafos (Cook & Holder, 1994), seja para compressão do grafo de entrada por meio de eliminação de pares de vértices frequentes (Yoshida, Motoda, & Indurkhya, 1994). Ambas as abordagens usam busca local gulosa para encontrar subestruturas frequentes.

- Classificação de Grafos, cujo objetivo é classificar um grafo como uma instância negativa ou positiva de acordo com um determinado conceito. A classificação de grafos normalmente não requer inferência coletiva, como é necessário na classificação de objetos e arestas, pois normalmente se assume que cada grafo como um todo é uma instância

[16] Diz que todos os subconjuntos não vazios de um conjunto de itens frequentes também são frequentes (Han & Kamber, 2006).

independente. Há três abordagens bastante usadas classificação de grafos:

- Mineração de Características em Grafos, que utiliza métodos da Descoberta de Subgrafos (apresentados anteriormente), e normalmente consiste em encontrar as subestruturas frequentes ou com informação relevante nos grafos de entrada. Essas subestruturas são usadas para transformar o grafo numa representação em formato de tabela, e classificadores tradicionais são usados para classificar as instâncias;

- Programação em Lógica Indutiva. Um exemplo de abordagem que usa programação em lógica indutiva (King et al., 1996) mapeia os dados de grafos numa representação relacional, depois aplica uma representação lógica sobre esses dados[17], e então usa um sistema de programação em lógica indutiva para encontrar hipóteses neste espaço;

- Definição de Kernels para Grafos. As técnicas baseadas em métodos de kernel (Kashima & Inokuchi, 2002) tem custo computacional menor, e assim são necessárias quando a busca por todas as subestruturas frequentes apresenta um custo computacional muito alto, o que acontece comumente em aplicações reais, principalmente quando se trata de mineração em grafos complexos (como aqueles que representam redes sociais).

3.5 Considerações Finais

Neste capítulo, foi destacado que modelagens tradicionais de análise de sentimento não são adequadas para problemas de classificação em cenários com informação relacional disponível, pois não fazem uso de uma importante informação que está disponível nesses cenários, que é o relacionamento entre as instâncias, e podem inclusive levar a conclusões imprecisas sobre os dados (Jensen D. , 1999). Por este motivo, a abordagem proposta neste trabalho

[17] Por exemplo, uma representação que define relações como vértice(idGrafo, idVertice, labelVertice, atributosVertice) e aresta(idGrafo, idVertice1, idVertice2, labelAresta),

procura explorar as informações adicionais providas pelos relacionamentos entre os objetos, de acordo com o princípio da homofilia, aplicando técnicas de Classificação Coletiva sobre o grafo que representa as instâncias que serão classificadas e suas interconexões.

O problema da Classificação Coletiva foi caracterizado, e as suas principais técnicas foram detalhadas. Foram discutidos ainda os aspectos relacionados à partida a frio, intrinsecamente ligados à Classificação Coletiva. Além disso, foi apresentada uma revisão superficial de técnicas relacionadas à área de Mineração de Links, que poderiam ser aplicadas em cenários semelhantes ao que este trabalho trata. Por fim, foi feita uma breve descrição de outras abordagens de análise de sentimento que fazem uso de informação relacional. No próximo capítulo, será apresentada em detalhes a abordagem proposta neste trabalho.

4 Classificação Coletiva Aplicada à Análise de Sentimentos em Redes Sociais

Como já mencionado em capítulos anteriores, a web (ou, mais especificamente, sites como aqueles dedicados à avaliação de produtos ou as redes sociais online) contem informação altamente valiosa para empresas, entidades públicas, políticos, personalidades ou quem quer que tenha interesse em monitorar opiniões ou mensurar a aceitação de um objeto. Eles oferecem uma forma de saber rapidamente a opinião do público sobre virtualmente qualquer tema de interesse, uma clara vantagem sobre pesquisas de opinião usuais, que são demoradas, trabalhosas, tem alto custo e alcançam um número bastante restrito de pessoas. Com o uso de ferramentas apropriadas, todo o processo de consolidação de opiniões pode ser realizado de forma automática, oferecendo um *feedback* imediato e mais representativo sobre o objeto avaliado.

A abordagem clássica usada para realizar o processamento dessa informação disponível online é baseada no processamento textual das opiniões dos usuários. Essa abordagem, entretanto, apresenta algumas limitações e complicadores de difícil tratamento, como o uso de construções gramaticais complexas, figuras de linguagem, ironia, sarcasmo, ambiguidade inerente à linguagem natural, resolução de contexto, etc (consulte o capítulo 2 para uma lista completa e um maior detalhamento sobre tais complicadores). Quando a análise de sentimento é realizada num cenário em que há informação relacional disponível (como é o caso das redes sociais online), métodos para aproveitamento dessa nova dimensão de informação são necessários, pois métodos tradicionais baseados em processamento do conteúdo textual das opiniões negligenciam esse tipo de informação.

A abordagem aplicada neste trabalho para realização de análise de sentimento em redes sociais é baseada no princípio da homofilia (Macskassy & Provost, 2007), que estabelece que indivíduos tendem a se conectar a outros com os quais apresentam alto grau de semelhança. A partir desse conceito, verificado em diversos casos na literatura (Yuan & Gay, 2006), (McPherson, Smith-Lovin, & Cook, 2001), (Bollen et al., 2011), o método proposto neste trabalho aplica técnicas de classificação coletiva sobre o grafo que representa a rede social. Para que a aplicação de um algoritmo de Classificação Coletiva

sobre um grafo seja possível, é necessário que o grafo seja parcialmente classificado (isto é, possua um conjunto inicial de nós com classes conhecidas). A classificação do conjunto inicial pode ser realizada de várias formas, inclusive manualmente. Este trabalho, contudo, utiliza classificadores textuais para produzir de forma automática esse conjunto inicial de nós classificados a partir do conteúdo das mensagens postadas nas redes sociais acerca de um determinado tema sob monitoramento. Uma arquitetura geral para sistemas que aplicam a técnica empregada neste trabalho é apresentada na figura abaixo:

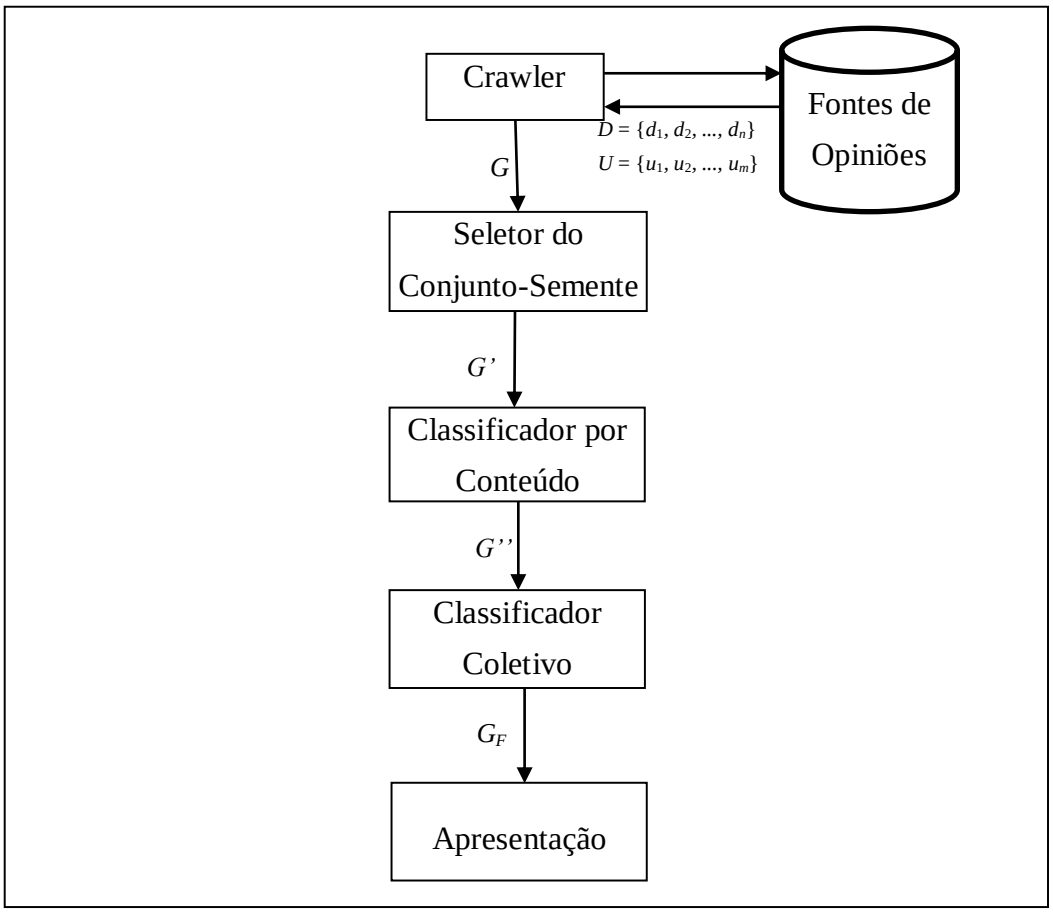

Figura 5: Macroarquitetura da abordagem proposta.

A função de cada um dos módulos da arquitetura acima é descrita abaixo:

1) Crawler: responsável por coletar postagens de uma rede social a partir de *hashtags* que servem de *gold standard* para as classes e por capturar as informações de relacionamento entre os usuários de uma rede social. A partir desses dados, o crawler monta o grafo *G* que será processado pelas demais etapas do algoritmo;

2) Seletor do Conjunto-Semente: tem o objetivo de escolher os nós que comporão o conjunto-semente de acordo com informações estruturais do grafo, seguindo os princípios da inferência ativa. O grafo que representa a rede social (ou o fragmento da rede social que tem as informações sobre o monitoramento realizado, no caso deste trabalho) é alterado, com a indicação dos nós que fazem parte do conjunto-semente;
3) Classificador por Conteúdo: responsável por realizar a classificação dos nós do conjunto-semente usando um classificador textual (tratamento para o problema da partida a frio). O grafo é mais uma vez modificado, agora recebendo as informações de classe para os nós do conjunto-semente;
4) Classificador Coletivo: módulo que aplica um algoritmo de classificação coletiva (como *relaxation labeling* - consulte a seção 3.2.2.2) para realizar a classificação coletiva dos demais nós autores do grafo a partir do conhecimento disponível no conjunto-semente, produzindo a versão final do grafo com todas as classificações realizadas;
5) Apresentação: representa um módulo genérico que é responsável por encaminhar o resultado final. Normalmente, seria uma interface gráfica para apresentação dos resultados ao usuário, mas poderia ser também uma interface de integração com outros sistemas, armazenamento do resultado para processamento/apresentação posterior, etc. Este trabalho não trata do uso que se faz sobre o grafo resultante do processamento e assim não aborda este módulo em detalhes.

Para avaliação da viabilidade do método proposto, foi implementado um protótipo que segue a arquitetura apresentada anteriormente (mais detalhes sobre cada um dos módulos implementados no protótipo podem ser encontrados nas subseções a seguir deste capítulo) e desenvolvido um estudo de caso para predição de preferência política de usuários do Twitter. Como mostram os resultados dos experimentos realizados (consulte o capítulo 5), esta abordagem permite realizar a análise de sentimento em redes sociais com maior precisão do que as técnicas que dependem exclusivamente da análise textual no cenário avaliado, pois a classificação baseada em conteúdo possui desafios

difíceis de contornar (como ironia, sarcasmo, resolução de contexto, apresentadas em detalhes na seção 2.5). O método proposto não é afetado por essas dificuldades, uma vez que é insensível a elas, pois não realiza análise sobre o conteúdo textual. Além disso, pela análise de relacionamento entre os nós de uma rede, é possível classificar a opinião de usuários que nem se manifestaram explicitamente sobre um determinado tema, o que amplia sobremaneira a cobertura da análise para aplicações reais, pois há uma parcela significativa de usuários que consomem informação mas não produzem conteúdo em volume suficiente para a aplicação de análise textual (Mustafaraj et al., 2011). Obviamente, este fato não pode ser aproveitado neste trabalho, pois usuários que não se manifestaram não tem classe conhecida e assim a precisão de classificação sobre eles não pode ser medida. A análise dos relacionamentos entre os usuários permite ainda reforçar a pontuação de confiança dada por classificadores textuais, ou até modificar a classificação de usuários que postaram opiniões em redes sociais.

A seguir, são apresentados detalhes dos quatro módulos que compõem a macroarquitetura do sistema apresentada na Figura 5.

4.1 Crawler

Este módulo é responsável por coletar postagens de redes sociais, juntamente a informações de relacionamento envolvendo os usuários autores das postagens coletadas (isto é, quais os usuários seguidores e os usuários seguidos por cada um dos autores). O resultado desta coleta é um grafo que contem todos os usuários autores, todos os usuários que seguem ou são seguidos pelos autores, e todas as arestas que representam cada relação de seguimento.

Este processo de construção normalmente produz um grafo com um conjunto-semente muito pequeno em relação ao tamanho total, o que impossibilita a propagação do conhecimento (isto é, as classes conhecidas do conjunto-semente) ao longo da rede. No experimento detalhado no capítulo 5, por exemplo, o número de usuários autores era de cerca de 1200, e com a agregação das informações de seguimento o grafo resultante ficou com aproximadamente 100 mil nós e 1 milhão de arestas. Isto significa que, mesmo se todas as classes dos nós autores fossem conhecidas, o conjunto-semente representaria aproximadamente 1% do total dos nós do grafo.

4.1.1 *Pruning* do Grafo

Obviamente, a aplicação de um algoritmo de classificação coletiva nesse tipo de grafo não é factível. Por isso, é aplicado um método de *pruning* automático do grafo, com o objetivo de eliminar o maior número possível de nós desconhecidos, ainda que mantendo uma quantidade mínima necessária para não quebrar a conectividade do grafo, e assim possibilitar a aplicação posterior de um algoritmo de classificação coletiva.

O algoritmo de *pruning* utilizado elimina o mínimo possível de nós autores (ou seja, nós com classe conhecida), reduzindo o grafo através da eliminação iterativa de usuários não-autores, desde que essa eliminação não quebre a conectividade do grafo. O seu funcionamento é ilustrado através do código abaixo:

```
/* São ordenados ascendentemente os nós de acordo com a
quantidade de conexões que possuem */
sortByDegree(nodes);

while (true) {
   /* se o número mínimo de nós desconhecidos for
atingido, retornar o grafo obtido */
   if (nodes.numUnknown() == MIN_UNKNOWN_COUNT)
      return nodes;
   /* selecione o primeiro nó desconhecido (ie, não
autor) */
   n = nodes.firstUnknown();
   newNodes = nodes.remove(n);
   /* remover o nó se isso não quebrar a conectividade
do grafo */
   if (isConnected(newNodes))
      nodes = newNodes;
   else
      return nodes;
}
```

Algoritmo 5: *Pruning* do grafo que representa os usuários de uma rede social.

4.2 Seleção do Conjunto-Semente

Conforme já citado anteriormente, algoritmos de classificação coletiva fazem a sua inferência com base num conjunto inicial de nós com classificação conhecida (o conjunto-semente) e as informações relacionais da rede. Dessa forma, para viabilizar a aplicação de classificação coletiva, é necessária alguma forma de definir quais nós comporão o conjunto-semente.

Se todos os nós autores tiverem classe conhecida, obviamente o ideal é que todos eles sejam incluídos no conjunto-semente. Isto incluirá mais conhecimento para ser propagado ao longo da rede. Entretanto, como normalmente se tem a intenção de reduzir o tamanho do conjunto-semente ao mínimo possível (seja pela dificuldade, pelo custo ou pelo tempo necessário para realizar a classificação – automática ou manual – com base nos textos dos usuários), é necessário que haja uma forma de selecionar alguns nós entre os autores que devem ser classificados. Uma abordagem imediata seria realizar uma seleção aleatória entre os nós com classe conhecida. Se o conjunto resultante tiver um tamanho suficiente, a classificação coletiva pode apresentar um bom desempenho. Na prática, entretanto, é possível aplicar heurísticas para melhorar a seleção dos nós.

Observe que, quanto menor o número de nós necessários para uma boa classificação coletiva, mais opções se tornam disponíveis para a realização da classificação inicial. Por exemplo, a marcação manual é uma opção viável se poucos nós forem necessários no conjunto-semente. Se for possível identificar quais nós do grafo, uma vez que tenham sua classe conhecida, produzem o melhor resultado de classificação coletiva, a classificação desses nós pode ser feita sob demanda, economizando recursos sem prejudicar a qualidade do resultado produzido.

O problema acima descrito é semelhante àquele tratado pela área de aprendizagem ativa (*active learning*) (Bilgic, Mihalkova, & Getoor, 2010): num cenário em que os exemplos de treinamento são abundantes, porém sua marcação é custosa, escolher adequadamente quais instâncias devem ser usadas para treinamento do modelo resulta em classificadores muitas vezes com qualidade superior, necessitando de menos exemplos de treinamento. Entretanto, no nosso caso, o modelo aqui usado já existe e não será retreinado

conforme os exemplos sejam selecionados. Porém, de forma análoga ao que acontece na aprendizagem ativa, a escolha adequada de quais nós devem ser inicialmente classificados impacta no resultado que a classificação coletiva produzirá. Este problema é estudado pela área de inferência ativa (*active inference*) (Rattingan, Maier, & Jensen, 2007), (Bilgic & Getoor, 2010).

4.2.1 Seleção Baseada em Informações Estruturais do Grafo

O mesmo princípio que norteou o desenvolvimento deste trabalho pode se aplicar à fase de seleção dos nós do conjunto-semente: assim como se podem usar informações relacionais do grafo para realizar a classificação dos usuários da rede, é possível usar essas mesmas informações para auxiliar na seleção dos nós que formam o conjunto-semente. A ideia aqui é escolher quais nós, formando o conjunto-semente, produzirão o melhor resultado para a classificação coletiva, uma vez que a seleção informada de nós iniciais produz melhores resultados do que uma seleção aleatória (Rattingan, Maier, & Jensen, 2007). Os principais critérios para a escolha de tais nós devem obedecer a dois princípios fundamentais:

(1) Os nós do conjunto-semente idealmente possuem caminhos curtos a muitos outros nós da rede;
(2) Esses nós precisar estar bem distribuídos na rede.

Isto é, idealmente a seleção inicial deve conter nós que são localmente influentes e globalmente dispersos.

Na seção 5.3, em que são detalhados os experimentos realizados para a seleção de nós do conjunto-semente seguindo estes métodos, a heurística usada para seleção dos nós foi baseada no grau, explorando o fato de que nós com muitas conexões são localmente influentes (Kempe, Kleinberg, & Tardos, 2003). Os resultados obtidos foram comparados com uma seleção aleatória, e confirmaram a expectativa de que a seleção informada produz melhores taxas de acerto.

4.2.2 Seleção Baseada em Classificação Textual

Em aplicações práticas, entretanto, normalmente não é suficiente considerar apenas informações estruturais, e dessa forma outro critério precisa ser observado para a seleção dos nós do conjunto-semente. Considere um cenário

real em que é necessário classificar os usuários de uma rede social de acordo com um determinado conjunto de classes. Na maioria dos casos, não há informação a priori sobre as classes de nenhum nó da rede, e assim pouco adianta identificar os nós localmente influentes e globalmente dispersos; a classe desses nós também é desconhecida e a aplicação da classificação coletiva é inviabilizada.

Dessa forma, é necessário primeiro avaliar quais nós podem ter a classificação inferida, o que é uma tarefa simples: são os nós autores, cujas mensagens foram coletadas na fase de *crawling* do sistema. Sobre essas mensagens, aplica-se algum método de classificação automática para inferir a orientação dos respectivos usuários. Como apontado anteriormente nesta tese, a classificação textual apresenta várias dificuldades difíceis de contornar (consultar seção 2.5), e assim é possível que a precisão obtida sobre a classificação neste passo seja insuficiente para uma boa classificação coletiva, pois o erro de classificação do conjunto-semente seria propagado ao longo da rede.

Entretanto, o objetivo neste passo não é usá-la para classificação de toda a rede, e sim para realizar apenas a seleção do conjunto inicial de usuários. Dessa forma, é possível usar heurísticas que permitam considerar as postagens com maior probabilidade de terem sido corretamente classificadas. Na seção 5.2, que apresenta os experimentos realizados com este método de seleção dos nós iniciais, detalha-se o uso de dois parâmetros que restringem quais postagens serão usadas efetivamente para definição do conjunto-semente: a pontuação mínima de confiança para cada classificação de postagem baseada em conteúdo, e o número mínimo de postagens por autor. Perceba que o uso da informação de confiança sobre a classificação textual combina a fase de seleção com a fase seguinte (classificação do conjunto-semente).

4.3 Classificação do Conjunto-Semente

Definidos os nós que compões o conjunto-semente, é preciso classificá-los para que posteriormente um algoritmo de classificação coletiva propague o conhecimento ao longo dos demais nós da rede. Há várias formas de realizar esta classificação:

a) Marcação manual. Sob o ponto de vista de qualidade, esta é sem dúvida a melhor opção, assumindo que os operadores possuem conhecimento suficiente para desempenhar esta atividade. Entretanto, seu alto custo e o longo tempo necessário para a marcação de uma quantidade suficiente de exemplos pode inviabilizar esta opção em aplicações reais (Bilgic, Mihalkova, & Getoor, 2010);

b) Uso de *hashtags*. Usada neste trabalho para montagem da base de experimentos[18], esta é uma boa alternativa para aplicações reais. Entretanto, só é aplicável a cenários nos quais seja usual o uso de *hashtags* indicativas de preferência (como no caso de política, estudado neste trabalho). Além disso, esta abordagem não é completamente automática, (pode ser entendida como semi-automática), pois envolve a identificação das *hashtags*, que normalmente é feita de forma manual.

c) Algoritmos de classificação textual, conforme o experimento detalhado na seção 5.2 deste trabalho. Esta abordagem tem um possível problema quanto à qualidade dos classificadores. No experimento da seção 5.2, foi treinado um classificador específico para o domínio tratado (uma vez que o foco é a classificação coletiva), porém a criação de um algoritmo que possa ser aplicado para qualquer domínio tende a produzir resultados menos precisos.

A seguir são apresentados os principais algoritmos de classificação textual usados neste trabalho.

4.3.1 Classificadores Disponíveis Online

Foram avaliadas APIs online para processamento de texto. A que apresentou melhores resultados foi o uClassify[19], que permite realizar uma série de operações sobre texto, como categorização de textos em tópicos pré-definidos, identificação do idioma de um texto, classificação de sentimento, classificação de humor (que identifica o estado de humor da pessoa ao escrever determinado texto, e não deve ser confundida com a classificação de sentimento), e até mesmo classificação de gênero ou idade do autor. Entretanto, esta API teve sua aplicação inviabilizada neste trabalho por dois motivos:

[18] Consulte a seção 5.1.1 para mais detalhes

[19] www.uclassify.com

1) A ferramenta impõe uma restrição sobre o número de chamadas por dia, o que dificulta a realização de experimentos;
2) Ausência de documentação científica sobre a ferramenta, o que implicaria num módulo "caixa preta" neste trabalho. A documentação disponível trata somente de informações técnicas, como integração, arquitetura ou aspectos transacionais, e apenas cita superficialmente que o núcleo do algoritmo é um classificador Naïve Bayes, com alguns passos adicionais que supostamente melhoram o processo de classificação, porém os detalhes são omitidos.

Essas razões impossibilitam o uso deste tipo de API online, pelo menos para trabalhos científicos em que é necessário pleno entendimento sobre todas as etapas do processo.

4.3.2 Classificadores da Plataforma Weka

Outra alternativa avaliada foi a criação de classificadores baseados na plataforma Weka (Witten, Frank, & Hall, 2011). Foram usados vários classificadores diferentes, sendo os que obtiveram os melhores resultados: árvores de decisão C4.5 (Quinlan, 1993), SVM (Cortes & Vapnik, 1995), Classificação Via Clustering (um algoritmo de metaclassificação que usa k-*means* para classificação) e Florestas Aleatórias (*Random Forests*) (Breiman, 2001) (este último, o que obteve a melhor precisão entre todos os algoritmos avaliados dentro da plataforma Weka).

A implementação de um classificador textual no Weka consiste em criar instâncias de treinamento a partir do texto, convertendo a *stream* de texto em características (palavras), filtrando *stopwords*[20] e aplicando redução ao radical (*stemming*) (Porter, 1980) no processo de tokenização. Os resultados produzidos por esses classificadores foram razoáveis, atingindo uma precisão de cerca de 70% nos melhores casos.

Outra opção de classificador textual identificada foi o TagHelper (Rose et al., 2007), que também é construído sobre o Weka, porém não está disponível na sua distribuição padrão. Este classificador apresentou taxas de precisão superiores às obtidas pelos classificadores padrão do Weka. Devido ao seu

[20] Palavras comuns que não possuem sentido semântico (como preposições e artigos).

desempenho superior, este classificador foi utilizado nos experimentos apresentados no capítulo 5. O TagHelper realiza o tratamento de pontuação e permite o uso de radicais (*stemming*), unigramas e bigramas[21] para definir as caraterísticas das instâncias de classificação, além de suportar bigramas morfológicos[22]. A classificação das instâncias é realizada através do método Naïve Bayes (John & Langley, 1995).

4.4 Classificação Coletiva

Antes de entrar em detalhes sobre o algoritmo de classificação coletiva usado neste trabalho, considere novamente as definições e notações apresentadas no capítulo 3, repetidas a seguir para facilidade de referência: o problema de classificação coletiva pode ser formalmente definido por $G(V; E)$, um grafo representando uma rede social composta de um conjunto $V=\{u_1, u_2, \ldots, u_n\}$ de usuários e um conjunto de arestas $E = \{(u_i, u_j)\} \forall\, u_i \in V$ e $u_j \in V$ | existe uma conexão entre u_i e u_j em G. Seja $K \subset V$ um subconjunto de usuários com orientação de opinião conhecida (isto é, os usuários marcados). Seja $Y = V - K$ o conjunto de usuários não classificados e $C=\{c_1, c_2, \ldots, c_m\}$ o conjunto de possíveis classes. A tarefa em questão é a criação de um modelo que seja capaz de classificar o conjunto de usuários cuja classificação é desconhecida Y com uma das classes em C usando apenas a informação disponível em G.

Uma vez que o conjunto-semente K de nós com classe conhecida tenha sido definido conforme detalhado na seção 4.3, o próximo passo é a aplicação da classificação coletiva sobre o grafo completo, com o objetivo de obter as classes do conjunto Y de nós cuja classificação ainda é desconhecida. Na nossa implementação, foi utilizado um framework (Macskassy & Provost, 2007) que realiza um processo de classificação sobre o grafo através do uso de algoritmos de inferência locais, relacionais e de classificação coletiva. O seu funcionamento é apresentado em detalhes abaixo:

Seja u_i um nó com classe desconhecida e $\vec{c_i}(t) = \langle c_t^1(t), c_t^2(t), \ldots, c_t^m(t) \rangle$ o vetor de probabilidade de classificações

[21] Unigrams são palavras isoladas, enquanto bigrams são pares de palavras que ocorrem em sequência no texto.

[22] Bigrams morfológicos (*part of speech bigrams*) são pares de classes morfológicas que ocorrem em sequência no texto (ex: VERBO, SUBSTANTIVO).

estimado para o nó u_i na iteração *t*. Cada componente do vetor $c_i^k(t), k = 1, \dots, m$ indica a probabilidade do nó u_i pertencer à classe c_k. No framework usado neste trabalho, essas probabilidades são estimadas iterativamente através da aplicação de três componentes, conforme descrito a seguir:

(a) Classificador local: este classificador é usado para gerar uma classificação inicial $\vec{c}_i(0)$ para os nós não marcados, que será posteriormente refinada pelos componentes de classificação relacional e coletiva. O classificador local pode ser um modelo de aprendizagem de máquina que usa os atributos dos nós para inferir de forma independente as classificações. No nosso caso, como são usadas apenas informações de relacionamento, o classificador local é um modelo simples que retorna a probabilidade a priori estimada a partir do conjunto inicial *K*.

(b) Classificador relacional: recebe como entrada o resultado produzido pelo classificador local e produz uma nova classificação através do uso de relações presentes no grafo. Na nossa implementação, foi usado o classificador relacional de voto ponderado, que estima as probabilidades das classes assumindo a existência de homofilia no grafo. Dado um nó u_i, sua probabilidade de classe depende das classes observadas na sua vizinhança N_i, isto é, $c_i^k = P(c_k | N_i)$. O classificador estima esta probabilidade como sendo a média ponderada das probabilidades das classes dos nós em N_i:

$$P(c_k | N_i) = \frac{1}{Z} \sum_{u_j \in N_i} w_{i,j} * P(c_k | N_j) \quad (1)$$

onde Z é um fator de normalização e w_{ij} é o peso da aresta que conecta os nós u_i e u_j.

(c) Inferência coletiva: este classificador aplica o classificador relacional iterativamente para classificar coletivamente todos os nós do grafo. Este algoritmo aproximadamente maximiza a probabilidade conjunta das classes de todos os nós no grafo cujas

classes eram inicialmente desconhecidas. Um algoritmo de *relaxation labeling* (Chakrabarti et al., 1998) (consultar a seção 3.2.2.2) foi usado neste passo, escolhido em detrimento a outros métodos de inferência coletiva (Sen et al., 2008) por apresentar melhores resultados nos experimentos. Isto provavelmente acontece porque, diferentemente de outros métodos, *relaxation labeling* não trata o grafo como se estivesse num estado específico de marcação num determinado ponto; ao invés disso, ele retem a incerteza, mantendo o registro das estimativas de probabilidade (Macskassy S. A., 2007).

4.4.1 Análise de Custo Computacional

O propósito de uma análise de custo computacional é prever o comportamento de um algoritmo (principalmente no que se refere ao tempo de execução) sem que seja necessário implementar tal algoritmo numa plataforma específica (Manber, 1989). Entretanto, normalmente é impossível prever o comportamento exato de um algoritmo, já que já muitos fatores que influenciam. Na análise de custo computacional, definem-se certos parâmetros que são os mais importantes para a análise, enquanto os detalhes de implementação são ignorados. A análise é, dessa forma, apenas uma aproximação. A seguir, serão apresentados os custos computacionais de cada fase do algoritmo detalhado no início desta seção.

4.4.1.1 Classificador Local

Como detalhado anteriormente, o classificador local utilizado apenas estima a probabilidade a priori sobre o conjunto-semente *K*. Dessa forma, seu custo computacional é da ordem $O(|K|)$.

4.4.1.2 Classificador Relacional

Para o classificador relacional, o custo é de, para cada nó com classe desconhecida, realizar o somatório dentro da respectiva vizinhança. Assumindo $n = |Y|$ e $\bar{m}$ é a média de tamanho de cada vizinhança N_i, temos que o custo computacional é de $O(n.\bar{m})$, com $\bar{m} \ll n$ normalmente. No pior caso (em grafos completos[23]), o custo é de $O(n^2)$.

[23] Todos os nós tem conexão com todos os outros nós.

4.4.1.3 Classificador Coletivo

Para facilitar a referência ao algoritmo de relaxation labeling, considere novamente o pseudocódigo apresentado anteriormente na seção 3.2.2.2:

For each $Y_i \in Y$ (etapa de inicialização)

    For each $y_i \in L$

        $b_i(y_i) = 1$

    End for

End for

Repeat

    For each $Y_j \in Y$

        For each $y_j \in L$

            $b_j(y_j) = \alpha\phi_j(y_j)\prod_{Y_i \in N_j \cap Y, y_i \in L} \psi_{ij}^{b_i(y_i)}(y_i, y_j)$

        End for

    End for

Until as mensagens $b_j(y_j)$ não apresentem alterações (etapa de troca de mensagens)

Algoritmo 6: Pseudocódigo para o algoritmo de Relaxation Labeling

Na etapa de inicialização, a operação realizada é apenas uma atribuição de valor, cujo custo computacional é praticamente irrelevante, porém seu custo computacional é de $O(n.l)$, onde $l = |L|$. Entretanto, o real custo do algoritmo está concentrado na fase de troca de mensagens. Para cada ciclo, o custo é de $O(n.l.\bar{m})$. O custo final dependerá de quando o algoritmo convergirá (isto é, quando não houver alterações significativas nas mensagens $b_j(y_j)$). Assumindo que $\bar{c}$ denota o número médio de iterações para convergência, o custo final do algoritmo será de $O(n.l.\bar{m}.\bar{c})$.

4.4.2 Escalabilidade do Método Proposto

Como se vê na análise de custo computacional apresentada na seção anterior, o método proposto apresenta um custo relativamente alto para cenários com uma grande quantidade de nós (por exemplo, se fosse aplicado a redes sociais inteiras ou a monitoramentos específicos mais longos). Dessa maneira, para esse tipo de aplicação, seria necessário desenvolver formas de escalar o método para que seu uso se torne viável. Duas alternativas são aqui propostas:

1) Implementar um mecanismo distribuído para o cálculo de valores independentes (por exemplo, mensagens que dependem de vizinhanças disjuntas poderiam ser calculadas em máquinas diferentes de forma independente);
2) De forma mais genérica, fazer uso de um framework de processamento distribuído como o Hadoop (White, 2009), adequando a implementação dos algoritmos usados seguindo o princípio de MapReduce (Dean & Ghemawat, 2004). Uma possibilidade a ser avaliada neste caso seria a extensão de frameworks específicos para processamento de algoritmos de aprendizagem de máquina de forma distribuída, como o Mahout (Owen et al., 2011), cuja versão atual já possui implementações de algoritmos de processamento de texto, classificação e *clustering*, mas precisaria de extensões para contemplar os algoritmos de classificação coletiva.

4.5 Considerações Finais

Neste capítulo, foi apresentado em detalhes o método proposto para classificação de usuários em redes sociais de acordo com suas opiniões emitidas através de postagens e suas conexões com outros usuários. O método combina classificação baseada em texto com classificação coletiva (explorando as informações relacionais disponíveis em redes sociais). Foram discutidos ainda os aspectos relacionados à seleção dos nós que compõem o conjunto-semente (conjunto inicial de nós com classe conhecida cujo conhecimento é propagado ao longo da rede mediante o algoritmo de classificação coletiva). No próximo capítulo, serão apresentados os experimentos realizados.

5 Estudo de Caso

Neste trabalho, foi desenvolvido um estudo de caso com o objetivo de classificar a preferência política de usuários do Twitter, que é um tópico relevante dado o crescente número de pesquisas com esse foco, bem como a ênfase que campanhas políticas tem dedicado recentemente a mídias sociais (Conover et al., 2011). Análise de sentimentos nesse contexto é importante para medir o interesse de usuários e para revelar tendências no cenário político. A escolha pelo Twitter se deve principalmente à sua popularidade e ao seu potencial de mobilização, sendo uma plataforma popularmente usada para engajamento político (Tumasjan et al., 2010), (Bekafigo & Mcbride, 2013). Além disso, o Twitter é uma plataforma que disponibiliza uma API[24] para a coleta de dados, o que simplifica a implementação do *crawler* (consulte a seção 4.1), necessário para a realização do experimento planejado. Recentes pesquisas na área de análise de sentimentos tem focado cada vez mais na execução de experimentos sobre dados disponíveis em redes sociais (Paltoglou, 2014), principalmente Facebook (Ortigosa, Martín, & Carro, 2014) e Twitter (Martinez-Camara et al., 2014).

5.1 Base de Dados Usada nos Experimentos

5.1.1 Procedimento de Coleta de Dados

A realização de experimentos com os algoritmos desenvolvidos depende da montagem de uma base marcada. Idealmente, tal base deveria ser manualmente avaliada, porém como isso é inviável, uma abordagem semi-automática foi adotada. Inicialmente, foi realizada uma coleta de dados reais do Twitter, a partir do desenvolvimento de uma plataforma de coleta de dados que permite monitorar, via palavras-chave, a base do Twitter, respeitando alguns limites de acessos impostos pela API. Várias coletas foram realizadas em temas distintos, focando em eventos específicos (política, esportes, artistas, etc).

Entretanto, para que fossem avaliados os resultados dos experimentos realizados, era necessária uma base de dados com informações de sentimento previamente conhecidas. Como não foram encontradas tais bases disponíveis para uso público, a primeira abordagem para solucionar esse problema foi usar

[24] https://dev.twitter.com/rest/public

um componente de análise de sentimento textual para classificar automaticamente os dados através de seu conteúdo. Para tanto, foi usado o Alchemy API[25], componente comercial de análise de sentimento com interface REST que pode ser usado gratuitamente em volumes reduzidos (até mil consultas por dia). Essa limitação impactou sobremaneira a velocidade da montagem das bases, porém era a melhor alternativa até então identificada.

Uma técnica de classificação coletiva sobre os dados coletados foi implementada e, depois de ajustar o sistema e executar o componente de análise de sentimento durante algumas semanas, foi avaliado o algoritmo implementado. O experimento consistiu em desconsiderar uma porcentagem das classificações textuais automaticamente calculadas, efetuar a classificação desses nós com o algoritmo de classificação coletiva e avaliar a taxa de acerto desse algoritmo de acordo com o resultado da classificação textual. O resultado deste experimento, entretanto, foi desanimador: menos de um terço dos nós eram corretamente classificados (isto é, de acordo com o resultado da classificação textual).

Todavia, analisando manualmente os resultados produzidos pela API de classificação textual, percebeu-se que a sua precisão não era tão alta quanto o seu website prometia. Dessa forma, ficou evidente que outra maneira de marcar os dados seria necessária. A única forma confiável de fazer isso foi através de análise manual dos textos coletados. Assim, foi implementada uma interface gráfica que apresenta os textos coletados e permite que um operador determine se ele é positivo, neutro ou negativo acerca de um dado tema. Essa abordagem, como se podia esperar, se mostrou excessivamente trabalhosa e lenta, e após algumas semanas de trabalho ela foi posta em segundo plano por se mostrar absolutamente inviável (pelo menos enquanto não houver uma equipe dedicada a desempenhar exclusivamente essa tarefa, como é o caso deste trabalho de pesquisa). Com isso, voltou-se ao ponto inicial: tem-se um algoritmo de classificação coletiva e os dados (texto e estrutura de links entre os usuários que postaram as informações), porém é necessária uma base que permita a realização de experimentos para a avaliação do algoritmo. Após um longo período de discussões sobre como isso poderia ser feito, verificou-se que, em várias situações, os próprios usuários fornecem "pistas" importantes sobre o que

[25] www.alchemy.com

falam e que podem ser facilmente interpretadas de forma automática através do uso *hashtags*.

Numa das bases coletadas (sobre política nos EUA), verificou-se a presença recorrente de *hashtags* como #obama2012 e #dems2012 entre os partidários do partido democrata, além de *hashtags* como #gop, #teaparty ou #reps entre os partidários do partido republicano. Dessa forma, foi realizada uma coleta especificamente com essas *tags*, e se considerou que elas próprias serviriam como marcadores para os textos que as contêm. Vale ressaltar que, em alguns casos, os partidários do lado contrário usam tais tags para comentar sobre o adversário, porém o volume de tais ocorrências é desprezível e pode ser considerado ruído da base. Essa base foi escolhida para a realização dos experimentos porque, além do tema (política americana) fomentar recorrentes discussões em redes sociais (e, dessa forma, gera bastante conteúdo que pode ser coletado para análise), possuía também *hashtags* usadas para (na maior parte dos casos) explicitar a opinião dos autores, tornando possível a realização de experimentos diversos sem a necessidade de marcação manual.

Assim, após mais uma sessão de alguns dias de coleta dos dados conforme descrito acima, foi obtida uma base de aproximadamente 8 mil tweets dividida em cerca de 50% para cada um dos lados. Como essa base é de tweets, e este trabalho usa a estrutura de links entre nós de um grafo para inferir as classes dos nós, foi necessário realizar uma segunda coleta para obter os seguidores e os seguidos de cada um dos autores que postaram textos com os termos monitorados. Assim, chegou-se a um grafo com mais de 97 mil nós e 900 mil conexões (na verdade, o grafo seria muito maior, porém só foram considerados autores com no mínimo duas postagens e nós com no mínimo 5 conexões).

5.1.2 Utilização dos Dados em Experimentos

O primeiro passo para viabilizar a realização de experimentos é a construção de um grafo $G(V, E)$ de usuários de redes sociais. Cada nó $u_i \in V$ representa um usuário do Twitter, e uma aresta direcionada (u_i, u_j) representa uma relação do tipo seguidor, isto é, o usuário u_i segue o usuário u_j no Twitter. Não foram considerados pesos nas arestas, e assim $w_{i,j} = 1$ para todo o grafo. Nos experimentos realizados, cada usuário é classificado de acordo com o conjunto

de classes *C*={Democrata, Republicano, Desconhecido}, indicando sua orientação política.

De forma a realizar a coleta de uma amostra de usuários polarizados referente ao tópico de interesse (política norte-americana), o Twitter foi monitorado inicialmente de 01/01/2012 a 09/01/2012, através de buscas por *hashtags* específicas que indicam a preferência política do usuário que postou a mensagem (consulte a seção 5.1.1 para uma descrição mais detalhada de como se chegou a este processo de coleta de dados). Especificamente, foram adotadas *hashtags* como #obama2012 e #dems2012 para o partido Democrata e *hashtags* como #teaparty e #tcot para o partido Republicano. A Tabela 1 apresenta o número de tweets coletados para cada uma das consultas realizadas.

Tabela 1: Dados coletados para os experimentos

Hashtag	**Número de tweets**	**Número de usuários**
#dems2012	1260	525
#obama2012	3159	1952
#teaparty	1535	742
#tcot	3144	1500
Total	**9098**	**4719**

Essas *hashtags* foram usadas para classificar os *tweets* (para montagem da base de experimentos) porque normalmente são usadas como uma forma de marcação manual, por parte dos próprios autores, das suas preferências políticas, como nos exemplos a seguir:

@adbridgeforth: Fall in US jobless claims paves way for stronger hiring trend http://t.co/NUGGcWag #Obama2012: http://t.co/L90ZhWey
@realDonaldTrump you wish you had the gumption to walk a day on his shoes #OBAMA2012
Is it me or is Obama whipping this country into shape....#Obama2012
I just can't get over how the GOP is campaigning on the position that providing

people with healthcare is un-American. #baffled #Obama2012

Exemplo 4: *Tweets* pró-democratas (marcados com a *hashtag* #obama2012)

Liberals As MSNBC's Chris Matthews, Arianna Huffington Abandon Obama For Poor Performance http://t.co/k8QfPV2B #teaparty #liberal
The Obama presidency: A Blatant, Frontal Assault on the Constitutional Separation of Powers #tcot #teaparty #iamthe53 #TNGop #dem #p21 #p2b
Obama is cutting our military to the bone as China is building up theirs. Good times! #tcot #teaparty
Obama Green Loan Scandal: Is Fisker the Next Solyndra?. #teaparty #tcot #tlot #ocra http://t.co/bC4VmMII

Exemplo 5: *Tweets* pró-republicanos (marcados com a *hashtag* #teaparty)

Obviamente, há alguns casos em que as *hashtags* são usadas por opositores num contexto de crítica, mas esses casos podem ser entendidos como ruído da base e não causam impacto significativo sobre os resultados. Abaixo são listados alguns exemplos de *hashtags* usadas dessa forma:

1	@MittRomney exposes another @BarackObama 'fighting 4 middle class' lie http://t.co/VJTO5t3F #BarackFighting4SocialistMiddleClass #Obama2012
2	Rick Santorum listed as one of the most corrupt Senators in 2006 #news #tcot #tlot #nwo #ows #teaparty #gop http://t.co/p2vGv0xQ

Exemplo 6: Exemplos de tweets com *hashtags* usadas num contexto de crítica. Por amostragem, constatou-se que este tipo de ocorrência é incomum e as *hashtags* podem ser usadas para obtenção de um gold standard consistente das postagens.

No exemplo 1 acima, apesar da tag #obama2012, o conteúdo da postagem é claramente contra Barack Obama (ou, mais genericamente, contra o partido Democrata). Já o exemplo 2 contem a tag #teaparty mas seu conteúdo possui uma notícia negativa sobre Rick Santorum, senador pelo partido Republicano. Por análise amostral da base coletada[26], pode-se perceber que esse

[26] Numa análise com amostra aleatória de 100 tweets da base (50 de cada classe), foram identificados 10 tweets com conteúdo não correspondente à *hashtag*. Para diminuir ainda mais a

tipo de ocorrência não é tão comum e não prejudica a realização dos experimentos, e que o uso das *hashtags* provê uma forma consistente de classificação para obtenção do *gold standard* das postagens, apesar desses casos excepcionais. Dessa forma, cada *tweet* é marcado como sendo pró-Democrata ou pró-Republicano de acordo com a presença das *hashtags*.

Como se pode ver na Tabela 1, 4.719 usuários diferentes foram identificados como autores das postagens coletadas. A partir desse conjunto inicial, foram excluídos os usuários como menos de cinco seguidores e menos de duas postagens, o que resultou num grafo com 1.244 nós autores. Para a construção do grafo final, foram então coletados todos os usuários que eram seguidores ou seguidos pelos autores. Esses usuários são chamados de não-autores, uma vez que não manifestaram explicitamente sua opinião a respeito do tópico monitorado. Um grafo direcionado formado por nós autores e não-autores é então construído, com as arestas modelando a relação de seguidor/seguido. É importante destacar que, embora não se possa usar os não-autores para avaliar a qualidade do método de classificação proposto, esses usuário são importantes para que seja possível a obtenção de um grafo conexo, que permite propagar as classificações durante o processo de classificação coletiva.

Um autor é marcado com a classe que possui a maior quantidade de tweets a seu favor, e os não autores são marcados como classe Desconhecida. Um possível refinamento futuro para essa abordagem de marcação seria considerar a proporção dos tweets para cada classe e definir a polaridade como um valor no intervalo [-1, +1], cujos limites representam as pontas do espectro. Dessa forma, usuários com uma posição política mais radical estariam próximos aos limites do intervalo, enquanto usuários mais neutros (isto é, que citam ambos os lados com frequência similar) ficariam próximos ao centro do intervalo.

O processo acima descrito resultou inicialmente num grafo com mais de 97 mil nós e quase 1 milhão de arestas (veja a Figura 6).

incidência desses tweets, pode ser implementado um método de classificação baseado em mais de uma *hashtag*, por exemplo.

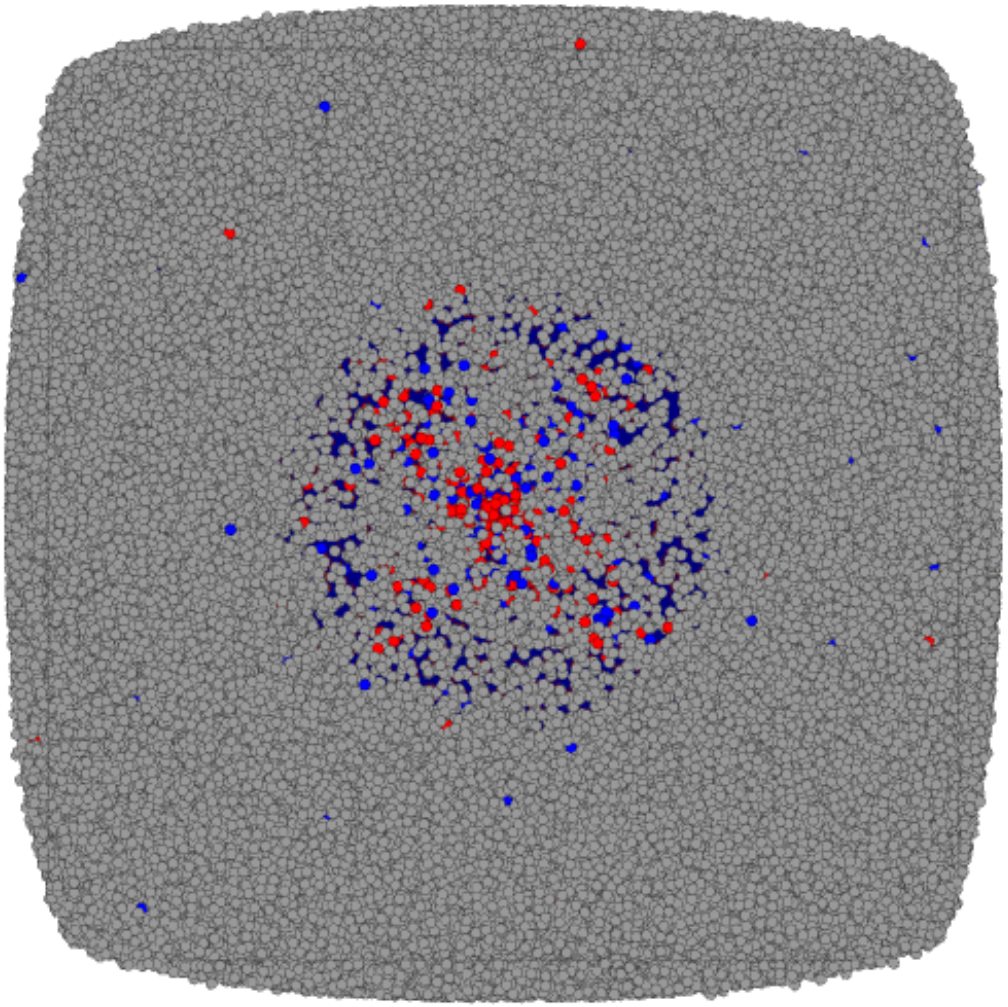

Figura 6: Grafo construído a partir da coleta de *tweets* contendo *hashtags* específicas (restrições: nós autores com no mínimo 2 *tweets* e nós quaisquer com no mínimo 5 conexões). Os nós azuis são os democratas, os vermelhos são os republicanos e os cinzas são os nós de usuários que não se manifestaram.

Conforme mencionado na seção 4.1.1, a aplicação de um algoritmo de classificação coletiva num grafo com essas características é inviável, visto que apenas uma pequena fração dos seus nós são autores (cuja classe pode ser determinada). Mesmo que todos os nós autores tenham classe conhecida, o conjunto-semente representaria apenas pouco mais de 1% do grafo total, quantidade insuficiente para propagar o conhecimento através de um algoritmo de classificação coletiva.

Por esse motivo, foi necessário desenvolver um método automático para redução da quantidade de nós desconhecidos do grafo. Simplesmente aumentar as restrições de mínimo de postagens e conexões não seria eficaz porque

diminuiria a quantidade de nós conhecidos na mesma proporção em que diminuiria os desconhecidos. Assim, foi desenvolvido o método de pruning automático do grafo detalhado na seção 4.1.1.

O grafo final produzido por esse algoritmo sobre o grafo de entrada ilustrado na Figura 6 possui um total de 1.244 nós e cerca de 78 mil arestas, sendo que aproximadamente 80% dos nós possuem classe conhecida. A Figura 7 apresenta esse grafo, visualizado através do Gephi (Bastian, Heymann, & Jacomy, 2009).

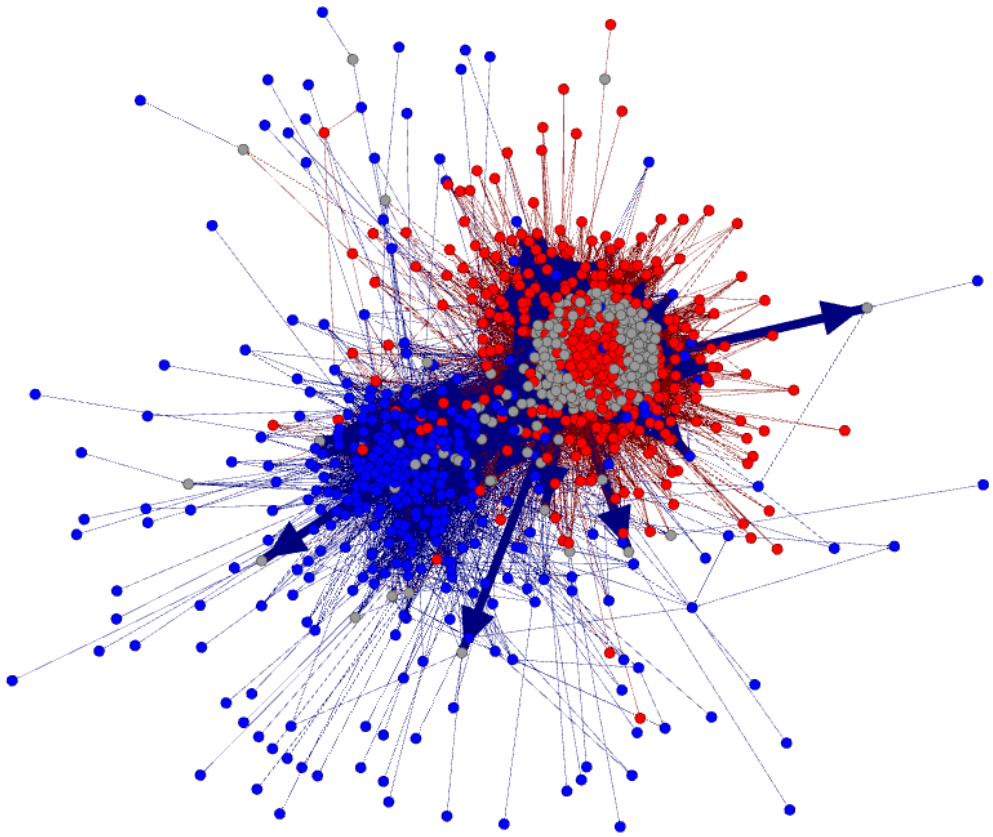

Figura 7: Grafo resultante da aplicação do algortimo de pruning sobre o grafo inicial apresentado na Figura 6. Os nós azuis são os democratas, os vermelhos são os republicanos e os cinzas são os nós de usuários que não se manifestaram. O grafo é direcionado e as setas representam a relação seguidor-seguido (as setas mais largas são devido a falha de renderização do aplicativo usado para visualização).

Como se pode perceber, os usuários no grafo são polarizados de acordo com sua preferência política. Aqueles usuários que se manifestaram

explicitamente tendem a seguir/serem seguidos por outros usuários com preferência similar. Há, entretanto, usuários ruidosos no grafo e uma fronteira na qual os nós apresentam um padrão misturado de relacionamento com ambos os lados. O coeficiente de assortatividade deste grafo, medido de acordo com a correlação de Pearson, foi de 0,72[27]. Este valor se mostrou estatisticamente significante a um nível de confiança de 95%. Em estudo (Bilgic & Getoor, 2008) para avaliar a aquisição de labels para classificação coletiva, os menores valores de coeficiente de assortatividade foram de 0,62. Dessa forma, o coeficiente do grafo usado no experimento é considerado conveniente para a execução de um processo de classificação coletiva[28].

5.2 Experimento I – Seleção do Conjunto-Semente a partir de Classificação Textual das Postagens

Este experimento consistiu em fazer uso de um classificador textual sobre os tweets coletados, e restringir quais serão usados efetivamente para definição do conjunto-semente, de acordo com a variação de dois parâmetros: a pontuação mínima de confiança θ_1 para cada classificação de tweet baseada em conteúdo, e o número mínimo de tweets por autor θ_2.. Ou seja, neste experimento foi realizada a seleção do conjunto-semente de acordo com a confiança de classificação sobre os nós autores, e foi avaliado o comportamento do sistema mediante a variação dos limiares usados. A orientação de opinião de um usuário pode ser definida de acordo com o sentimento observado nos textos postados por ele para os quais o classificador textual tenha obtido uma pontuação de confiança suficientemente alta. Em outras palavras: um usuário autor é classificado neste experimento de acordo com as opiniões fortes que ele emitiu acerca de um determinado tema sob monitoramento, e este conhecimento inicial é propagado pela rede através do algoritmo de classificação coletiva.

Seja $T_i=\{t_1, t_2, \ldots, t_{nt}\}$ o conjunto de *nt* postagens produzidas por um determinado usuário u_i. Inicialmente é aplicado um classificador textual para

[27] O coeficiente de assortatividade é o coeficiente da Correlação de Person entre pares de nós ligados. Valores positivos indicam correlação entre nós similares. Quando o valor é 1, diz-se que o grafo possui padrões assortativos perfeitos. Quando o valor é 0, diz-se que o grafo é não-assortativo. Quando o valor é -1, a rede é completamente desassortativa.

[28] Observe que o princípio da homofilia é avaliado nos grafos gerados de acordo com o tema sob monitoramento, e não sobre a rede social inteira.

inferir a orientação de opinião de cada texto $t_j \in T_i$. Cada classificação gerada pelo classificador textual está associada com uma pontuação de confiança. Se a confiança do classificador para uma determinada postagem estiver abaixo de um limiar pré-definido θ_1, aquela postagem não será considerada para classificar o usuário correspondente (isto é, seu autor). Após a aplicação do limiar de confiança, um número *nf* <= *nt* de postagens classificadas é coletado. De forma a evitar que a orientação dos usuários seja definida a partir de um número pequeno de postagens (o que tornaria o modelo pouco robusto a postagens esdrúxulas), só são considerados aqueles usuários para os quais existe um número mínimo θ_2 de postagens classificadas com pontuação de confiança acima do limiar estabelecido (isto é, se *nf* >= θ_2). A classificação para um usuário é, enfim, definida por votação a partir de tais postagens.

É importante destacar que os valores de θ_1 e θ_2 refletem a precisão e cobertura do processo de marcação inicial dos usuários da rede. Estabelecer valores muito pequenos para esses limiares pode fazer com que o número de nós inicialmente marcados seja alto (ou seja, com boa cobertura), porém com pouca precisão, o que pode acontecer com frequência, dada a dificuldade inerente da classificação textual de sentimentos. De forma contrária, se forem estabelecidos valores muito altos para os limiares, o número de nós inicialmente marcados pode ser insuficiente para o passo seguinte do método, que é a aplicação de um algoritmo de classificação coletiva. No caso ideal, valores adequados para esses parâmetros levarão a uma classificação inicial precisa, ao mesmo tempo em que produzirão usuários classificados em número suficiente para o processo de classificação coletiva. Conforme apresentado nos resultados deste experimento, o método proposto se mostrou robusto quanto à escolha dos limiares. A Tabela 2 apresenta a distribuição das classes em cada um dos conjuntos-semente usados nos experimentos.

Tabela 2: Distribuição das classes nos conjuntos-semente.

	θ_1=0		θ_1=0,9		θ_1=0,95		θ_1=0,99	
θ_2	**Rep**	**Dem**	**Rep**	**Dem**	**Rep**	**Dem**	**Rep**	**Dem**
1	48.40%	51.60%	49.40%	50.60%	52.40%	47.60%	54.40%	45.60%
2	49.40%	50.60%	57.40%	42.60%	50.40%	49.60%	53.40%	46.60%
3	54.40%	45.60%	52.40%	47.60%	53.40%	46.60%	48.40%	51.60%
4	56.40%	43.60%	55.40%	44.60%	51.40%	48.60%	49.40%	50.60%
5	55.40%	44.60%	56.40%	43.60%	54.40%	45.60%	52.40%	47.60%
6	52.40%	47.60%	50.40%	49.60%	52.40%	47.60%	56.40%	43.60%
7	48.40%	51.60%	55.40%	44.60%	51.40%	48.60%	50.40%	49.60%
8	47.40%	52.60%	52.40%	47.60%	52.40%	47.60%	50.40%	49.60%
9	51.40%	48.60%	56.40%	43.60%	57.40%	42.60%	51.40%	48.60%
10	48.40%	51.60%	55.40%	44.60%	53.40%	46.60%	54.40%	45.60%

A Tabela 2 lista os resultados da execução do classificador textual sobre todos os nós autores (isto é, o passo de marcação inicial descrito na seção 4.1). Para cada combinação de valores dos parâmetros, é apresentada a porcentagem de usuários classificados corretamente (C), a porcentagem dos usuários classificados incorretamente (E) e a porcentagem de usuários não classificados ou com classe desconhecida (D) (isto é, aqueles usuários que não atingiram os limiares estabelecidos). Perceba que a combinação de θ_1=0 e θ_2=1 corresponde ao uso isolado do classificador textual sobre todos os usuários, como numa abordagem clássica do problema de análise de sentimento que não possui uma fase de classificação coletiva. A precisão alcançada nesse caso foi de 81,2%, valor que será usado como benchmark de comparação.

Tabela 3: Resultados obtidos pelo classificador textual (marcação inicial dos nós)

	θ_1=0			θ_1=0,9			θ_1=0,95			θ_1=0,99		
θ_2	C	E	D	C	E	D	C	E	D	C	E	D
1	81,2%	18,8%	0,0%	83,4%	15,0%	1,6%	82,4%	14,3%	3,4%	80,8%	11,4%	7,7%
2	81,2%	18,6%	0,2%	73,6%	12,0%	14,4%	70,4%	10,8%	18,7%	59,7%	7,9%	32,4%
3	57,7%	9,3%	32,9%	48,6%	6,0%	45,5%	45,5%	5,1%	49,4%	37,4%	2,8%	59,8%
4	41,2%	6,6%	52,2%	34,2%	4,4%	61,4%	31,9%	3,5%	64,5%	23,9%	1,9%	74,1%
5	32,5%	4,4%	63,1%	34,2%	4,4%	61,4%	22,5%	2,0%	75,5%	16,7%	1,0%	82,3%
6	25,1%	3,4%	71,4%	17,1%	2,2%	80,7%	16,0%	1,3%	82,8%	12,7%	0,7%	86,6%
7	20,1%	2,8%	77,1%	14,3%	1,3%	84,5%	13,4%	1,2%	85,4%	10,8%	0,7%	88,6%
8	16,3%	2,4%	81,3%	11,8%	1,3%	87,0%	10,6%	0,9%	88,5%	8,3%	0,6%	91,1%
9	13,8%	1,6%	84,6%	10,3%	1,0%	88,7%	8,8%	0,8%	90,3%	6,7%	0,6%	92,7%
10	11,6%	1,2%	87,2%	8,3%	1,0%	90,7%	7,1%	0,8%	92,1%	5,2%	0,6%	94,2%

O gráfico da Figura 8 apresenta os mesmos dados, representando o percentual dos nós no eixo Y e o parâmetro θ_2 no eixo X. Para representar a variação do parâmetro θ_1, foram usadas cores (verde para θ_1=0, roxo para θ_1=0,9, vermelho para θ_1=0,95 e azul claro para θ_1=0,99). Os nós classificados corretamente tem linha tracejada, os nós com classificação errada tem linha contínua e os nós desconhecidos são representados por linhas pontilhadas.

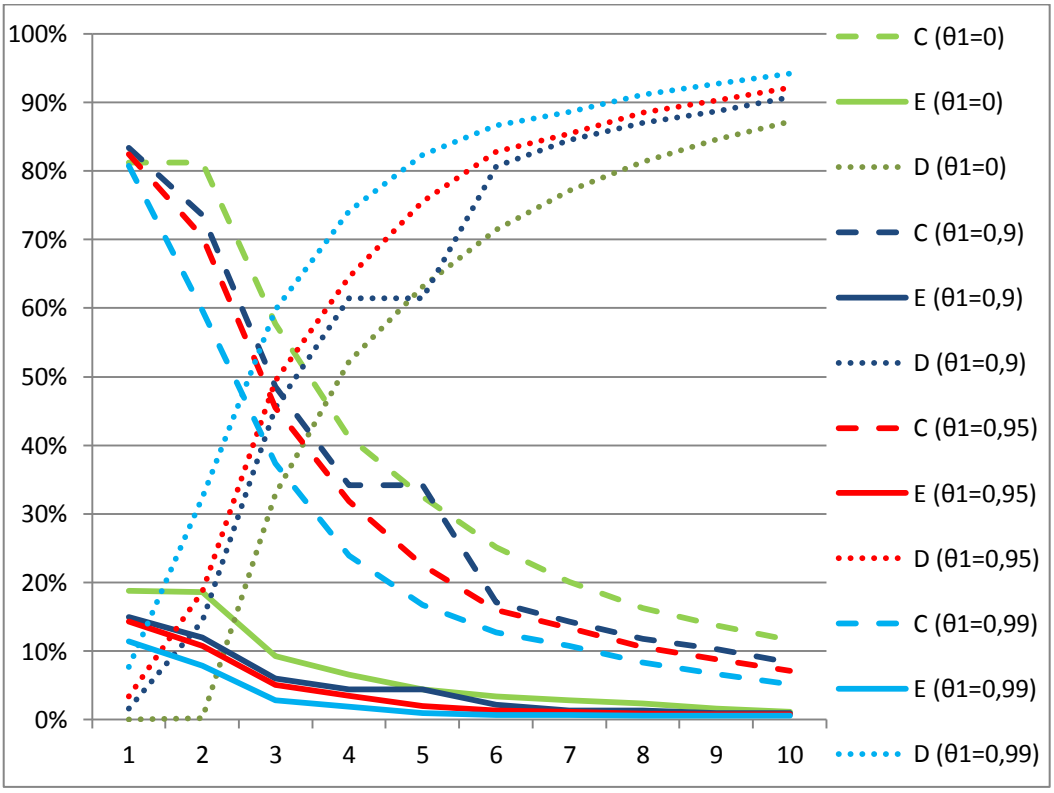

Figura 8: Representação gráfica dos dados da Tabela 2 (precisão do classificador textual). O eixo horizontal representa a variação do parâmetro θ_2.

Como esperado, à medida que os parâmetros se tornam mais restritivos, o número de usuários não marcados pelo classificador textual aumenta. E, mais importante, a precisão relativa do classificador textual também cresce. Isso fica evidente na Tabela 3, que mostra a precisão do classificador textual considerando apenas os usuários que atingiram os limiares estabelecidos. Dessa forma, ajustando adequadamente esses parâmetros, é alcançada uma precisão de acima de 90% em vários casos, o que é uma marca melhor do que a obtida pelos melhores trabalhos da literatura[29]. Entretanto, a cobertura nos casos de maior precisão é baixa se for considerado o conjunto total de usuários, já que há muitos usuários marcados como desconhecidos (isto é, quão mais restritivos os parâmetros, mais nós ficam fora do conjunto cuja classificação será

[29] Vale salientar um fato relevante neste ponto: em tarefas subjetivas como análise de sentimento, não existe um gold-standard que possa ser usado como base de comparação para resultados obtidos. Há trabalhos (Wilson, Wiebe, & Hoffmann, 2005) que realizaram experimentos para determinar a concordância de marcação entre operadores manuais para uma tarefa de análise de sentimento em nível de frase. A conclusão foi de que houve concordância em apenas 82% dos casos, o que pressupõe que sistemas nessa faixa de precisão podem ser considerados tão bons quanto um time de operadores humanos para a realização desta tarefa.

considerada). Este é justamente o ponto em que entra a etapa de classificação coletiva. A ideia é usar um classificador textual, estabelecendo parâmetros θ_1 e θ_2 adequados (isto é, que obtenham uma boa precisão relativa, mas mantendo uma cobertura suficiente) e aplicar o algoritmo de classificação coletiva sobre os nós restantes, combinando assim os pontos fortes de cada método: uma alta precisão de classificação a partir dos textos postados para uma fração pequena do total de usuários, e usar esse resultado como as classes iniciais para a classificação coletiva inferir as classes dos nós restantes.

Tabela 4: Taxas de acerto (C) e erro (E) obtidas pelo classificador textual na etapa de marcação inicial

	θ_1=0		θ_1=0,9		θ_1=0,95		θ_1=0,99	
θ_2	C	E	C	E	C	E	C	E
1	81,2%	18,8%	84,8%	15,2%	85,2%	14,8%	87,6%	12,4%
2	81,4%	18,6%	86,0%	14,0%	86,7%	13,3%	88,3%	11,7%
3	86,1%	13,9%	89,1%	10,9%	89,9%	10,1%	93,1%	6,9%
4	86,1%	13,9%	88,7%	11,3%	90,0%	10,0%	92,5%	7,5%
5	88,2%	11,8%	88,7%	11,3%	91,8%	8,2%	94,3%	5,7%
6	87,9%	12,1%	88,7%	11,3%	92,7%	7,3%	95,0%	5,0%
7	87,9%	12,1%	91,9%	8,1%	92,0%	8,0%	94,1%	5,9%
8	87,0%	13,0%	90,3%	9,7%	92,0%	8,0%	93,4%	6,6%
9	89,6%	10,4%	91,0%	9,0%	91,3%	8,7%	92,0%	8,0%
10	90,8%	9,2%	89,2%	10,8%	89,4%	10,6%	89,9%	10,1%

O gráfico a seguir apresenta os mesmos dados da Tabela 3, usando a mesma notação já aplicada anteriormente:

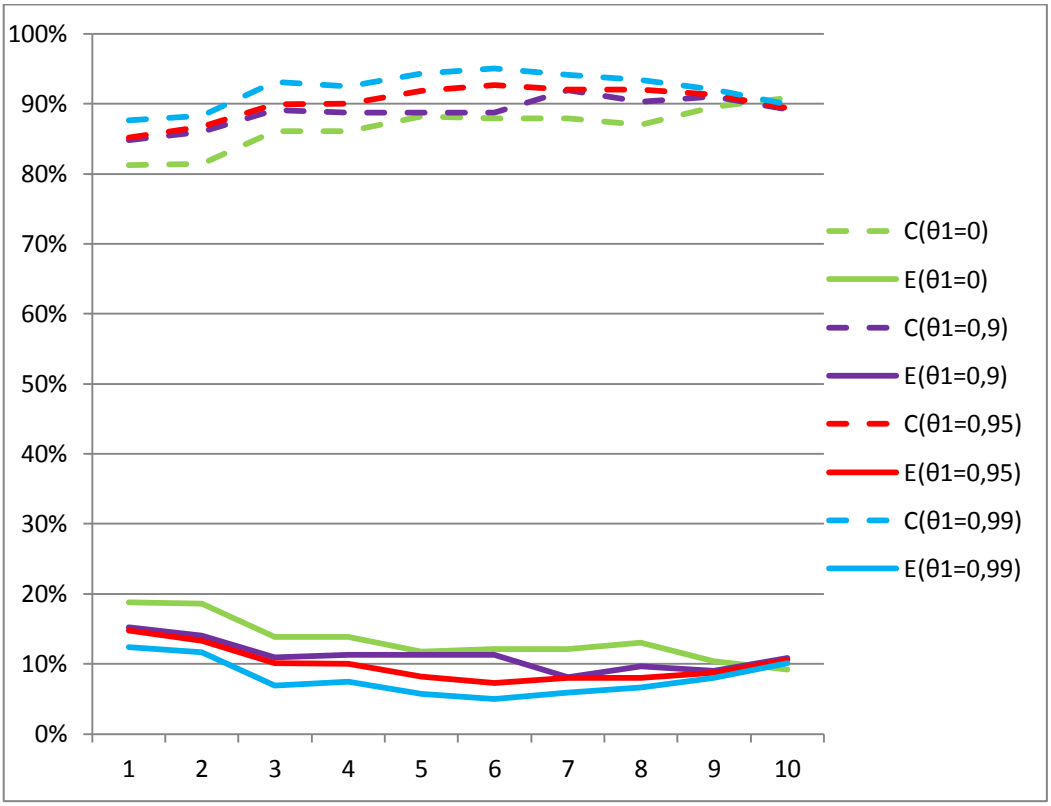

Figura 9: Representação gráfica dos dados da Tabela 3 (precisão do classificador textual considerando apenas os nós que atendem aos limiares definidos). O eixo horizontal representa a variação do parâmetro θ_2.

A Tabela 4 apresenta os resultados da classificação coletiva sobre os nós marcados como desconhecidos pelo classificador textual, usando como classes iniciais os nós que atingem os limiares estabelecidos. Perceba que é atingida uma alta taxa de precisão de acima de 80% na maior parte das vezes. O desempenho se mostrou robusto considerando os valores dos limiares, porém a precisão começa a degradar quando os parâmetros são demasiadamente restritivos. Esse comportamento é esperado, já que nesses casos o algoritmo recebe uma quantidade insuficiente de nós inicialmente classificados (por exemplo, para a combinação de parâmetros θ_1=0,99 e θ_2=10, o classificador coletivo precisa classificar 94% dos nós, partindo de apenas 6% de nós com classe conhecida).

Tabela 5: Precisão (C) e erro (E) obtidos pelo classificador coletivo sobre os nós não marcados pelo classificador textual

	θ_1=0		θ_1=0,9		θ_1=0,95		θ_1=0,99	
θ_2	**C**	**E**	**C**	**E**	**C**	**E**	**C**	**E**
1	-	-	100%	0,0%	93,1%	6,9%	89,6%	10,4%
2	66,7%	33,3%	91,1%	8,9%	88,8%	11,2%	88,6%	11,4%
3	87,0%	13,0%	86,8%	13,2%	86,3%	13,7%	87,7%	12,3%
4	86,3%	13,7%	87,5%	12,5%	87,4%	12,6%	87,8%	12,2%
5	87,1%	12,9%	88,1%	11,9%	87,7%	12,3%	86,9%	13,1%
6	88,0%	12,0%	87,6%	12,4%	88,4%	11,6%	87,4%	12,6%
7	88,9%	11,1%	87,6%	12,4%	86,0%	14,00%	85,2%	14,8%
8	88,0%	12,0%	85,1%	14,9%	84,8%	15,2%	83,6%	16,4%
9	88,2%	11,8%	83,7%	16,3%	82,9%	17,1%	72,7%	27,3%
10	84,8%	15,2%	80,7%	19,3%	64,3%	35,7%	60,5%	39,5%

O gráfico abaixo apresenta graficamente os dados da Tabela 4, usando a notação dos demais gráficos acima.

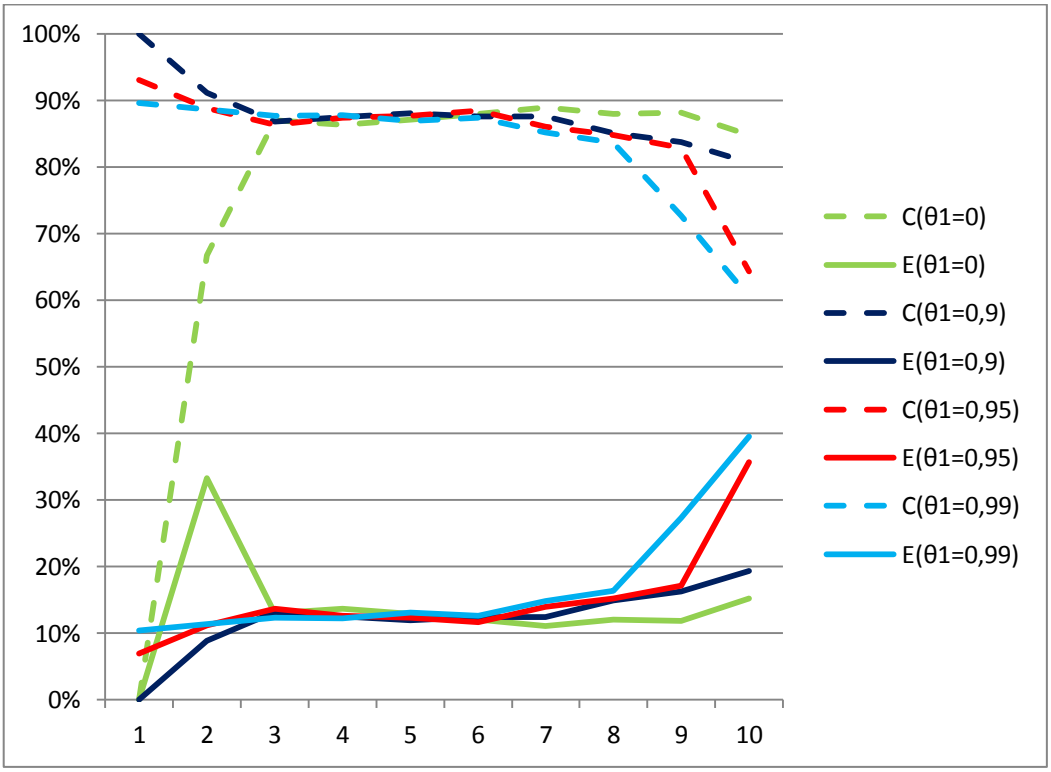

Figura 10: Precisão e erro do classificador coletivo sobre os nós que não atenderam aos limiares definidos e por isso foram desconsiderados na classificação textual (isto é, nós que não pertencem ao conjunto-semente). O eixo horizontal representa a variação do parâmetro θ_2.

O gráfico apresentado na Figura 10 precisa ser avaliado em conjunto com os dados que mostram a quantidade de nós desconsiderados pelo classificador textual (isto é, os nós desconhecidos) para que se tenha uma ideia de quanto a precisão do classificador coletivo influencia no resultado global. Por isso, é apresentado a seguir outro gráfico que acrescenta a linha pontilhada do número de nós que foram classificados pelo classificador coletivo.

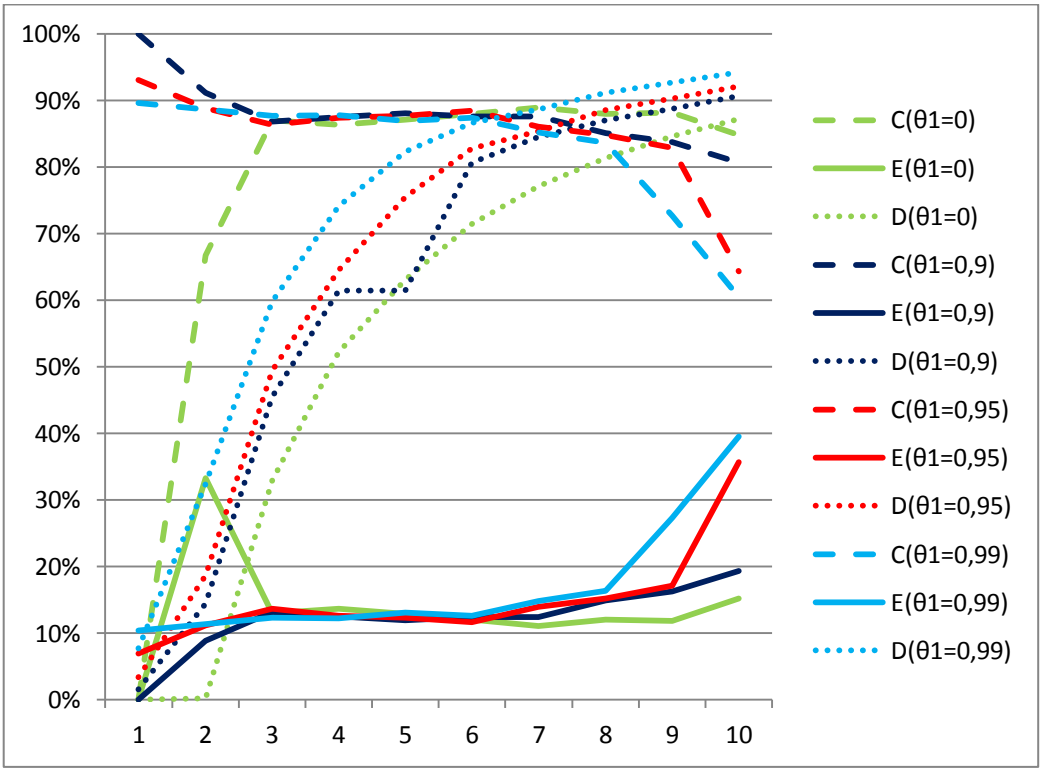

Figura 11: Gráfico com os dados de precisão e erro do classificador coletivo, acrescido da linha que informa a quantidade de nós efetivamente classificada para cada combinação de limiares. O eixo horizontal representa a variação do parâmetro θ_2.

Observe que o gráfico da Figura 11 evidencia que o classificador coletivo possui o melhor impacto quando o limiar θ_2 está no intervalo entre 6 e 8, pois é essa faixa que contem a melhor combinação entre precisão e quantidade de nós.

Finalmente, a Tabela 5 apresenta o resultado final da combinação das classificações textual e coletiva. Para quase todas as combinações de limiares a precisão observada foi maior do que o *benchmark* de 81,2%. As poucas exceções ocorreram para valores extremos dos limiares, como discutido anteriormente. O melhor resultado final foi observado para a combinação de limiares θ_1=0,99 e θ_2=3. A Tabela 5 mostra que, através da combinação de classificadores aqui proposta, é possível aumentar a precisão de 81,2% (*benchmark*) para 89,9% (uma diferença de 8,7 pontos percentuais ou 10,7%).

Além disso, os resultados indicam que o classificador coletivo pode ser consistentemente aplicado em cenários semelhantes. Em muitos casos, a precisão pode não ser o único indicador de qualidade. Supondo cenários em que a marcação inicial é viável porém demasiadamente custosa (por exemplo:

impossibilidade de uso de um classificador textual), é possível aplicar o método proposto combinando uma marcação inicial mais restrita que, se não é a ideal, é suficiente para alcançar uma precisão final aceitável no conjunto completo. Por exemplo, os resultados do experimento mostram que seria necessário marcar menos que 10% dos nós para atingir uma precisão de classificação de mais de 83% nos nós restantes, chegando a uma precisão combinada de 84,4%. Isto é: é possível produzir uma análise sobre um conjunto dez vezes mais representativo do que a amostra manualmente avaliada através do método proposto.

Tabela 6: Precisão (C) e erro (E) obtidos pela combinação do classificador textual com o classificador coletivo em todos os usuários

	θ_1=0		**θ_1=0,9**		**θ_1=0,95**		**θ_1=0,99**	
θ_2	**C**	**E**	**C**	**E**	**C**	**E**	**C**	**E**
1	81,2%	18,8%	85,0%	15,0%	85,5%	14,5%	87,8%	12,2%
2	81,3%	18,7%	86,7%	13,3%	87,1%	12,9%	88,4%	11,6%
3	86,4%	13,6%	88,0%	12,0%	88,1%	11,9%	89,9%	10,1%
4	86,2%	13,8%	87,9%	12,1%	88,3%	11,7%	89,0%	11,0%
5	87,5%	12,5%	88,3%	11,7%	88,7%	11,3%	88,2%	11,8%
6	88,0%	12,0%	87,8%	12,2%	89,1%	10,9%	88,4%	11,6%
7	88,6%	11,4%	88,3%	11,7%	86,9%	13,1%	86,2%	13,8%
8	87,8%	12,2%	85,8%	14,2%	85,6%	14,4%	84,5%	15,5%
9	88,4%	11,6%	84,5%	15,5%	83,7%	16,3%	74,1%	25,9%
10	85,6%	14,4%	81,5%	18,5%	66,3%	33,7%	62,2%	37,8%

Os dados da Tabela 5 são apresentados graficamente abaixo:

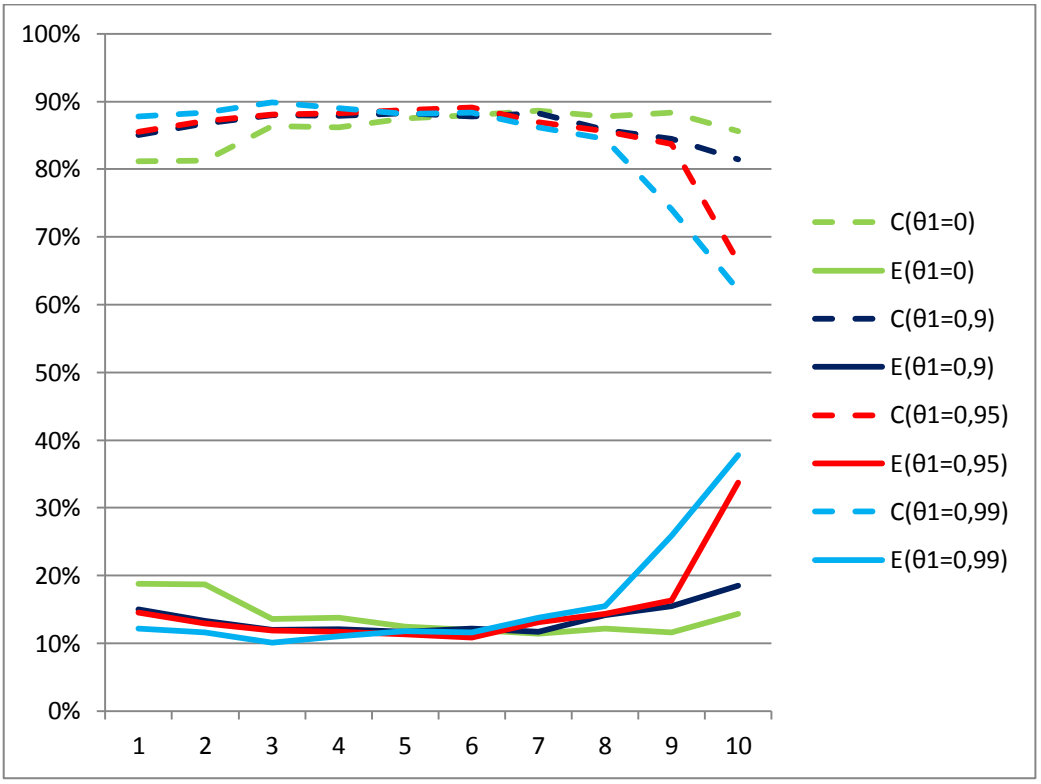

Figura 12: Precisão (C) e erro (E) obtidos pela combinação do classificador textual com o classificador coletivo em todos os usuários. O eixo horizontal representa a variação do parâmetro θ_2.

5.3 Experimento II – Seleção do Conjunto-Semente através de Análise Estrutural do Grafo

Além dos experimentos da seção anterior, cujo foco é a avaliação da abordagem proposta combinando classificação coletiva com um algoritmo de classificação inicial dos nós (nesse caso, um classificador textual), também foram realizados experimentos para avaliar a melhor forma de selecionar os nós que devem ser usados inicialmente para o processo de classificação coletiva de acordo com informações estruturais do grafo. Esse novo experimento consistiu em usar heurísticas de seleção dos nós iniciais e medir a precisão alcançada pela classificação coletiva a partir de diferentes quantidades de nós iniciais, selecionados de acordo com tais heurísticas (usando a classificação dada pelas hashtags para os nós iniciais, ou seja, assumindo que os nós iniciais possuem classificação correta). A ideia é avaliar se é possível estabelecer critérios objetivos para seleção dos nós iniciais que permitam uma maior precisão da classificação coletiva.

A heurística usada para seleção dos nós foi baseada no grau, isto é, a quantidade de conexões de cada nó, para explorar o fato de que nós com muitas conexões são localmente influentes (Kempe, Kleinberg, & Tardos, 2003)[30]. Para efeito de comparação, foi usada também uma seleção inicial aleatória (na verdade, a média de cinquenta execuções com conjuntos iniciais aleatórios diferentes para cada tamanho de conjunto inicial avaliado). O resultado é apresentado no gráfico abaixo, que mostra o desempenho da classificação coletiva a partir de conjuntos iniciais com tamanhos variados (em passos de 5% do total de nós conhecidos), que são formados com os nós mais conectados (curva azul), os menos conectados (curva verde), além da seleção aleatória já mencionada (curva vermelha).

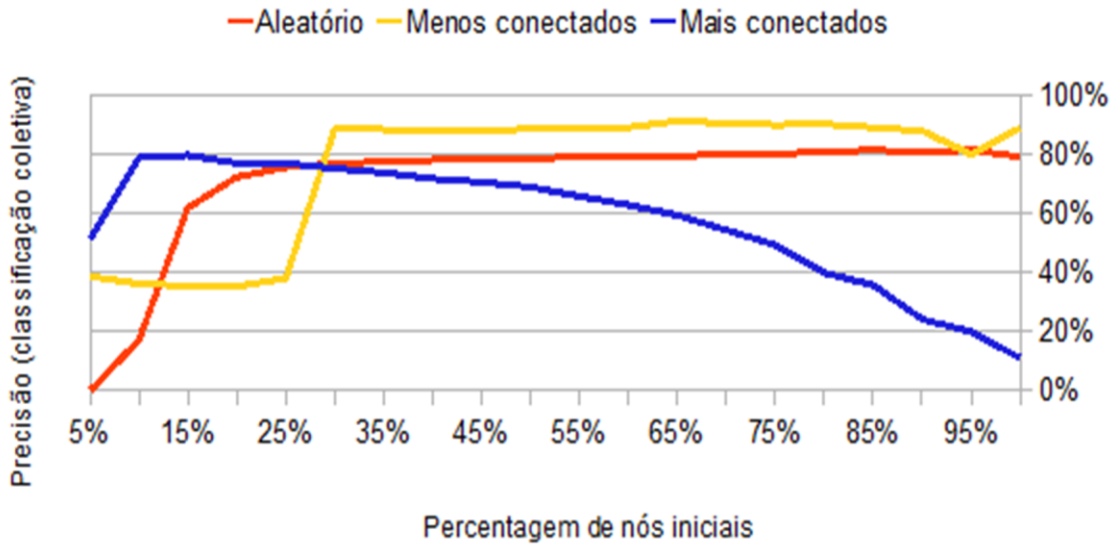

Figura 13: Precisão da classificação coletiva de acordo com o conjunto inicial de nós selecionados de três formas

Como esperado, se a classificação coletiva receber um número muito pequeno de nós iniciais, mesmo que sejam nós muito conectados, o resultado da classificação não será tão bom. Porém, partindo dos 10% mais conectados, a classificação coletiva já consegue uma precisão de cerca de 80% nos nós restantes. Isso indica que essa abordagem pode ser usada, produzindo um resultado bastante satisfatório, para classificar 90% dos nós de um grafo, uma vez que os 10% usados como conjunto-semente sejam os mais conectados e

[30] Outras opções poderiam ser usadas para a mesma finalidade, como medidas de *betweeness*, *closeness*, autovetores e *page rank* dos nós (Wikipedia, 2013).

tenham classificação conhecida previamente. Numa eventual aplicação real, um algoritmo de classificação automática como o que foi usado neste trabalho poderia ser aplicado aos 10% mais conectados, deixando os 90% restantes para a classificação coletiva. No caso de indisponibilidade de algoritmos com boa taxa de acerto para realizar a classificação inicial de forma automática, poderia ser avaliada até mesmo a possibilidade de usar uma equipe de operadores para analisar manualmente as postagens dos 10% usuários mais conectados. De forma inversa, partindo de um conjunto-semente com os nós menos conectados, vê-se que a precisão alcançada é inicialmente muito baixa, porém a partir de um conjunto de cerca de 30% dos nós menos conectados conhecidos, o resultado passa a ser muito bom. Isso acontece porque a rede já tem informação relacional suficiente, e precisa classificar nós que possuem muitas conexões (o que produz melhores resultados). O uso dos nós menos conectados como conjunto-semente em aplicações reais, entretanto, é pouco provável porque a relação custo *x* benefício não é favorável (isto é, o ganho em precisão normalmente não compensa o esforço extra de marcação necessário).

O gráfico da Figura 13 também mostra uma queda substancial da qualidade da classificação coletiva conforme o conjunto-semente formado pelos nós mais conectados cresce em tamanho. Isso acontece porque, tendo a classificação conhecida para os mais conectados, o algoritmo precisa classificar os nós com poucas conexões. Dessa forma, a informação relacional disponível entre os nós é insuficiente para que essa classificação seja feita de forma satisfatória. Essa análise é corroborada pela curva que mostra a precisão para o conjunto-semente definido a partir dos nós menos conectados. Quando temos um conjunto-semente com um número suficientemente grande de nós pouco conectados cuja classificação é conhecida, o resultado da classificação coletiva tem uma alta precisão, pois os nós que serão classificados pela classificação coletiva são aqueles com muitas conexões. Chama atenção, ainda, a queda pontual de precisão na curva que considera os nós menos conectados quando o conjunto-semente tem 95% do total dos nós. Isto aconteceu devido a ruído na base: como a precisão alcançada é medida sobre os nós restantes, que neste caso corresponde a apenas 5% dos nós, o resultado fica mais sensível a este tipo de anomalia na ocorrência de ruído.

Apesar da seleção de nós iniciais realizada neste trabalho ser baseada apenas em informações locais dos nós (grau), o melhor método de seleção de nós iniciais de acordo com um estudo específico (Rattingan, Maier, & Jensen, 2007) é baseado em alguma forma que permita a identificação de nós localmente centrais (ou seja, tenham grande influência sobre sua vizinhança) ao mesmo tempo que sejam globalmente dispersos (para que consigam propagar a informação de forma global). De acordo com experimentos realizados (Rattingan, Maier, & Jensen, 2007), o uso de centroides de clusters definidos através do k-*means* consegue combinar essas características. Como descrito adiante na no capítulo 6, uma avaliação do k-*means* (e possivelmente de outros algoritmos de agrupamento em grafos) é sugerida como possível extensão deste trabalho.

5.4 Limitações

O protótipo desenvolvido foi avaliado num cenário que possui duas características importantes:

1) Polarização de opiniões entre os usuários;

2) Invariabilidade das opiniões dos usuários ao longo do tempo.

Dessa forma, pelo menos inicialmente, é de se esperar que os resultados observados nos experimentos se repitam em cenários que apresentem características semelhantes. A seção 6.2 traz uma discussão mais aprofundada sobre esta questão.

5.5 Considerações Finais

Neste capítulo, foram detalhados os experimentos realizados sobre uma base coletada do Twitter sob o tema de política americana. Foram, ainda, apresentados os resultados alcançados pelo protótipo usando diferentes parâmetros de configuração para os limiares de número de tweets emitidos por usuário e pontuação de confiança do classificador textual. Por fim, foram realizados experimentos que determinam a formação do conjunto-semente através apenas de características estruturais (grau dos nós). Embora não se possa afirmar que esses números se repitam nos resultados das eleições reais, outros trabalhos que realizaram análises sobre política em redes sociais (através

apenas de classificadores textuais) atestam que os dados coletados a partir de redes sociais apresentam alta correlação com pesquisas de opinião convencionais (Ceron et al., 2014). No próximo capítulo, serão apresentadas as conclusões, principais contribuições e trabalhos futuros propostos como extensão à pesquisa desenvolvida.

6 Conclusões e Trabalhos Futuros

Esta tese tem por objetivo a classificação de usuários de redes sociais de acordo com suas opiniões acerca de um determinado objeto de interesse. A abordagem usual de classificar as opiniões emitidas pelos usuários através de processamento textual não foi aplicada aqui (na verdade, foi aplicada num experimento apenas para definição do conjunto inicial de nós). Nesta proposta, foi explorada a existência de conexões entre usuários das redes sociais para que, através da análise dos relacionamentos e afinidades entre os usuários, seja possível inferir as opiniões de usuários a partir da opinião de seus contatos.

A abordagem proposta neste trabalho é baseada no princípio da homofilia (Macskassy & Provost, 2007), que estabelece que indivíduos tendem a se conectar a outros com os quais apresentam alto grau de semelhança. A partir desse conceito, são aplicadas técnicas de classificação coletiva (Rattingan, Maier, & Jensen, 2007), (Sen et al., 2008) sobre o grafo que representa a rede social com a intenção de explorar o fato de que a classificação coletiva não realiza inferência utilizando apenas as características locais dos nós, mas também as características dos nós relacionados. Além disso, a classificação é executada de forma simultânea sobre todas as instâncias, o que permite considerar as influências que cada instância exerce sobre outras às quais está relacionada. Uma vez que essa abordagem não depende do conteúdo textual, ela é insensível aos vários complicadores inerentes às abordagens clássicas, como ironia, sarcasmo e resolução de contexto, entre outros.

Para viabilizar a aplicação de classificação coletiva, o grafo deve ser parcialmente classificado, isto é, precisa dispor de um conjunto-semente cujas classificações são conhecidas. A este problema que ocorre quando um sistema depende de um volume suficiente de informação coletada a priori (normalmente via intervenção manual), dá-se o nome de partida a frio (*cold start*), que é comum em sistemas de recomendação (Lam et al., 2008), (Park et al., 2006). No contexto de classificação coletiva de opiniões, o problema da partida a frio se reflete na necessidade de conhecimento prévio das opiniões de parte dos usuários da rede a respeito do tópico sob monitoramento.

Dado o contexto acima, esta proposta de tese lida com o problema de classificação de opiniões de usuários através de classificação coletiva, além da

dificuldade da geração do conjunto inicial de usuários etiquetados (partida a frio), através de duas abordagens: na primeira, classificadores textuais são usados para produzir de forma automática um conjunto inicial de nós de acordo com o conteúdo das mensagens postadas nas redes sociais; na segunda, técnicas de inferência ativa são usadas para identificar os nós mais relevantes a serem etiquetados usando medidas estruturais da rede.

Como mostram os experimentos do capítulo 5 e os resultados preliminares listados na seção 6.4, esta abordagem permite realizar a análise de sentimento em redes sociais com maior precisão do que as técnicas que dependem exclusivamente da análise textual, pois o método proposto é insensível às dificuldades inerentes à abordagem de classificação textual, tornando-se uma alternativa viável para tratamento do problema de análise de sentimento em cenários em que haja informações relacionais disponíveis.

6.1 Principais Contribuições

Podem-se destacar os seguintes resultados alcançados após o desenvolvimento deste trabalho:

- Definição de um método para aplicação de algoritmos de classificação textual, classificação coletiva e de princípios de inferência ativa para o problema de análise de sentimento em redes sociais;

- Criação de um framework de análise de sentimento para redes sociais (validado em experimentos de política, mas que pode ser aplicado a diferentes domínios), cujo principal componente de classificação não está sujeito aos problemas intrínsecos à análise textual;

- Criação de algoritmo de descarte seletivo de nós em grafos, com potencial para aplicação em qualquer cenário que envolva manipulação de grafos muito grandes;

- Criação de um *corpus* sob o tema de política americana, contendo citações (conteúdo textual) e informação relacional entre os usuários emissores de opinião.

6.2 Aplicabilidade

Como explicado anteriormente, o método aqui proposto é baseado no princípio da homofilia, isto é, a tendência que as pessoas apresentam de se relacionar mais frequentemente a outras pessoas com as quais possuem mais afinidades. A hipótese aplicada foi que essa propriedade pode ser usada para propagar as preferências e opiniões a respeito de um determinado tema numa rede de pessoas interconectadas, e experimentos conduzidos sobre dados reais coletados a partir de uma rede social online mostraram que isso de fato acontece.

Entretanto, há que se salientar que os experimentos foram realizados num cenário que apresenta duas características importantes:

1) Polarização de opiniões, isto é, existe uma separação natural entre os usuários a respeito do tópico monitorado;

2) Invariabilidade das opiniões dos usuários ao longo do tempo, uma vez que é incomum que pessoas mudem radicalmente sua orientação sobre política ao longo de um curto período de tempo (ainda mais quando se trata de pessoas que manifestam tal opinião em redes sociais, que tendem a ser mais engajadas e mais convictas sobre tais opiniões) (Effing, Hillegersberg, & Huibers, 2011).

À luz dessas características observadas nos experimentos, surge um questionamento sobre a aplicabilidade do método proposto em cenários nos quais essas características não se verifiquem. Por exemplo, é possível propagar com o mesmo sucesso as opiniões de usuários sobre produtos específicos, como telefones celulares ou carros? Nesses casos, não necessariamente há uma polarização de opiniões, pois um usuário que se manifesta a favor de um determinado produto não necessariamente é contra um produto concorrente. Além disso, nesse cenário é mais comum que um usuário mude sua opinião a respeito do produto em curtos períodos de tempo (por exemplo, satisfeito logo após a compra e insatisfeito após problemas apresentados em alguns dias).

Entretanto, a intenção ao realizar monitoramentos nesses cenários normalmente é avaliar a aceitação do produto ou serviço, e não comparar diretamente com outros produtos ou serviços concorrentes. Ou seja, os

experimentos realizados são indícios de que o método proposto possa ser aplicado nesses cenários monitorando o tópico de interesse, calculando o sentimento dos usuários acerca do tópico (positivo, negativo ou neutro) e propagando esse sentimento na rede de acordo com as conexões entre os usuários. No caso de comparações entre diferentes produtos, pode ser realizado o mesmo procedimento de forma isolada para cada um (com a possibilidade de um mesmo usuário ter a mesma opinião a respeito de produtos diferentes, como normalmente acontece).

Em relação à possível variabilidade de opiniões ao longo do tempo, pode ser criado um mecanismo de reavaliação das opiniões postadas ao longo do tempo, com atribuição de peso maior às opiniões mais recentes. No caso de mudanças de opiniões de uma parte relevante da rede, o método proposto seria capaz de capturar e propagar a informação na rede. A extensão do protótipo implementado, considerando essas características em cenários diferentes do que foi usado nos experimentos deste trabalho, é listada como possível trabalho futuro na próxima seção.

Como detalhado no capítulo 3, uma premissa para aplicabilidade do método é o uso de grafos com valores do coeficiente de assortatividade próximos a 1. Dessa forma, espera-se que o desempenho em grafos que não apresentem esta propriedade seja inferior ao verificado nos experimentos realizados.

6.3 Trabalhos Futuros

As pesquisas, implementações e execução de experimentos realizados até o momento foram importantes para validar o direcionamento deste trabalho. A ideia inicial seria desenvolver pesquisas especificamente na área de análise de texto, mas os experimentos realizados demonstram que há possibilidade de aplicação de uma abordagem alternativa e inovadora, baseada no uso das informações de links entre usuários, para o problema de análise de sentimento, possivelmente com resultados mais precisos e confiáveis do que as técnicas existentes que usam apenas o conteúdo textual.

Entretanto, várias possibilidades de extensão foram identificadas durante o desenvolvimento deste trabalho, com potencial para melhorar os

resultados alcançados ou tornar o modelo aplicável a uma gama maior de cenários. A seguir, são listadas as principais possibilidades de trabalhos futuros:

- Geração de grafos com o mínimo possível de nós desconhecidos para que seja avaliada a qualidade da classificação coletiva. Nesse caso, os nós desconhecidos serviriam apenas para manter a conectividade do grafo, e uma porcentagem de nós com classificação conhecida (isto é, nós autores) teriam sua classificação desconsiderada para que o algoritmo calculasse as classes com base apenas na estrutura de links, tornando possível a comparação dos resultados.

- Estudo de técnicas para generalização da abordagem de coleta de dados, de forma que o trabalho aqui desenvolvido seja aplicável ao maior número possível de cenários. Um exemplo de extensão que permitiria a generalização foi citado ao fim da seção 6.2: para tratar a variabilidade de opiniões ao longo do tempo, seria implementado um mecanismo de reavaliação periódica das opiniões postadas. Se houvesse mudança de opiniões de uma parte relevante da rede, esta extensão capturaria e propagaria a informação na rede. Algo semelhante foi desenvolvido para processamento contínuo de opiniões no Twitter (Wang et al., 2012);

- Avaliação de outros tipos de algoritmos de classificação coletiva (a exemplo daqueles apresentados em detalhes no capítulo 3) ou de técnicas de mineração de links que possam ser aplicadas neste cenário;

- Inclusão de novas informações nos grafos, como retweets (que normalmente significam uma concordância ou endosso do post retwitado), relações de amizade (seguimento recíproco), etc, e avaliação do impacto dessas novas informações na qualidade da resposta produzida;

- Variação do conjunto de treinamento dos classificadores textuais. Já que normalmente será difícil conseguir um conjunto grande de marcações manuais (ou mesmo um conjunto grande classificado satisfatoriamente através de uma heurística simples), outros

experimentos podem ser realizados reduzindo a quantidade de tweets no conjunto de treinamento do classificador textual.

- Seguindo a abordagem de trabalhos existentes (Rattingan, Maier, & Jensen, 2007), podem ser avaliadas as seguintes heurísticas para a seleção dos nós que devem ser inicialmente classificados: aleatória e grau (já avaliadas neste trabalho, com resultados apresentados no capítulo 5) e *closeness*, *betweeness* e *k-means*. Serão realizados experimentos usando a classificação dada pelas *hashtags* nos nós iniciais (que é considerada aqui a marcação correta dos nós) e também usando o classificador textual, para medição de quão distante ficará o resultado final do melhor possível. Caso o resultado observado em outros trabalhos se repita nesses experimentos (isto é, o *k-means* seja a melhor opção para seleção dos nós iniciais (Rattingan, Maier, & Jensen, 2007)), também podem ser avaliados outros algoritmos de agrupamento em grafos, inclusive algoritmos de detecção de comunidade;

- Realização de experimentos com grafos apresentando coeficientes de assortatividade inferiores àqueles dos grafos usados nos experimentos atuais, para avaliação do desempenho do método em cenários que apresentam condições diferentes (e a princípio inadequadas para sua aplicação).

6.4 Divulgação de Resultados

Foram publicados três artigos como resultado do trabalho desenvolvido durante a pesquisa:

- Juliano C. B. Rabelo, Ricardo B. C. Prudêncio, Flávia de Almeida Barros: Using link structure to infer opinions in social networks. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 681-685. Seul. 2012.

- Juliano Rabelo, Ricardo Prudêncio, Flávia Barros. Leveraging Relationships in Social Networks for Sentiment Analysis. In Proceedings of the 18th Brazilian symposium on Multimedia and the web (Webmedia). pp 181-188. São Paulo. 2012.

- Juliano Rabelo, Ricardo Prudencio, Flavia Barros. Collective Classification for Sentiment Analysis in Social Networks. In Proceeedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI). Atenas. 2012.

Os comentários enviados pelos avaliadores foram importantes tanto para validar o direcionamento da pesquisa quanto para realização de ajustes. O próximo passo em relação a este ponto é a submissão de um artigo sumarizando a última versão do trabalho desenvolvido para um periódico.

7 Bibliografia

Abbasi, A., Chen, H., & Salem, A. (Junho de 2008). **Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums**. *ACM Transactions on Information Systems, 26*(3), pp. 1-34.

Aggarwal, C. C. (2011). **Social Network Data Analytics**. Springer.

Agrawal, R., & Srikant, R. (1994). **Fast algorithms for mining association rules**. *In Proceedings of the International Conference on Very Large Data Bases*, (pp. 487-499).

Agrawal et al. (2003). **Mining newsgroups using networks arising from social behavior**. *Proceedings of the International World Wide Web Conference*, (pp. 529–535). Budapeste, Hungria.

Ananthakrishna, R., Chaudhuri, S., & Ganti, V. (2002). **Eliminating fuzzy duplicates in data warehouses**. *In Proceedings of the International Conference on Very Large Databases (VLDB), .* Honk Kong, China.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). **Modern Information Retrieval**. Addison Wesley.

Balahura, A., & Turchib, M. (Janeiro de 2013). **Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis**. *Computer Speech & Language, 28*(1), pp. 56-75.

Bastian, M., Heymann, S., & Jacomy, M. (2009). **Gephi: an open source software for exploring and manipulating networks**. *International AAAI Conference on Weblogs and Social Media.*

Bekafigo, M. A., & Mcbride, A. (2013). **Who Tweets About Politics?: Political Participation of Twitter Users During the 2011 Gubernatorial Elections**. *Journal Social Science Computer Review, 31*(5), pp. 625--643.

Bhattacharya, I., & Getoor, L. (2006). **A Latent diric hlet model for unsupervised entity resolution**. *In Proceedings of SIAM International Conference on Data Mining .*

Bilgic, M., & Getoor, L. (2008). **Effective Label Acquisition for Collective Classification**. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Las Vegas.

Bilgic, M., & Getoor, L. (2010). **Active Inference for Collective Classification**. In: M. Fox, & D. Poole (Ed.), *AAAI.* AAAI Press.

Bilgic, M., Mihalkova, L., & Getoor, L. (2010). **Active learning for networked data**. *Proceedings of the International Conference on Machine Learning*, (pp. 79-86).

Bollen et al. (Agosto de 2011). **Happiness is assortative in online social networks**. *Artificial Life, 17*(3), pp. 237-251.

Breiman, L. (2001). **Random Forests**. *Machine Learning Journal, 45*(1), pp. 5-32.

Carvalho et al. (2009). **Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-)**. *1st international ACM CIKM workshop on Topic-sentiment analysis for mass opinion.* Nova Iorque.

Chakrabarti et al. (1998). **Automatic Resource list Compilation by Analyzing Hyperlink Structure and Associated Text**. *Proceedings of the International World Wide Web Conference (WWW)*, (pp. 65-74).

Conover et al. (2011). **Predicting the political alignment of twitter users**. *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom).*

Cook, D. J., & Holder, L. B. (1994). **Substructure discovery using minimum description length and background knowledge**. *Journal of Artificial Intelligence Research*, pp. 231-255.

Cortes, C., & Vapnik, V. (Setembro de 1995). **Support-Vector Networks**. *Machine Learning Journal, 20*(3), pp. 273 - 297.

Craven et al. (2000). **Learning to construct knowledge bases from the world wide web**. *Artificial Intelligence*, pp. 69-114.

Culotta, A., & McCallum, A. (2005). **Joint deduplication of multiple record types in relational data**. *In Proceedings of the Fourteenth Conference on Information and Knowledge Management (CIKM).*

Ding, X., Liu, B., & Yu, P. (2008). **A holistic lexicon-based approach to opinion mining**. *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 231-240). New York, NY, USA: ACM.

Effing, R., Hillegersberg, J. v., & Huibers, T. (2011). **Social Media and Political Participation: Are Facebook, Twitter and YouTube Democratizing Our Political Systems?** In: E. Tambouris, A. Macintosh, & H. de Bruijn, *Electronic Participation* (pp. 25-35). Springer Berlin Heidelberg.

Efron, M. (2004). **Cultural orientation: Classifying subjective documents by cociation analysis**. *In Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design (2004), pp. 41-48 Key: citeulike:1444093*, (pp. 41-48).

Ermakov, S., & Ermakova, L. (2013). **Sentiment Classification Based on Phonetic Characteristics**. In: P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, et al., *Advances in Information Retrieval* (pp. 706-709). Springer Berlin Heidelberg.

Esuli, A., & Sebastiani, F. (2006). **SentiWordNet A Publicly Available Lexical Resource for Opinion Mining**. *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, (pp. 417-422). Gênova.

Feldman, R., & Sanger, J. (2006). **The Text Mining Handbook**. Nova Iorque: Cambridge University Press.

Ferligoj, A., Doreian, P., & Batagelj, V. (2005). **Generalized Blockmodeling (Structural Analysis in the Social Sciences)**. Cambridge University Press.

Ganapathibhotla, M., & Liu, B. (2008). **Mining opinions in comparative sentences**. *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 241-248). Manchester, United Kingdom: Association for Computational Linguistics.

Geman, S., & Geman, D. (1984). **Stochastic relaxation, Gibbs distributions and the bayesian restoration of images**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 721 - 741). IEEE.

Getoor, L., & Diehl, C. P. (Dezembro de 2005). **Link Mining: A Survey**. *ACM SIGKDD Exploration Newsletter, VII*(2), pp. 3-12.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). **Inferring web communities from link topology**. *ACM Conference on Hypertext and Hypermedia*, (pp. 225-234).

Goldberg, A. B., & Zhu, X. (2006). **Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization**. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, (pp. 45-52).

Guerra, P. (2013). **Multipolarized social networks: bridging graph mining, opinion mining and social sciences**. PhD Thesis Proposal.

Han, J., & Kamber, M. (2006). **Data Mining: Concepts and Techniques**. Morgan Kaufmann.

Hatzivassiloglou, V., & Wiebe, J. (2000). **Effects of adjective orientation and gradability on sentence subjectivity**. *Proceedings of the International Conference on Computational Linguistics (COLING).*

Hogenboom et al. (2011). **Determining Negation Scope and Strength in Sentiment Analysis**. *In Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2589 - 2594). Anchorage, AK: IEEE.

Hu et al. (2013). **Exploiting social relations for sentiment analysis in microblogging**. *Proceedings of the sixth ACM international conference on Web search and data mining*, (pp. 537-546).

Hu, M., & Liu, B. (2004). **Mining and summarizing customer reviews**. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 168-177). Seattle, WA, USA: ACM.

Huang, C., & Darwiche, A. (1996). **Inference in belief networks: A procedural guide**. *International Journal of Approximate Reasoning*, pp. 225-263.

Hummel, R. A., & Zucker, S. W. (1983). **On the Foundations of Relaxation Labeling Processes**. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(3), pp. 267-287.

Indurkhya, N., & Damerau, F. (2010). **Handbook Of Natural Language Processing**. Boca Raton, FL: CRC Press.

Jensen, D. (1999). **Statistical challenges to inductive inference in linked data**. *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics.*

Jensen, D., Neville, J., & Gallagher, B. (2004). **Why collective inference improves relational classification**. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 593-598). Nova Iorque, NY, EUA: ACM.

John, G. H., & Langley, P. (1995). **Estimating Continuous Distributions in Bayesian Classifiers**. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338-345). San Mateo: Morgan Kaufmann.

Jurafsky, D., & Martin, J. H. (2008). **Speech and Language Processing: An Introduction to Natural Language Processing**. Prentice-Hall.

Kalashnikov, D. V., Mehrotra, S., & Chen, Z. (2005). **Exploiting relationships for domain-independent data cleaning**. *In Proceedings of SIAM International Conference on Data Mining, .*

Kao, A., & Poteet, S. (2007). **Natural Language Processing and Text Mining**. Springer.

Kashima, H., & Inokuchi, A. (2002). **Kernels for graph classification**. *In Proceedings of the ICDM Workshop on Active Mining.*

Kempe, D., Kleinberg, J., & Tardos, E. (2003). **Maximizing the spread of influence through a social network**. *Proceedings of the 9th ACM SIGKDD international*, (pp. 137–146).

King et al. (Janeiro de 1996). **Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming**. *National Academy of Sciences*, pp. 438-442.

Kittler, J., & Illingworth, J. (1985). **Relaxation labelling algorithms — a review**. *Image and Vision Computing, 3*(4), pp. 206 - 216.

Kleinberg, J. (1999). **Authoritative sources in a hyperlinked environment**. *Journal of the ACM, 46*(5), pp. 604-632.

Laliwala, Z., & Shaikh, A. F. (2013). **Web Crawling and Data Mining with Apache Nutch**. Birmingham, UK: Packt Publishing.

Lam et al. (2008). **Addressing cold-start problem in recommendation systems**. *Proceedings of the 2nd international conference on Ubiquitous information management and communication (ICUIMC '08)*, (pp. 208-211). Suwon, Korea.

Li, X., Morie, P., & Roth, D. (2005). **Semantic integration in text: From ambiguous names to identifiable entities**. *AI Magazine*.

Liben-Nowell, D., & Kleinberg, J. (2003). **The link prediction problem for social networks**. *In Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, (pp. 556-559).

Liu, B. (2006). **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data**. Springer.

Liu, B. (2010). **Sentiment Analysis and Subjectivity**. In: N. Indurkhya, & F. J. Damerau, *Handbook of Natural Language Processing.* Chapman and Hall/CRC.

Lu, Q., & Getoor, L. (2003). **Link Based Classification**. *Proceedings of the Twentieth International.* Menlo Park, CA: AAAI Press.

Macskassy, S. A. (2007). **Improving Learning in Networked Data by Combining Explicit and Mined Links**. *Proceedings of the 22nd national conference on Artificial Intelligence (AAAI'07)* (pp. 590-595). AAAI Press.

Macskassy, S. A., & Provost, F. (2003). **A Simple Relational Classifier**. *In Proceedings of the Multi-Relational Data Mining Workshop at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Macskassy, S. A., & Provost, F. (Maio de 2007). **Classification in networked data: A toolkit and a univariate case study**. *Journal of Machine Learning Research, 8*, pp. 935-983.

Manber, U. (1989). **Introduction to Algorithms**. Addison-Wesley.

Martin, J. R., & White, P. (2005). **The Language of Evaluation: Appraisal in English**. Palgrave Macmillan.

Mcdonald et al. (2007). **Structured Models for Fine-to-Coarse Sentiment Analysis**. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.*

McDowell, L. K., Gupta, K. M., & Aha, D. W. (2007). **Cautious Inference in Collective Classification**. *In Proceedings of the Twenty-Second Conference on Artificial Intelligence.* Menlo Park, CA: AAAI Press.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). **Birds of a feather: Homophily in social networks**. *Annual Review of Sociology*, *27*(1), pp. 415-444.

Meena, A., & Prabhakar, T. V. (2007). **Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis**. *Proceeding ECIR'07 Proceedings of the 29th European conference on IR research* (pp. 573-580). Springer-Verlag Berlin.

Morinaga et al. (2002). **Mining product reputations on the web**. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 341–349). Edmonton, Alberta, Canada: ACM.

Mullen, T., & Collier, N. (2004). **Sentiment analysis using support vector machines with diverse information sources**. *Proceedings of Conference on Empirical Methods in Natural Language Processing.*

Mullen, T., & Malouf, R. (2006). **A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse**. *AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.*

Murphy, K. P., Weiss, Y., & Jordan, M. I. (1999). **Loopy Belief Propagation for Approximate Inference: An Empirical Study**. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 467--475). Estocolmo, Suécia: Morgan Kaufmann Publishers Inc.

Mustafaraj et al. (2011). **Vocal minority versus silent majority: Discovering the opinions of the long tail**. *SocialCom/PASSAT* (pp. 103-110). IEEE.

Neville, J., & Jensen, D. (2007). **Relational dependency networks**. *Journal of Machine Learning Research*.

Newman, M. E. (2004). **Detecting community structure in networks**. *European Physical Journal B*, pp. 321-330.

Ogneva, M. (19 de 04 de 2010). **How Companies Can Use Sentiment Analysis to Improve Their Business**. Acesso em 18 de 04 de 2013, disponível em mashable.com: http://mashable.com/2010/04/19/sentiment-analysis

O'Madadhain, J., Hutchins, J., & Smyth., P. (2005). **Prediction and ranking algorithms for even-based network data**. *SIGKDD Explorations*.

Page et al. (1999). **The PageRank citation ranking: Bringing order to the web**.

Pak, A., & Paroubek, P. (2010). **Twitter as a Corpus for Sentiment Analysis and Opinion Mining**. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Pan, S. J., & Yang, Q. (Outubro de 2010). **A Survey on Tranfer Learning**. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), pp. 1345-1359.

Pang, B., & Lee, L. (2004). **A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts**. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* Stroudsburg.

Pang, B., & Lee, L. (Janeiro de 2008). **Opinion Mining and Sentiment Analysis**. *Foundations and Trends in Information Retrieval, 2*(1-2), pp. 1-135.

Pardo, T. A., & Nunes, M. d. (2008). **On the Development and Evaluation of a Brazilian Portuguese Discourse Parser**. *Journal of Theoretical and Applied Computing, 15*(2), pp. 43-64.

Park et al. (2006). **Naïve filterbots for robust cold-start recommendations**. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 699-705). Philadelphia, PA, USA: ACM.

Parrott, W. G. (2000). **Emotions in Social Psychology: Key Readings** . Psychology Press.

Pasula et al. (2003). **Identity uncertainty and citation matching**. *Advances in Neural Information Processing Systems*.

Popescu, A., & Etzioni, O. (2005). **Extracting product features and opinions from reviews**. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 339-346). Vancouver, British Columbia, Canada: ACM.

Popescul, A., & Ungar, L. H. (2003). **Statistical relational learning for link prediction**. *In Proceedings of IJCAI Workshop on Learning Statistical Models from Relational Data.*

Porter, M. (1980). **An Algorithm for Suffix Stripping**. *Program, 14*(3), pp. 130–137.

Quinlan, J. R. (1993). **C4.5: Programs for Machine Learning**. Morgan Kaufmann Publishers.

Rabelo, J. C., Prudêncio, R. B., & Barros, F. d. (2012a). **Collective Classification for Sentiment Analysis in Social Networks**. *IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 958 - 963). Atenas, Grécia: IEEE.

Rabelo, J. C., Prudêncio, R. B., & Barros, F. d. (2012b). **Leveraging Relationships in Social Networks for Sentiment Analysis**. *Proceedings of the 18th Brazilian symposium on Multimedia and the web (Webmedia)* (pp. 181-188). São Paulo, Brazil: ACM.

Rabelo, J. C., Prudêncio, R. B., & Barros, F. d. (2012c). **Using link structure to infer opinions in social networks**. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 681-685). Seul, Korea: IEEE.

Rao, D., & Ravichandran, D. (2009). **Semi-supervised polarity lexicon induction**. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 675-682).

Rattingan, M., Maier, M., & Jensen, D. (2007). **Exploiting Network Structure for Active Inference in Collective Classification**. *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops* (pp. 429-434). IEEE Computer Society.

Rose et al. (Novembro de 2007). **Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning**. *International Journal of Computer Supported Collaborative Learning*.

Rosenfeld, A., Hummel, R., & Zucker, S. (1976). **Scene labeling by relaxation operations**. *IEEE Transactions on Systems, Man and Cybernetics, VI*(6).

Sarawagi, S. (Março de 2008). **Information Extraction**. *Foundations and Trends in Databases, 1*(3), pp. 261-377.

Sen et al. (2008). **Collective classification in network data**. *AI Magazine, 29*(3), pp. 93–106.

Sindhwani, V., & Melville, P. (2008). **Document-Word Co-regularization for Semi-supervised Sentiment Analysis**. *Proceedings of the Eighth IEEE International Conference on Data Mining*, (pp. 1025 - 1030). Pisa.

Sniedovich, M. (2010). **Dynamic Programming: Foundations and Principles**. Taylor & Francis.

Snyder, B., & Barzilay, R. (2007). **Multiple aspect ranking using the Good Grief algorithm**. *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, (pp. 300–307).

Soliman et al. (2013). **Mining Social Networks' Arabic Slang Comments**. *In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13).* Praga, República Tcheca.

Su et al. (2008). **Hidden Sentiment Association in Chinese Web Opinion Mining**. *Proceedings of the 17th international conference on World Wide Web* (pp. 959-968). Beijing, China: ACM.

Titov, I., & Mcdonald, R. (2008). **A Joint Model of Text and Aspect Ratings for Sentiment Summarization**. *Proceedings of the ACL-08*, (pp. 308--316).

Tumasjan et al. (2010). **Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment**. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178-185). AAAI Press.

Wang et al. (2011). **Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach**. *Proceedings of the 20th ACM Conference on Information and Knowledge Management* (pp. 1031-1040). ACM.

Wang et al. (2012). **A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle**. *Proceedings of the ACL 2012 System Demonstrations* (pp. 115-120). Jeju Island, Korea: Association for Computational Linguistics.

Wasserman, S., & Faust, K. (1994). **Social Network Analysis: Methods and Applications**. Cambridge: Cambridge University Press.

Whitelaw, C., Garg, N., & Argamon, S. (2005). **Using appraisal groups for sentiment analysis**. *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05)*, (pp. 625-631). Nova Iorque.

Wiebe et al. (Setembro de 2004). **Learning Subjective Language**. *30*(3), pp. 277-308.

Wiebe, J., & Mihalcea, R. (2006). **Word sense and subjectivity**. *Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL).*

Wiegand et al. (2010). **A survey on the role of negation in sentiment analysis**. *Proceeding NeSp-NLP '10 Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 60-68). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wikipedia. (11 de Novembro de 2013). Acesso em 1 de Fevereiro de 2014, disponível em Wikipedia: http://en.wikipedia.org/wiki/Centrality

Wilks, Y., & Stevenson, M. (1998). **The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation**. *Journal of Natural Language Engineering*, *4*(2), pp. 135–144.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). **Recognizing contextual polarity in phrase-level sentiment analysis**. *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, (pp. 347–354).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). **Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis**. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347-354). Stroudsburg, PA: ACM.

Wilson, T., Wiebe, J., & Hwa, R. (2004). **Just how mad are you? Finding strong and weak opinion clauses**. *Proceedings of AAAI*, (pp. 761–769).

Witten, I. H., Frank, E., & Hall, M. A. (2011). **Data Mining: Practical Machine Learning Tools and Techniques (3rd edition)**. Morgan Kaufmann Publishers.

Yedidia, J. S., Freeman, W., & Weiss, Y. (2005). **Constructing free-energy approximations and generalized belief propagation algorithms**. *In Proceedings of the IEEE Transactions on Information Theory* (pp. 2282-2312). IEEE.

Yoshida, K., Motoda, H., & Indurkhya, N. (Julho de 1994). **Graph-based induction as a unified learning framework**. *Journal of Applied Intelligence*, pp. 297-316.

Yuan, Y. C., & Gay, G. (2006). **Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams**. *Journal of Computer-Mediated Communication, 11*(4), pp. 1062-1084.

Zhang et al. (2013). **Joint Naïve Bayes and LDA for Unsupervised Sentiment Analysis**. *Advances in Knowledge Discovery and Data Mining* (pp. 402-413). Springer-Verlag Berlin Heidelberg.

Zhang, N. L., & Poole, D. (1996). **Exploiting causal independence in Bayesian network inference**. *Journal of Artificial Intelligence Research*, pp. 301-328.