UNIVERSIDADE DE SÃO PAULO
FFCLRP-DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

“SentiElection: análise de sentimento no Twitter baseada em
centralidade de palavras”

George Narita Vilarinho

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Ribeirão Preto-SP

2019

George Narita Vilarinho

“SentiElection: análise de sentimento no Twitter baseada em
centralidade de palavras”

Versão Revisada

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Orientador: Evandro Eduardo Seron Ruiz

Ribeirão Preto–SP

2019

George Narita Vilarinho

SentiElection: análise de sentimento no Twitter baseada em centralidade de palavras. Ribeirão Preto–SP, 2019.

91p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Evandro Eduardo Seron Ruiz

1. Análise de sentimento no Twitter. 2. Classificação binária. 3. Medidas de centralidade. 4. Grafo de palavras.

George Narita Vilarinho

“SentiElection: análise de sentimento no Twitter baseada em
centralidade de palavras”

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Trabalho aprovado. Ribeirão Preto–SP, 30 de julho de 2019:

Evandro Eduardo Seron Ruiz
Orientador

Professor
Convidado 1

Professor
Convidado 2

Ribeirão Preto–SP
2019

*Para mim, é muito melhor compreender
o universo como ele realmente é
do que persistir no engano, por mais satisfatório e
tranquilizador que possa parecer.
(Carl Sagan)*

Agradecimentos

Aqui deixo meus mais profundos agradecimentos a todos que contribuíram de alguma forma para a conclusão desta etapa de grande aprendizado em minha vida.

A toda minha família, Solange Narita, Marina Narita, Ivan Narita e Ivan Vilarinho que me apoiaram e muito se esforçaram para que eu tivesse condições de cumprir este objetivo de vida. Também gostaria de agradecer aos familiares, especialmente Oscar Narita e Mitiê Osaka, que me receberam com carinho em Ribeirão Preto e sempre me apoiaram do início ao fim.

Ao meu orientador, Prof. Dr. Evandro Eduardo Seron Ruiz pela consideração, amizade, orientação, paciência e ensinamentos que foram essenciais para a idealização deste trabalho e para meu aprendizado pessoal e profissional.

Aos grandes amigos que fiz durante esses últimos anos e que espero levar por toda minha vida: Marília Melo, Diogenes Laertius, Patrícia Fukushima, Rafael Delalibera, Gabriel Gusmão, Angélica Ribeiro, Wellington Zhao e Raquel Candido. A todos os amigos que fiz recentemente na Gryfo, empresa onde eu trabalho atualmente, especialmente aos seus gestores por terem acreditado em mim, justamente em um momento crucial da minha vida em termos financeiros.

E por fim, um agradecimento mais que especial à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro dado a mim durante dois anos completos, o qual foi diretamente responsável pelo alcance dos resultados obtidos e aos dois trabalhos publicados.

*“Acredite em si próprio e chegará um dia
em que os outros não terão outra escolha
senão acreditar com você.”
(Cynthia Kersey)*

Resumo

As redes sociais, a exemplo do Twitter, são um fenômeno que atraíram rapidamente milhões de usuários ativos. O uso destas redes permite que diariamente sejam trocadas milhões de postagens referentes a uma infinidade de tópicos. Toda esta pletora de textos tem motivado várias pesquisas na área de Processamento de Língua Natural, principalmente no tópico de análise de sentimento (AS) que busca extrair, analisar e quantificar a opinião do usuário sobre um determinado assunto. Devido ao alto custo envolvido no processo de extração e análise manual desses dados, diversos estudos têm focado na busca de soluções para a automatização dessa tarefa. Esta pesquisa propõe uma nova metodologia de AS sobre *tweets* baseada na Teoria de Grafos chamada SentiElection. Nossa hipótese é que textos positivos e textos negativos guardam, entre si, semelhanças no encadeamento de palavras. Calculamos essa importância através do uso de medidas de centralidade de vértices em grafos de palavras, sendo assim, a classe a qual determinado texto obtiver o maior valor de importância corresponderá a seu sentimento. O SentiElection é uma medida composta pelas medidas de centralidade de autovetor, *Katz* e *PageRank*. Em nossos experimentos o SentiElection apresentou resultados competitivos em relação a métodos tradicionais de AS, atingindo valores de acurácia e medida-F superiores a 70%. Além disso, nesses mesmos experimentos, o método aqui proposto obteve desempenho superior à técnica que o inspirou, a qual também faz uso de similaridade de grafos para a classificação de sentimentos.

Palavras-chave: Análise de sentimento no Twitter. Classificação binária. Medidas de centralidade. Grafo de palavras.

Abstract

Social networks, like Twitter, are a phenomenon that has quickly attracted millions of active users. The use of these social networks allows millions of posts to be exchanged daily on a multitude of topics. This plethora of texts has motivated several research topics in the area of Natural Language Processing, mainly in the topic of sentiment analysis (SA). SA seeks to extract, analyze, and quantify the user's opinion on a particular subject. Due to the high cost involved in the manual extraction and analysis of social network data, several studies have focused on finding solutions to automate this task. This research proposes a new SA methodology over tweets based on Graph Theory analysis called SentiElection. We hypothesize that positive and negative tweet texts are similar in their own class if these messages are compared based on the sequence of words in a sentence. Vertex centrality measures are applied to word graphs to measure how an incoming tweets relates either to a positive set of tweets or to a negative one. SentiElection is a compilation of centrality measures, such as, eigenvector, Katz similarity and PageRank. In our experiments, the SentiElection presented competitive results compared to traditional SA methods, reaching accuracy and F-measurement values higher than 70%. Moreover, in these same experiments, the method proposed here obtained superior performance to the technique that inspired it, which makes use of graph similarity for sentiment classification.

Keywords: Twitter sentiment analysis. Binary classification. Centrality measures. Word Graphs.

Lista de figuras

Figura 1 – Exemplo de <i>tweet</i>	24
Figura 2 – Exemplo de grafo não direcionado.	34
Figura 3 – Exemplo de grafo direcionado.	34
Figura 4 – Exemplo de multigrafo.	34
Figura 5 – Exemplo de grafo direcionado e sua respectiva matriz de adjacência. . .	35
Figura 6 – Isomorfismo entre os grafos G_1 e G_2	36
Figura 7 – Máximo subgrafo comum H dos grafos G_1 e G_2	37
Figura 8 – Exemplo de rede complexa.	38
Figura 9 – Centralidade de vértices baseado no grau.	39
Figura 10 – Partindo do nó A, o usuário avança aleatoriamente para uma das páginas B, C ou D, com probabilidade de 1/3 cada.	41
Figura 11 – Ciclo fechado durante o cálculo do <i>PageRank</i> das páginas A, B, C e D. . .	42
Figura 12 – Grafo de palavra com <i>word frame</i> igual a 2.	51
Figura 13 – Grafo de palavra com <i>word frame</i> igual a 3.	51
Figura 14 – Transformação dos textos e união dos grafos de palavras de cada texto contido no conjunto de dados positivos.	52
Figura 15 – Representação das redes de sentimento P e N antes da união temporária com o grafo de palavras T	53
Figura 16 – Representação das redes de sentimento P e N após a adição de T em ambos.	54
Figura 17 – Representação geral do funcionamento do SentiElection.	55
Figura 18 – Cálculo a nível de sentença da centralidade de uma frase inserida nas duas redes de sentimento.	56
Figura 19 – Cálculo a nível de palavra da centralidade de uma frase inserida nas duas redes de sentimento.	57
Figura 20 – Sequência de tarefas da análise de sentimento usando o método SentiE- lection.	59
Figura 21 – Versão mais simples do SVM denominado SVM Linear.	64
Figura 22 – Representação geral do funcionamento do <i>Random Forest</i>	66
Figura 23 – Exemplo de validação cruzada K-Fold utilizando 5 <i>Folds</i>	71
Figura 24 – Valores de acurácia e medida-F média dos classificadores sobre o con- junto de dados US Airline 2015.	81
Figura 25 – Valores de acurácia e medida-F média dos classificadores sobre o con- junto de dados Coachella 2015.	81
Figura 26 – Valores de acurácia e medida-F média dos classificadores sobre o con- junto de dados GOP Debate 2016.	82

Figura 27 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados Deflategate 2015.	82
Figura 28 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados SemEval 2017.	83

Lista de tabelas

Tabela 1 – Conjuntos de dados adquiridos para esta pesquisa.	48
Tabela 2 – Conjuntos de dados balanceados que serão usados para construção de grafos de palavras e testes, e seus respectivos número de <i>tweets</i> positivos e negativos.	49
Tabela 3 – <i>Datasets</i> divididos em conjuntos de treino e teste, e seus respectivos número de <i>tweets</i> positivos e negativos.	50
Tabela 4 – Tabela de frequência das palavras contabilizadas por documento. . . .	62
Tabela 5 – Tabela do inverso da frequência das palavras contabilizadas por documento.	62
Tabela 6 – Matriz de confusão.	69
Tabela 7 – Resultados do classificador SO-CAL em porcentagem (%)	73
Tabela 8 – Resultados do classificador Naïve Bayes em porcentagem (%)	73
Tabela 9 – Resultados do classificador SVM em porcentagem (%)	73
Tabela 10 – Resultados do classificador <i>Random Forest</i> em porcentagem (%)	74
Tabela 11 – Resultados do classificador CS (<i>word frame</i> = 2) em porcentagem (%)	74
Tabela 12 – Resultados do classificador CS (<i>word frame</i> = 3) em porcentagem (%)	75
Tabela 13 – Resultados do classificador CS (<i>word frame</i> = 4) em porcentagem (%)	75
Tabela 14 – Resultados do classificador MCSNS (<i>word frame</i> = 2) em porcentagem (%)	75
Tabela 15 – Resultados do classificador MCSNS (<i>word frame</i> = 3) em porcentagem (%)	75
Tabela 16 – Resultados do classificador MCSNS (<i>word frame</i> = 4) em porcentagem (%)	76
Tabela 17 – Resultados do classificador MCSUES (<i>word frame</i> = 2) em porcentagem (%)	76
Tabela 18 – Resultados do classificador MCSUES (<i>word frame</i> = 3) em porcentagem (%)	76
Tabela 19 – Resultados do classificador MCSUES (<i>word frame</i> = 4) em porcentagem (%)	76
Tabela 20 – Resultados do classificador MCSDES (<i>word frame</i> = 2) em porcentagem (%)	77
Tabela 21 – Resultados do classificador MCSDES (<i>word frame</i> = 3) em porcentagem (%)	77
Tabela 22 – Resultados do classificador MCSDES (<i>word frame</i> = 4) em porcentagem (%)	77

Tabela 23 – Resultados do classificador SentiElection a nível de sentença (<i>word frame</i> = 2) em percentagem (%)	78
Tabela 24 – Resultados do classificador SentiElection a nível de sentença (<i>word frame</i> = 3) em percentagem (%)	78
Tabela 25 – Resultados do classificador SentiElection a nível de sentença (<i>word frame</i> = 4) em percentagem (%)	78
Tabela 26 – Resultados do classificador SentiElection a nível de palavra (<i>word frame</i> = 2) em percentagem (%)	78
Tabela 27 – Resultados do classificador SentiElection a nível de palavra (<i>word frame</i> = 3) em percentagem (%)	79
Tabela 28 – Resultados do classificador SentiElection a nível de palavra (<i>word frame</i> = 4) em percentagem (%)	79

Lista de símbolos

$\cap$	Intersecção
α	Letra grega Alfa
λ	Letra grega Lambda
μ	Letra grega Mi
$\in$	Pertence
Σ	Somatório
$\cup$	União

Sumário

	Introdução	17
1	REFERENCIAL TEÓRICO	22
1.1	Twitter	22
1.2	Análise de Sentimento	23
1.3	Análise de Sentimento no Twitter	25
1.4	Níveis de Análise Textual	26
1.5	Pré-processamento de Texto	26
1.6	Classificação das Abordagens de AS	29
1.7	Áreas de Aplicação da AS	30
1.8	Desafios	32
1.9	Grafos e Redes Complexas	33
1.10	Conceitos de Grafo	33
1.11	Redes Complexas	36
1.12	Centralidade de Nó	38
1.13	Níveis de Centralidade de Nó	39
1.14	Medidas de Centralidade	39
2	TRABALHOS RELACIONADOS	43
3	MATERIAIS E MÉTODO	46
3.1	Conjuntos de Dados	46
3.2	Desbalanceamento do Conjunto de Dados	48
3.3	Pré-processamento de Textos	49
3.4	Grafos de Palavras: Construção	50
3.5	Método	52
3.6	SentiElection	54
3.7	Resumo do SentiElection	58
4	MÉTODOS DO ESTADO-DA-ARTE	60
4.1	Semantic Orientation Calculator (SO-CAL)	60
4.2	Algoritmos de Aprendizado	61
4.3	Análise de Sentimento com Grafos de Palavras (ASGP)	65
5	AVALIAÇÃO DOS MÉTODOS	69
5.1	Métricas de Desempenho	69
5.2	Validação Cruzada K-Fold	70

6	RESULTADOS E DISCUSSÃO	72
6.1	SO-CAL	72
6.2	Aprendizado de Máquina Supervisionado	73
6.3	Análise de Sentimento com Grafo de Palavras	74
6.4	SentiElection	77
6.5	Comparação entre Métodos	80
7	CONCLUSÃO	84
7.1	Trabalhos Futuros	85
	Referências	87

Introdução

Os últimos anos testemunharam o notável crescimento do uso de *smartphones* na maioria das sociedades modernas. O sucesso dos *smartphones* somado ao grande número de usuários conectados à internet transformaram aplicações da *Web 2.0*, como as redes sociais, nos principais veículos de informação no dia-a-dia das pessoas. Hoje, temos milhões de usuários presentes na *web* seja consumindo e/ou produzindo, a todo instante, os mais diversos tipos de conteúdo, os quais podem ser representados por meio de uma simples mensagem de texto ou até mesmo através de vídeos de longa duração.

O *boom* das redes sociais está diretamente ligado ao surgimento do conceito de *Big Data*, um conjunto de dados de alta dimensão formado por informações geradas por usuários (FAHAD et al., 2014). Tais informações podem ser subjetivas, e assim, representar a opinião ou sentimento de uma pessoa em relação a um ou mais assuntos, tais como: produtos, serviços, esportes, entretenimento, saúde, lugares, economia, educação, política, ciência, tecnologia e etc (RAVI; RAVI, 2015).

Muitos pesquisadores consideram o *Big Data* como uma valiosa fonte de informações, pois a partir dele é possível extrair conhecimentos e percepções que retroalimentam os mesmos usuários que as inserem. Como exemplo, podemos citar empresas que utilizam o *feedback* de seus clientes para avaliar a reputação de seu produto. Informações semelhantes podem, por exemplo, auxiliar a formulação de estratégias de marketing e decisões de negócio. Podemos mencionar também os mais variados sistemas de predição capazes de antecipar algum acontecimento em contextos dos mais diversos, como: o resultado de uma campanha eleitoral, a identificação de fraudes, a previsão do tempo, a detecção de áreas de risco e regiões com surtos de doenças, a predição de fenômenos socioeconômicos como variações na bolsa de valores, entre outros (MEDHAT; HASSAN; KORASHY, 2014).

As redes sociais oferecem uma plataforma conveniente para que usuários possam se expressar e interagir com outras pessoas de forma fácil e ágil. Como consequência, bilhões de postagens são criadas a cada mês contribuindo diretamente para o engrandecimento do *Big Data*, e também para o aumento da complexidade de sua análise e extração de conhecimento. Esse fato evidencia a necessidade de métodos automatizados para que seja possível a análise das informações contidas em um imenso conjunto de dados, dentro de um intervalo de tempo aceitável.

No contexto desta miríade de informações, os dados na forma de texto livre atraem a atenção dos pesquisadores da área de Processamento de Língua Natural (PLN)¹. Ao longo dos últimos anos, muitos pesquisadores e empresas têm tido interesse em analisar publicações de internautas em redes sociais com o propósito de obter opiniões dos usuários sobre determinado assunto, produto ou serviço. Para realizar tais tarefas, sistemas baseados em processos de PLN têm auxiliado os estudiosos na extração e análise automática de conteúdo em larga escala, dando origem a uma área específica de conhecimento conhecida como Análise de Sentimento (AS) ou Mineração de Opinião (MO).

Conceitualmente a análise de sentimentos é o estudo computacional de opiniões, sentimentos, atitudes e emoções expressas em textos sobre determinado assunto (LIU, 2012). Vários autores consideram AS e MO como termos intercambiáveis, ou seja, possuem o mesmo significado. Contudo, alguns pesquisadores distinguem a AS como a área que identifica e analisa o sentimento contido em textos com o objetivo de classificá-los como neutro, positivo ou negativo; enquanto que a MO extrai e processa a opinião das pessoas em relação às entidades, isto é, de quem ou sobre o quê está sendo feito um comentário. Embora exista esta sutil diferença, consideramos aqui que AS e a MO representam o mesmo campo de estudo, e por conta disso, neste trabalho ambos os termos serão utilizados de forma equivalente.

Brevemente podemos explicar que uma aplicação de mineração de opinião compreende duas grandes etapas, que são: a) a etapa de extração ou recuperação dos sentimentos, ou ainda, dos atributos textuais que envolvem os sentimentos e, posteriormente; b) a etapa de classificação destes sentimentos. A classificação usualmente atribui a um texto um dos seguintes rótulos: ‘positivo’, ‘negativo’, ou ‘neutro’. Focando na etapa de classificação, notamos que a AS dispõe de diversos métodos para analisar e processar as informações contidas em grandes conjuntos de dados. As abordagens mais conhecidas também podem ser divididas em três grandes grupos, que são:

1. As baseadas em léxicos, ou seja, abordagens dependentes de um acervo de palavras rotuladas, ou pré-classificadas, de uma determinado língua;
2. As supervisionadas através de aprendizado de máquina (AM), as quais, embora muito eficientes, demandam um trabalho de rotulação de textos para montar uma base de aprendizado (treinamento); e, finalmente,
3. As abordagens híbridas que associam duas ou mais tipos de abordagens para classificação de sentimento em textos, por exemplo a junção de recursos léxicos e algoritmos de aprendizado de máquina.

¹ Subárea da Inteligência Artificial que estuda formas de se utilizar computadores para compreensão e manipulação da linguagem humana.

A existência de léxicos e corpora² rotulados ajuda a amenizar as dificuldades encontradas durante estes últimos processos citados de análise de sentimento, porém tais recursos ainda são escassos para muitos idiomas, como o português. Em outros idiomas, como o inglês, a língua explorada na maioria dos trabalhos científicos da AS, estes recursos são mais abundantes. Enfatizamos que para expandir as pesquisas sobre AS em outros idiomas é necessário criar novos recursos ou achar uma forma de se aproveitar os recursos que já existem.

À parte as abordagens citadas anteriormente, a utilização da Teoria de Grafos para AS emergiu com mais vigor nos últimos anos em aplicações diversas (SCHEIBLE et al., 2010; AISOPPOS et al., 2012; GAO et al., 2015; GIACHANOU; CRESTANI, 2016). Lembramos que o uso de grafos na área de AS surgiu há cerca de duas décadas como sugestão para elaboração de léxicos de sentimentos (HATZIVASSILOGLOU; MCKEOWN, 1997), mas só recentemente vimos essa abordagem receber uma atenção maior da comunidade acadêmica, principalmente em casos que o processamento de textos deve considerar a estrutura original da sequência das palavras no texto. Além disso, recentemente surgiu a metodologia de analisar sentimentos através de grafos de palavras (ASGP) que, por sua vez, combina a construção de grafos de palavras, e seus algoritmos de similaridade, com classificadores automáticos (VIOLOS et al., 2016). Poderemos ver no Capítulo 2, Trabalhos Relacionados, página 43, uma revisão comentada destes trabalhos acadêmicos nesta área.

Neste trabalho atacamos o problema de análise de sentimento em mensagens do Twitter utilizando a abordagem de grafos de palavras e, através de nossa pesquisa bibliográfica, inovamos com a utilização do conceito de medidas de centralidade global de nós em grafos. Apresentamos o método chamado SentiElection, o qual classifica determinado texto com base nas medidas de centralidade de seus termos em relação à medidas de grafos formados a partir de textos previamente categorizados. Além disso, fazemos a comparação da performance do SentiElection com abordagens de aprendizado de máquina, léxica e grafos.

Objetivo

Sabemos que a centralidade de nó pertencente a uma rede complexa é uma medida capaz de identificar os vértices centrais dessa rede, ou seja, seus pontos focais (OPSAHL; AGNEESSENS; SKVORETZ, 2010). Nosso objetivo é, a partir de um conjunto de *tweets* já rotulados, construir dois grafos distintos de palavras, um representando a classe de *tweets* positivos e o outro a classe de *tweets* negativos. Nossa hipótese é que um *tweet* desconhecido ao ser inserido nos dois grafos anteriores, dos *tweets* positivos e negativos, altera de modo diferente as centralidades dos nós destes grafos.

² Plural de corpus. Corpus: conjunto de documentos sobre determinado tema.

Também faz parte dos objetivos deste trabalho responder às seguintes questões:

1. É possível e viável realizar uma classificação de texto baseado na centralidade de seus termos?
2. O quão eficiente seria a classificação feita utilizando o método proposto em comparação às outras abordagens existentes como a léxica, aprendizado de máquina e a ASGP de Violos et al. (2016)?
3. Quais as limitações deste projeto e seus cenários de aplicação?

Contribuições Científicas Esperadas

A extração de sentimento baseada em grafos é um estudo ainda recente na área de AS. Os trabalhos já realizados nessa área geralmente usam a estrutura de grafos com a intenção de formar conjuntos de léxicos rotulados ou, com mais frequência, utilizam grafos para explorar as propriedades de relacionamentos entre indivíduos em redes sociais (GIACHANOU; CRESTANI, 2016).

Esperamos que a análise de sentimento baseada em grafos, usando medidas de centralidades de nós, torne-se uma abordagem factível e seja considerada uma solução alternativa para os problemas de classificação de textos no mundo real. Além disso, também contamos que este projeto e seus resultados sejam importantes para a literatura científica, servindo de inspiração para o surgimento de novos trabalhos que fortaleçam as pesquisas em análise de sentimento.

Organização do Projeto

Este trabalho está organizado da seguinte maneira: no Capítulo 1 são introduzidos conceitos fundamentais para a compreensão desta dissertação, tais como o conceito de análise de sentimento, seus desafios e área de aplicação. Neste capítulo também explicamos definições acerca de grafos e redes complexas, bem como explicitamos as medidas de centralidade usadas nesta pesquisa.

No Capítulo 2 apresentamos trabalhos relacionados a análise de sentimento dando ênfase àquelas pesquisas que utilizaram Teoria dos Grafos. Nestes trabalhos podemos analisar outras formas de utilização dessas estruturas de dados no estudo de mineração de opinião.

O Capítulo 3 detalha os conjuntos de dados utilizados neste trabalho, bem como as etapas do método proposto. Já no Capítulo 4, descrevemos algumas técnicas de AS, já presentes no estado-da-arte, e que tiveram seus desempenhos comparados ao SentiElection.

O Capítulo 5 apresenta as formas e métricas de avaliação dos métodos implementados, produzindo uma série de resultados, os quais são exibidos e discutidos no Capítulo 6, e por fim; no Capítulo 7 apresentamos as conclusões obtidas após os experimentos, tornando-nos capazes de responder os questionamentos levantamos no objetivo do trabalho. Além deste tópicos, neste capítulo também discorreremos sobre eventuais trabalhos futuros propondo algumas melhorias no SentiElection.

Referencial Teórico

Visando uma melhor compreensão do projeto, neste capítulo iniciamos apresentando uma visão geral sobre a plataforma Twitter e a definição de análise de sentimento juntamente de diferentes tipos de abordagem, áreas aplicação e desafios. Além disso, introduzimos os principais conceitos-chave relacionados à Teoria dos Grafos e Redes Complexas, ressaltando a importância da centralidade de vértices em grafos.

1.1 Twitter

Twitter¹ é um serviço popular de *microblog* em tempo real que, atualmente, possibilita seus usuários criarem e compartilharem publicações de até 280 caracteres. Esta rede social é um fenômeno importante diante do seu grandioso número de usuários, além de ser utilizada como ferramenta de comunicação e *marketing* por diversas empresas. Segundo Aslam (2018), em 2018 o Twitter contava com cerca de 326 milhões de usuários mensalmente ativos, produzindo em torno de 500 milhões de mensagens por dia, sendo que 80% desses usuários acessam o serviço através de um dispositivo móvel. Esses números reforçam o motivo pelo qual o Twitter é uma das plataformas mais utilizadas em estudos que objetivam compreender e abordar a opinião pública em relação a determinado tema.

Um *tweet*, como é chamado uma publicação feita no Twitter, aceita praticamente qualquer tipo de postagem, desde que esta respeite os termos de uso da plataforma. Geralmente esse tipo de publicação possui os seguintes elementos em sua estrutura:

1. Texto: a mensagem em forma de texto que se deseja publicar, podendo ser escrita sem qualquer restrição quanto à gramática;
2. *Hashtag*: palavras ou texto precedidos pelo símbolo ‘#’ associados a assuntos ou discussões que se deseja indexar em redes sociais. A capacidade indexadora de

¹ <http://twitter.com/>

um *hashtag* pode direcionar o usuário para uma página que contenha publicações relacionadas ao mesmo tema.

3. Menção de usuário: é possível mencionar perfis de usuários numa publicação para que estes sejam notificados assim que a publicação for criada. Um perfil pode ser mencionado num *tweet* através do nome de usuário precedido pelo símbolo '@';
4. Link: hiperlink que redireciona o usuário para uma página interna ou externa ao Twitter. Devido à limitação de caracteres do *tweet*, geralmente um link é encurtado antes de ser postado; e
5. Imagem ou vídeo: conteúdo multimídia que é reproduzido dentro do mesmo *frame* de um *tweet*.

A Figura 1 ilustra uma publicação do Twitter contendo muitos elementos estruturais possíveis.

O Twitter, por si só, facilita quem deseja obter uma parte de seu gigantesco conjunto de dados, uma vez que a plataforma permite que *tweets* criados por contas públicas possam ser coletados em tempo real, através do mecanismo de *streaming* contido em sua API. É possível determinar certos parâmetros na busca como, termos-chaves, localização geográfica, idioma e data de publicação, a fim de se filtrar ainda mais o conteúdo retornado pela aplicação. Essa comodidade possibilitou a criação de diversos conjuntos de dados de *tweets*, hoje existentes na *web*.

1.2 Análise de Sentimento

Atualmente é um hábito bastante comum consultarmos a *web* em busca de informações e opiniões sobre determinado produto ou serviço antes de adquiri-lo. Desejamos ter conhecimento da experiência de alguém que já possui o objeto desejado ou já usou o serviço pretendido para avaliarmos se este corresponderá às nossas expectativas. Consequentemente, há a necessidade de visitarmos diversos sítios da internet, preferencialmente os blogs, as redes sociais e as páginas de produtos e/ou serviços de *e-commerces* em busca dessas experiências. Por outra via, existem também diversas empresas interessadas em saber como seus produtos e serviços estão sendo avaliados pelo público, bem como quais são as suas preferências. Todo esse conhecimento opinativo colocado na *web* é útil se desejamos ter ciência sobre o que está bom, ruim ou o que pode ser melhorado para atingir o maior grau de satisfação e recomendação dos usuários ou clientes.

Entretanto, o excessivo número de comentários (*feedbacks*) dificulta a análise manual de cada experiência, tornando muitas vezes o processo inviável de ser executado

Figura 1 – Exemplo de *tweet*.

por humanos. Como agravante, temos a existência de comentários que nada acrescentam à avaliação final do produto, o chamado *spam* (JINDAL; LIU, 2007). Também há a existência de opiniões tendenciosas enquanto outras são expressas de forma irônica. Cabe à área de análise de sentimentos a tarefa de lidar com essas dificuldades e tentar chegar à uma opinião generalizada sobre determinado assunto ou objeto.

De acordo com Liu e Zhang (2012) a análise de sentimento, ou mineração de opinião, é a área que analisa e extrai opiniões, sentimentos, atitudes e emoções expressas em textos acerca de entidades, as quais podem ser: eventos, assuntos, produtos e serviços. Os sistemas de análise de sentimento tentam classificar de forma automática o sentimento central e inerente ao conteúdo analisado. A maioria dos pesquisadores considera sentimentos assumindo valores de ‘positivo’ e ‘negativo’, enquanto outros vão além desses dois valores, reconhecendo o ‘neutro’ como uma terceira polaridade.

Formalmente, o problema de classificação de sentimentos pode ser definido como um problema de classificação binária:

Dado um conjunto D de documentos e um conjunto de n classes $C = \{C_1, C_2, \dots, C_n\}$ representando seus respectivos rótulos, um classificador de sentimentos é uma função binária $F : D \times C \rightarrow \{0, 1\}$ (BAEZA; RIBEIRO-NETO, 2011). Essa função é encarregada de atribuir o valor 0 ou 1 para uma série de pares (d_j, c_i) , tal que $d_j \in \{D\}$ e $c_i \in \{C\}$. Portanto, se o documento $d_j \in c_i$, então $F(d_j, c_i) = 1$, caso contrário, será 0 (FERREIRA et al., 2014).

Considerando uma classificação ternária, frequentemente a polaridade de textos é mapeada para o intervalo $[-1, 1]$, assumindo que -1 é a polaridade mais negativa possível e 1 é a mais positiva. Para a polaridade neutra é atribuído o valor 0 , porém há uma ambiguidade em relação a este valor pois ao mesmo tempo que 0 representa um texto ausente de sentimento, também pode representar uma mistura relativamente balanceada de conteúdo positivo e negativo em um mesmo documento (KOPPEL; SCHLER, 2006). É sabido que textos neutros são difíceis de serem avaliados, mesmo por seres humanos e, por conta disso, dados dessa categoria são, por vezes, omitidos em estudos a fim de simplificar o problema.

1.3 Análise de Sentimento no Twitter

A mineração de opinião é um dos desafios da Computação. Diversos trabalhos nas áreas de Processamento de Língua Natural e Inteligência Artificial têm sido realizados com base em textos opinativos escritos por pessoas que compartilharam alguma experiência ou opinião sobre determinado assunto ou acontecimento (AISOPOS et al., 2012; CUI et al., 2011; GIACHANOU; CRESTANI, 2016; LIU; ZHANG, 2012; THAKKAR, 2013; FANG; ZHAN, 2015).

Essas opiniões, popularmente chamadas de *reviews*, podem ser facilmente encontradas em *blogs*, *webforums*, sites de críticas e principalmente na área de depoimentos sobre produtos em sites de varejistas; porém quando o assunto é analisar textos publicados em plataformas de rápida interação, incluindo aqui textos curtos, a análise de sentimento se torna ainda mais difícil. Os *tweets* são formados por mensagens curtas, ausentes de padrões de escrita, e de característica coloquial, além de serem conhecidos como ‘dados não estruturados’ por parte dos pesquisadores da AS (MUKHERJEE et al., 2013). Essa natureza informal dos dados possibilita que uma mesma palavra tenha diferentes formas de escrita, o que resulta numa dificuldade para a análise computacional do texto quando comparado à mineração de opinião em documentos menos coloquiais.

Para amenizar os impactos da informalidade textual e a não estruturação dos dados, podemos recorrer à tarefa de pré-processamento de texto, também conhecida como normalização textual, a qual é responsável pela atenuação do ruído presente nos dados ao aplicar uma série de técnicas de normalização textual sobre eles. Além disso, o pré-processamento textual apresenta o benefício de uniformizar os dados eliminando características do texto que podem dificultar seu processamento e classificação.

Apesar dos problemas citados, podemos destacar duas vantagens de se trabalhar com dados do Twitter, que são: 1) fácil aquisição de dados recentes, em grande volume, através de sua API, e; 2) textos curtos, dado o número de termos ou palavras, normalmente

exigem menos tempo computacional para serem processados se comparados a textos longos como críticas de filmes ou resenhas de livros.

1.4 Níveis de Análise Textual

Liu (LIU, 2010) afirma que a classificação de sentimento pode ser feita seguindo três níveis de abstração. São eles:

- **Nível de documento:** Neste tipo de análise, tem-se o objetivo de classificar a polaridade expressa num documento, se positiva, negativa ou neutra, em relação a uma entidade. Vale ressaltar que é preciso que os documentos processados opinem a respeito de apenas uma entidade (PANG; LEE; VAITHYANATHAN, 2002);
- **Nível de frase:** A análise textual é executada através da segmentação de todo o conteúdo em frases, seguida de uma verificação com o objetivo de selecionar as frases que expressam subjetividade, para então definir a polaridade de cada uma (YU; HATZIVASSILOGLU, 2003), e;
- **Nível de aspecto:** O nível de aspecto é uma análise mais detalhada e bastante útil quando se deseja saber o sentimento de cada atributo ou característica da entidade comentada. Para melhor compreensão, considere a frase ‘Eu gosto deste restaurante, embora seja distante, a comida é sempre ótima!’, podemos notar que existe três opiniões presentes na frase, cada uma sobre um aspecto, duas positivas, sobre o restaurante e a comida, e uma negativa, sobre a sua localização (THET; NA; KHOO, 2010). Caso a classificação de sentimento fosse centrada apenas no atributo ‘localização’, este documento seria classificado como negativo.

1.5 Pré-processamento de Texto

O pré-processamento (ou normalização) de textos é uma das principais atividades precedentes à análise dos dados. Esse processo consiste em fazer um tratamento dos documentos removendo ou substituindo tudo o que não for essencial ou que prejudique a inspeção de seu conteúdo, seja para uma tarefa de classificação, agrupamento, extração de conhecimento, seleção de características, tradução automática, entre outras (UYSAL; GUNAL, 2014). Basicamente, o principal objetivo dessa atividade é transformar sentenças informais em textos escritos na sua forma padrão e completamente gramatical (KAUR; MANN, 2016).

Atualmente a normalização de textos é uma tarefa imprescindível para os sistemas de Processamento da Linguagem Natural. Como mencionado anteriormente, o crescimento tecnológico permitiu que as pessoas pudessem se expressar de forma livre na *web*,

principalmente em redes sociais e aplicativos de mensagens instantâneas, necessitando apenas de um *smartphone* com acesso à internet. Consequentemente a linguagem natural, especialmente em redes sociais, foi adquirindo novas características marcantes chegando ao ponto de se tornar um novo ‘idioma’, caracterizado como textos curtos, repleto de gírias, erros ortográficos e de digitação, abreviações informais, palavras enfatizadas, *hashtags*, links, imagens, vídeos, *emoticons* e *emojis*.

A existência de tais características somada à liberdade da forma de expressão, seja pela escrita formal ou não, torna o conteúdo presente no *Big Data* das redes sociais bastante diversificado. Por exemplo, é possível representar uma mesma palavra de diversas formas diferentes seja pelo uso de *emojis*, de abreviação, ou através da escrita não formal ou mesmo em sua forma original. Para o cérebro humano, compreender uma palavra expressa em diferentes formatos não costuma ser um desafio, mas para as máquinas essa não é uma tarefa tão simples. Em virtude dessa dificuldade, são raras as situações que dispensam a etapa de pré-tratamento dos dados antes de uma análise mais profunda.

A seguir iremos descrever as principais técnicas de normalização textual utilizadas na área de PLN:

- **Remoção de elementos irrelevantes:** Por conta da variabilidade de fontes de informação presentes na *web*, o conjunto de textos recuperado pode conter diferentes tipos de elementos como: símbolos, *emojis*, links, abreviações coloquiais, *stop words*², *hashtags*, marcação HTML, números, pontuação, caracteres especiais, entre outros. É necessário que o pesquisador defina quais elementos que compõe o texto serão importantes para seu estudo e remova aqueles que não interessam para se evitar um processamento desnecessário (DENNY; SPIRLING, 2017) ou um prejuízo ao processamento;
- **Substituição de elementos:** Para alguns estudos, elementos não usualmente reconhecidos como palavras podem ser relevantes, fazendo parte da informação ou do sentido que o texto transmite. Tais elementos podem ser dígitos, *emoticons*, *emojis*, símbolos, gírias e outros. Para facilitar o processamento desse tipo de texto, recomenda-se substituir os itens em questão por palavras que expressem o mesmo sentido. Por exemplo, na frase “O filme Gravidade foi vencedor de 7 Oscars deste ano.”, o número ‘7’ poderia ser substituído por sua forma extensa, então a frase ficaria: “O filme Gravidade foi vencedor de sete Oscars deste ano.”. Também podemos trocar *emoticons* que representam determinada expressão por um termo, como é o caso de ‘=(’ que poderia ser facilmente trocado pela palavra ‘triste’ ou algum de seus sinônimos;

² Termos de baixa relevância, ou seja, termos que não acrescentam informações importantes ao texto. Mesmo com a eliminação dessas palavras o sentido da frase permanece inalterado.

- **Palavras com o mesmo tamanho de caixa:** Com o objetivo de reduzir a complexidade no processamento de textos, esta técnica propõe que todas os termos (palavras) do texto devem estar representados de forma minúscula ou maiúscula em sua totalidade (DENNY; SPIRLING, 2017). Essa técnica não deve ser usada caso a aplicação que analisará os dados necessite ser sensível ao tamanho da caixa da fonte;
- **Stemização:** Segundo (PORTER, 1980), stemização ou *stemming* refere-se ao processo de remoção das inflexões dos vocábulos, tornando-os assim mais simples. Os vocábulos processados são representados pelo seu *stem*. Tecnicamente, busca-se por uma sub-palavra única e não ambígua que represente um vocábulo e suas variações, mais conhecida como radical (ALVARES, 2005). Para uma maior clareza, se aplicarmos o stemização sobre os termos ‘gato’, ‘gata’, ‘gatos’ e ‘gatas’ o radical a ser obtido será ‘gat’. É vantajoso utilizar este método se quisermos identificar similaridade a partir da morfologia das palavras de um documento, além de se reduzir a complexidade de busca por termos correspondentes;
- **Lematização:** Seu objetivo é reduzir os elementos derivados de uma mesma palavra para sua forma canônica (lema) (JOEL; NADA; DUNJA, 2004), ignorando afixos e tempos verbais. Para exemplificar, se aplicássemos a lematização sobre as palavras ‘vizinha’, ‘vizinhas’, ‘vizinhos’ teríamos como resultado a palavra ‘vizinho’. Stemização e lematização são bastante úteis quando se deseja diminuir a quantidade de vocábulos e abstrair seus significados;
- **Expansão de vocábulos:** É comum nos dialetos a existência de duas palavras representadas por apenas uma. Denomina-se essa característica de ‘forma contraída’ e esta ocorre frequentemente na língua inglesa. Geralmente a contração é feita pela combinação de um pronome ou substantivo com um verbo (CAMBRIDGE, 2017). A presente técnica recomenda que palavras contraídas sejam expandidas para se evitar contrações ambíguas que acabariam prejudicando a análise do texto. Podemos citar como contração ambígua o termo ‘He’s’ que pode ser a forma reduzida de ‘He is’ ou ‘He has’;
- **Thesaurus:** O *thesaurus* é uma coleção pré-compilada de palavras relacionadas a um determinado contexto. Normalmente para cada palavra pertencente a esse dicionário existe um conjunto de sinônimos, o tipo mais comum, mas outras vezes temos a representação de antônimos também além de outras palavras correlacionadas. A motivação para se utilizar o *thesaurus* está na possibilidade de fazer uma indexação e buscas a partir de um conjunto padrão de termos (KATARIYA et al., 2015);
- **Normalização de palavras enfatizadas:** É muito comum as pessoas expressarem seus sentimentos de felicidade ou raiva enfatizando palavras no texto. Essa ênfase normalmente é feita digitando todas as letras de uma palavra usando letras maiúsculas

ou prolongando-se alguma(s) de suas letras. Por exemplo: na frase ‘It is awesome!’ com ênfase pode ser escrita nas formas: ‘It is AWESOME!’, ou ‘It is awesoomeee!’, ou até mesmo ‘It is AWEEEEESOOOME!!!!’. Dadas estas diferentes grafias, é preciso normalizar a palavra enfatizada para sua forma regular, ou seja, sem letras repetidas e de preferência escrita usando-se letras de mesmo tamanho, a fim de se evitar ter mais de uma representação para determinada palavra num documento.

Vale ressaltar que os métodos do pré-processamento de textos são opcionais ficando a cargo do pesquisador selecionar quais métodos deverão ser utilizados em seu projeto. Também é preciso estabelecer uma sequência de processamento, visto que, a variabilidade da ordem de execução das técnicas pode produzir resultados distintos.

1.6 Classificação das Abordagens de AS

Existem diversos tipos de abordagens envolvendo a mineração de opinião, e a escolha por qual delas se utilizar deve ser baseada na natureza da tarefa a ser desempenhada e também no tipo de conhecimento que os interessados desejam obter (THAKKAR, 2013). Tais abordagens podem ser classificadas em 4 grandes grupos:

- **Léxica:** Esta abordagem utiliza um dicionário de palavras pré-rotuladas com seus respectivos sentimentos (THAKKAR, 2013). Por exemplo, a palavra ‘excelente’ possui a polaridade positiva, enquanto ‘ruim’ tem sentimento negativo. Os trabalhos que usam essa técnica geralmente atribuem um sentimento a um documento ou sentenças contabilizando o número de palavras positivas e negativas que ele contém. Portanto, se em um documento existir mais palavras positivas do que negativas, ele será considerado um texto positivo e vice-versa.

Alguns trabalhos vão além de uma simples contabilidade de vocábulos positivos ou negativos ao utilizarem pesos que representam a intensidade do sentimento de cada palavra, isto é, uma escala de sentimento. A título de exemplo, a palavra ‘excelente’ possui uma intensidade positiva maior que ‘bom’. Representando essa intensidade numa escala decimal, poderíamos atribuir o valor 3 para ‘excelente’ e, 1 para ‘bom’. Já para o termo ‘horrível’ um valor adequado seria -3 e para o vocábulo ‘ruim’ -1 . Assim a opinião de um documento é definida através do somatório dos pesos do número total de palavras mapeadas no dicionário léxico. Resumidamente, podemos dizer que a classe léxica de aplicações de AS tem o objetivo de calcular a orientação semântica de palavras ou frases de um documento;

Aprendizado de máquina (AM): Essa forma de classificação é uma das mais utilizadas nos trabalhos de análise de sentimento. Os algoritmos baseados em

aprendizagem podem ser divididos em duas grandes categorias: a) Algoritmos supervisionados: consistem em escolher atributos (posteriormente agrupados em vetores de características) de um corpus já rotulado para o treinamento de um classificador. Esse classificador posteriormente irá classificar novos textos apresentados; e b) Algoritmos não-supervisionados: são os algoritmos capazes de classificar um novo texto sem a necessidade de se basear nas anotações e nos atributos de textos previamente classificados. Estes algoritmos tem como base o relacionamento entre os dados para então agrupá-los em categorias de acordo com suas características comuns;

Híbrida: A evolução da AS permitiu que vários pesquisadores apostassem em abordagens híbridas, ou seja, que combinam duas ou mais abordagens para a construção de uma técnica desempenho esperado melhor que as técnicas anteriores. Podemos citar o trabalho de Mudinas, Zhang e Levene (2012) no qual os autores fizeram uso da eficiência do AM e da rapidez e estabilidade do método léxico para então construir o sistema de análise de sentimento chamado *pSenti*. O *pSenti* usa léxicos provindos de fontes abertas para detecção de expressões e *emoticons* de uso comum, e assim pode gerar características como negações, intensificadores e palavras rotuladas, as quais foram usadas para o treinamento de um classificador tipo *Support Vector Machine*, ou SVM. Classificadores assim são geralmente responsáveis pela realização da predição de sentimento de textos. Essa combinação trouxe excelentes resultados chegando a obter 82,3% de acurácia.

Grafos: A AS baseada em grafos de relacionamento entre usuários é bastante utilizada no contexto de redes sociais, especialmente no Twitter, a fim de se descobrir o grau de influência que um usuário tem sobre outro ou mesmo sobre toda a rede (GIACHANOU; CRESTANI, 2016). Para esse tipo de trabalho, normalmente são selecionados *tokens* (*emojis*, *hashtags*, palavras) de mensagens postadas por um usuário muito referenciado da rede (um *hub*) para se avaliar como determinado assunto está sendo recebido pelas pessoas. Outros trabalhos utilizam a estrutura de grafos para representar e agrupar léxicos com sentimento já marcado. Para formar um grafo, as palavras ou as frases são os nós da estrutura. Estes grafos são comparados com outros grafos já rotulados em busca de um valor de similaridade que irá definir o sentimento daqueles grafos ainda não rotulados (GIACHANOU; CRESTANI, 2016). Neste trabalho usamos este tipo de grafo, os grafos de palavras.

1.7 Áreas de Aplicação da AS

Por ser uma área que ganhou grande destaque nos últimos anos, a análise de sentimento está sendo aplicada a muitas áreas de atividade. Esta subsecção apresenta algumas das principais aplicações desse estudo.

- **Comércio eletrônico:** Grande parte dos sites de comércio eletrônico reservam um espaço da página de cada produto para comentários de usuários que já o adquiriram (VOHRA; TERAIIYA, 2013). Além da área de texto para o depoimento, também é bastante comum haver campos onde o opinador deixa sua nota acerca de atributos ou aspectos específicos do produto ou serviço como: preço, custo benefício, tempo de entrega, atendimento, etc. A medida que uma opinião é deixada, uma nota geral do produto é calculada através do número de estrelas ou nota que cada atributo recebe do usuário, fazendo assim a análise do produto para futuros novos consumidores. Essas informações são ideais para aqueles que desejam recuperar e analisar dados com opiniões sobre diversos aspectos de uma entidade num único texto, e também bastante úteis para empresas que desejam saber a reputação de suas marcas;
- **Cenário político:** Também é muito comum o uso da AS para descobrir o sentimento relacionado à cada partido ou candidato durante debates políticos e campanhas eleitorais. AS pode também monitorar conteúdos transmitidos ao vivo pela televisão ou rádio. Geralmente, aplicam-se essas técnicas em redes sociais, por ser o meio mais utilizado pelo público para expressar seu posicionamento e também mais prático. Também é comum a construção de sistemas preditivos que utilizam um grande conjunto de opiniões da população expressas durante as campanhas eleitorais com a finalidade de predizer quem irá ganhar uma eleição (VOHRA; TERAIIYA, 2013);
- **Sistemas de decisão:** A mineração de opinião pode ser aplicada para auxiliar sistemas de decisão como é o caso dos investimentos financeiros na bolsa de valores. É sempre possível encontrar na *web* textos relevantes e atualizados sobre as mais diversas empresas do mercado. O conjunto dessas informações recuperadas poderá ser processado por uma aplicação de mineração de opinião que terá como resultado as possíveis informações necessárias para o funcionamento de um sistema de negociação automatizado (THAKKAR, 2013);
- **Surto de doenças e áreas de risco:** Possíveis regiões com surtos de doenças e áreas propensas a ocorrerem desastres naturais podem ser mapeadas com a ajuda da AS. Através de postagens em blogs e redes sociais é possível identificar as regiões de maior concentração de sentimentos que expressem perigo ou sofrimento. Tais áreas podem ser evacuadas e isoladas de modo que uma epidemia não se alastre para outras regiões ou que pessoas consigam escapar de prováveis tragédias naturais (CHEW; EYSENBACH, 2010);
- **Saúde:** Na área de saúde, é frequente a utilização da AS através da abordagem de aprendizado de máquina sobre textos de pacientes em hospitais. Através de uma profunda análise dos textos referentes ao tratamento recebido pelos profissionais, exames

realizados, condições de limpeza do local e outras características, os responsáveis pela gerência do hospital poderão ter uma visão mais detalhada sobre as percepções dos pacientes e ter ciência de onde as mudanças podem ter maior impacto para a melhoria da experiência do paciente (KHAN; KHALID, 2016). Informações mais amplas sobre disseminação da informação sobre doenças é uma outra aplicação da AS sobre tweets (ALMANSA et al., 2014).

1.8 Desafios

Embora já exista uma considerável quantidade de pesquisas relacionadas à área de pesquisa sobre análise de sentimento, como toda grande área de estudo sempre existirão problemas que ainda são um desafio para a comunidade. Assim como a subsecção anterior, iremos aqui discorrer sobre algumas das principais questões consideradas ‘em aberto’ na área de AS.

- **Ambiguidade:** O verdadeiro sentido de uma palavra ou conjunto de palavras é inerente ao contexto onde está inserida. Podemos citar a palavra ‘gelado’ que pode ser ter um sentido positivo se estivermos falando de sorvete e negativo se falamos de pizza;
- **Negações:** É preciso tratar negações em frases com bastante cuidado, caso contrário, poderá se obter resultados completamente incorretos. Na frase ‘Existe boa chance de eu não pegar trânsito ruim.’ a presença de uma palavra negativa pode afetar o resultado da avaliação dessa frase caso seja retornado a polaridade negativa quando na verdade deveria ser positiva (KOLKUR; DANTAL; MAHE, 2015);
- **Sarcasmo:** Algumas vezes, até para nós, humanos, é difícil a percepção de uma ironia presente em um texto, visto que a entonação da voz é um fator que na maioria das vezes nos ajuda a compreender a ironia presente na fala.
- **Deteção de Spam:** Um dos maiores problemas quando vamos consultar a reputação de um produto é a presença de depoimentos falsos e tendenciosos. Em busca de promover positivamente seus produtos, várias empresas fazem uso de *spam* automático contendo notícias tendenciosas ou até mesmo contratam pessoas para criarem falsos depoimentos e vídeos para então atrair mais clientes (VOHRA; TERAIYA, 2013);
- **Conteúdo multimodal:** Em alguns casos a orientação semântica de um documento é resultado da união semântica de suas palavras e complementos como imagens e vídeos. Tais complementos podem conter um sentido mais enfático do que as palavras, entretanto extrair características semânticas desses elementos ainda não é

uma tarefa trivial e consequentemente essa é uma das áreas pouco exploradas pela AS (GIACHANOU; CRESTANI, 2016);

- **Textos multilíngues:** A possibilidade de se escrever textos em diversos idiomas, e até mesmo misturá-los na mesma frase, acaba dificultando a análise de sentimento do texto. É preciso escolher uma estratégia de processamento que seja independente da língua do documento ou usar ferramentas de tradução para se conseguir uma maior precisão da análise (GIACHANOU; CRESTANI, 2016).

1.9 Grafos e Redes Complexas

O presente trabalho propõe uma abordagem híbrida para AS usando a estrutura de grafos de palavras para a representação de redes de sentimento. Propomos também a utilização de um conjunto de medidas de centralidade de nós em grafos para calcular a importância de um documento para uma classe de documentos, seja esta uma classe de documentos positivos ou negativos. Mas antes de apresentar a metodologia do projeto, é necessário elucidar algumas definições básicas acerca da Teoria de Grafos e sobre redes complexas para um melhor aproveitamento na leitura das próximas seções.

Assim como ocorre na AS e MO, grafos e redes complexas podem ser usados como termos equivalentes na literatura, porém há pesquisadores da área que preferem utilizar o termo ‘grafos’ em pesquisas matemáticas, enquanto utilizam ‘redes complexas’ em estudos que objetivam representar e analisar o relacionamento não trivial entre entidades do mundo real (ESTRADA, 2013). Em outras pesquisas também utiliza-se ‘redes complexas’ para grafos com um elevado número de nós e/ou arestas. Para este projeto, decidiu-se adotar os dois termos de forma equivalente, porém preferencialmente em momentos distintos; sendo a palavra ‘grafos’ usada com mais frequência na parte conceitual e ‘redes complexas’ quando nos referimos a um grafo que, com o passar do tempo, adquiriu uma estrutura de alta complexidade e robustez.

1.10 Conceitos de Grafo

Muitas destas definições são apresentadas com mais detalhes em (ESTRADA, 2013; STEEN, 2010).

Definição 1.1. Um **grafo não-direcionado** $G = (V, E)$ é composto de um par ordenado de dois conjuntos, $V = \{v_1, v_2, \dots, v_n\}$ e $E = \{e_1, e_2, \dots, e_k\}$, tal que V representa a coleção de *vértices*, ou nós, e E simboliza o conjunto de *arestas*. Uma aresta $e \in E$ é um par ordenado de vértices $\{u, v\} \in V$. Veja Figura 2.

Figura 2 – Exemplo de grafo não direcionado.

Definição 1.2. Um **grafo direcionado**, ou digrafo, $G = (V, A)$ é composto de um par ordenado de conjuntos $V = \{v_1, v_2, \dots, v_n\}$ e $A = \{a_1, a_2, \dots, a_k\}$, tal que V representa a coleção de vértices e A simboliza o conjunto de arestas, na qual cada aresta possui uma direção associada. Veja Figura 3.

Figura 3 – Exemplo de grafo direcionado.

Definição 1.3. Um **multigrafo** é um grafo não dirigido que pode possuir arestas múltiplas para os mesmos vértices finais, ou seja, arestas no formato $\{v, v\}$. Assim os multigrafos também podem possuir *laços*, aresta para uma mesmo vértice. Um grafo sem laços é chamado de grafo simples. Veja Figura 4.

Figura 4 – Exemplo de multigrafo.

Definição 1.4. Uma aresta e conecta dois vértices adjacentes u e v , portanto $e = \{u, v\}$. Se u é adjacente a v , então dizemos que u é **vizinho** de v . Quando o grafo é ponderado cada uma de suas ligações terá um valor, também conhecido como peso. O conjunto de vértices conectados a u através de uma aresta é chamado *vizinhança* de u .

Definição 1.5. O **grau** de um vértice $v \in V(G)$ é definido como o somatório do número de arestas adjacentes a v . Num digrafo existem dois tipos de graus de vértices, o grau de saída $d_{\text{out}}(v)$ que representa o número de arcos que partem de v , e o grau de entrada $d_{\text{in}}(v)$, ou seja, o número de arcos que incidem em v .

Definição 1.6. Um grafo P na forma de $E = \{v_0v_1, v_1v_2, \dots, v_{n-1}v_n\}$ é conhecido como **caminho**, sendo v_0 o nó inicial e v_n o nó final. O tamanho desse caminho é representado pelo número de arestas presentes em E .

Definição 1.7. A **matriz de adjacência** A de um grafo é uma de suas representações alternativas ao modelo tradicional composto por nós e arestas. Veja Figura 5. Nessa representação utiliza-se uma matriz quadrada $n \times n$, para qual n representa o número de nós do grafo G , enquanto cada linha e cada coluna correspondem a um de seus vértices. Num grafo direcionado consideramos o valor $A[u, v] = 1$ se existir uma aresta que emerge do nó u e incide sobre o nó v . De maneira análoga, se num grafo não-direcionado existir uma ligação entre os vértices u e v teremos tanto $A[u, v] = 1$ quanto $A[v, u] = 1$, portanto podemos dizer que a representação matricial A de um grafo não-direcionado é simétrica. Quando não existir uma ligação entre u e v , então a $A[u, v] = 0$. Devemos mencionar também que em grafos ponderados, os elementos diferentes de zero da matriz de adjacência recebem o valor do peso de cada aresta.

Figura 5 – Exemplo de grafo direcionado e sua respectiva matriz de adjacência.

Definição 1.8. Seja G_1 e G_2 dois grafos. Uma função bijetora entre os vértices de G_1 e G_2 , $f : V_{G_1} \rightarrow V_{G_2}$, preserva a matriz de adjacência se para cada par de nós vizinhos u e v no grafo G_1 , os nós $f(u)$ e $f(v)$ também forem adjacentes no grafo G_2 . A bijeção entre G_1 e G_2 é denominada **isomorfismo** de grafos, como mostrado na Figura 6.

Definição 1.9. Um grafo H é um **subgrafo** de um grafo G se seu conjunto de vértices e arestas for subconjunto do conjunto de vértices e arestas de G . Dessa forma

Figura 6 – Isomorfismo entre os grafos G_1 e G_2 .

podemos dizer que um grafo G contém um outro grafo H se algum subgrafo de G é H ou é isomorfo a H .

Definição 1.10. Seja G_1 e G_2 dois grafos. Um **subgrafo comum** de G_1 e G_2 , $cs(G_1, G_2)$, é um grafo H tal que existe subgrafos isomórficos de H a G_1 e de H a G_2 . Chamamos H de ‘máximo subgrafo comum’ de G_1 e G_2 , $mcs(G_1, G_2)$, se não houver outro $mcs(G_1, G_2)$ que tenha mais vértices do que H . Veja Figura 7.

1.11 Redes Complexas

De acordo com Knöbel (2009) ainda não se pode definir com exatidão quando um grafo passa a ser considerado uma rede complexa. De modo impreciso, podemos dizer que uma rede é vista como complexa a partir do momento em que ela cresce de forma natural, ou seja, sua estrutura evolui baseada em eventos aleatórios e constantes, sem o seguimento de um plano de construção que conduza esse crescimento.

A maioria das redes complexas apresentam um grande dinamismo em sua estrutura, além do crescimento constante, ela está sempre sob efeito de mudanças em sua configuração de arestas e vértices, as quais à medida que algumas delas são destruídas outras são adicionadas ao longo do processo garantindo sua robustez (KNÖBEL, 2009).

Podemos citar também duas outras características marcantes das redes complexas como sua topologia não-trivial, a qual dificulta sua visualização e a análise não-automática;

Figura 7 – Máximo subgrafo comum H dos grafos G_1 e G_2 .

e também a variabilidade quanto ao tamanho, direção e capacidade dos tipos de conexões existentes na rede (SILVA; ZHAO, 2016). Outras características podem ser vistas com mais detalhes em Knöbel (2009) e Silva e Zhao (2016). Veja uma ilustração de uma rede complexa na Figura 8.

Redes complexas (RC) é uma subárea de conhecimento de sistemas complexos (SC), sendo este um campo de estudo interdisciplinar e bastante estudado, principalmente por matemáticos, físicos, biólogos, engenheiros e diversos outros profissionais e cientistas. Como consequência, o estudo de RC é constantemente aplicado tanto em situações do mundo real como em demais situações científicas que necessitam de uma abstração formal para representar entidades (nós) e seus mais diversos tipos de relacionamentos (arestas); além de ser uma técnica que auxilia na detecção e compreensão de características e fenômenos específicos presentes nessas redes. Algumas destas características justificam o uso de redes complexas numa vasta gama de exemplos, tais como: mapeamento de usuários e computadores conectados através da *web*, redes de reações bioquímicas, redes neurais do cérebro, sistemas biológicos interativos e acoplados, redes financeiras, tráfego aéreo e terrestre, redes de propagação de informação, além de muitas outras situações (SILVA; ZHAO, 2016).

Alguns atributos ou características destas redes serão úteis para o desenvolvimento deste projeto, tal como as medidas de centralidade.

Figura 8 – Exemplo de rede complexa.

1.12 Centralidade de Nó

A centralidade avalia o quão central ou importante um vértice (uma entidade) ou um relacionamento é para a rede a qual estão inseridos (SILVA; ZHAO, 2016). Assim, pode-se imaginar que um vértice que contém um grande número de arestas adjacentes é ‘mais central’ ou importante do que um outro vértice com menor número de arestas. Trazendo esse conceito para um contexto real, podemos pensar numa rede complexa baseada na estrutura de uma rede social, em que os vértices mais importantes, também chamados de *hubs*, são perfis de pessoas ou empresas famosas que ao transmitirem determinada informação, esta será propagada ao longo da rede com rapidez e poderá atingir milhões de pessoas em questão de segundos.

Na Figura 9 temos um exemplo de cálculo de centralidade dos nós de um grafo baseado no número de conexões que cada um possui. Supondo que o grafo da imagem represente uma sociedade, podemos dizer que C é o indivíduo mais popular, ou mais importante, por ter ligação com cinco pessoas, enquanto A e D são os menos conhecidos tendo uma conexão cada e assim, com grau de influência menor na rede.

Figura 9 – Centralidade de vértices baseado no grau.

1.13 Níveis de Centralidade de Nó

Na literatura existem diversas medidas de centralidade que se diferenciam pela heurística escolhida para identificar pontos e ligações importantes da rede. Estas medidas podem ser classificadas em dois níveis, local e global.

- **Centralidade local:** Representa a conectividade de um nó considerando apenas o seu grau, ou seja, o número de conexões que esse nó possui com seus vizinhos diretos.
- **Centralidade global:** Indica a importância de um nó, dentro de uma rede, considerando não apenas o seu próprio grau mas também o grau de seus vizinhos diretos e indiretos.

Portanto, podemos dizer que um nó é globalmente central quando possui uma posição de importância estratégica na estrutura geral da rede, por outra via, a centralidade local está preocupada com a importância relativa de um ponto focal apenas em sua vizinhança.

1.14 Medidas de Centralidade

Para esta pesquisa o uso de métricas globais de centralidade é mais interessante do que a medida local, pois elas indicam os vértices que se destacam, em termos de conectividade, ao longo de toda uma rede. Assim, podemos calcular a importância geral que um determinado *tweet* tem quando inserido em diferentes grafos de palavras, um grafo composto apenas por *tweets* marcados como positivo e outro grafo contendo apenas *tweets* marcados como negativos.

O SentiElection, classificador proposto neste trabalho, combina o uso de três medidas globais de centralidades de vértices: autovetor, *Katz* e *PageRank*. Tais medidas são apresentadas a seguir.

Autovetor

Segundo Bonacich e Lloyd (2001) essa medida parte do princípio o qual afirma que: um vértice é importante se ele está ligado a outros vértices importantes, mesmo que seu número de ligações não seja alto. A centralidade de autovetor calcula a conectividade de um nó baseado na centralidade de seus vizinhos, ou seja, a importância de um vértice v_i é dependente do grau de importância de seus vértices vizinhos e não só do número de vértices adjacentes que ele possui. Assim, podemos dizer que a centralidade de autovetor de um nó é a soma das centralidades dos seus vizinhos diretos (BONACICH, 2007). Matematicamente, a centralidade de autovetor de um vértice v_i (x_i) de um grafo G é dada pela fórmula:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^n A_{i,j} x_j, \quad (1.1)$$

onde λ é uma constante fixa (autovalor); A é a matriz de adjacência que contém os vizinhos v_j de v_i , x_j a centralidade (autovetor) de um vértice v_j e n é o total de vértices da rede. De forma matricial temos:

$$Ax = \lambda x,$$

sendo que os valores de λ correspondem aos autovalores do grafo e os valores de x correspondem aos seus autovetores. Os valores do autovetor associado ao maior autovalor da matriz representam os nós de maior importância na rede (BONACICH, 1987).

Katz

Baseada na centralidade de autovetor e introduzida pelo estatístico norte-americano Leo Katz (KATZ, 1953), a centralidade de *Katz* mensura o grau de influência relativa de um vértice v dentro de uma rede G através centralidade do seu conjunto de vizinhos imediatos A , e também da centralidade de todos os vizinhos não imediatos que se conectam com v por meio de cada elemento da matriz de adjacência A (AGGARWAL, 2011). Esta medida se diferencia de outras métricas de centralidades que na maioria das vezes consideram somente o menor caminho entre dois pares de nós. Para cada aresta é atribuído um fator de atenuação α^k , que penaliza as conexões feitas entre vizinhos distantes, tal que $\alpha \in [0, 1]$ e k representa o comprimento da distância entre dois vértices comunicantes. Por exemplo, se v_i é vizinho imediato de v_j então $k = 1$, caso contrário $k > 1$. Resumindo, quanto mais distante um vértice v_i for de v_j , a ligação entre eles vai perdendo força.

A centralidade de *Katz* é dada pela seguinte fórmula:

$$C_{Katz}(v_i) = \sum_{k=1}^{\infty} \sum_{j=1}^n \alpha^k (A^k)_{ij}; \quad (1.2)$$

tal que n é o número de vértices de G , α é o peso da ligação entre dois vértices separados por uma distância k , e A é o conjunto de vértices v_j adjacentes a v_i .

PageRank

A medida de centralidade conhecida como *PageRank* foi criada em 1998 pelos fundadores do buscador Google, Larry Page e Sergey Brin, como base de seu motor de busca para o ranqueamento de websites (PAGE et al., 1999).

Considerando um cenário em que temos um usuário que está navegando aleatoriamente pela internet (grafo) e acesse um determinado site A (nó) (ver Figura 10); o *PageRank* do nó A corresponde à probabilidade que o usuário acesse essa página, de forma aleatória, através de um *hiperlink* contido na página anterior. A cada fração de tempo, o usuário prossegue de sua página atual para uma próxima página da *web* escolhida arbitrariamente, na qual é referenciada pela página anterior. A Figura 10 mostra o internauta num nó A, do qual existem ligações para os nós B, C e D; cada uma dessas páginas tem probabilidade de 1/3 de serem visitadas partindo de A.

Figura 10 – Partindo do nó A, o usuário avança aleatoriamente para uma das páginas B, C ou D, com probabilidade de 1/3 cada.

À medida que o usuário prossegue nessa caminhada aleatória de nó para nó, ele visita alguns nós com mais frequência que outros; intuitivamente, esses são nós com muitas arestas vindas de outros nós visitados com frequência. O conceito envolvendo o *PageRank* é que as páginas visitadas com mais frequência nessa caminhada são as mais importantes. Além disso, o *PageRank* de uma página aumenta quando esta é referenciada por uma página com alto valor de *PageRank*, ou seja, o nível de importância de uma página é medido pelo número e a ‘qualidade’ de *links* que apontam para ela.

No algoritmo do *PageRank* é previsto que um usuário que visita aleatoriamente diversas páginas de um grafo acabe se cansando de navegar por entre esses vértices e acabe

digitando manualmente o endereço de uma outra página qualquer no navegador, que pode ser totalmente desconectada do grafo anterior. Assim, a probabilidade de que, a cada passo, o internauta continuar a seguir as ligações do grafo é o fator de amortecimento d . Esse valor pode variar entre 0 e 1, sendo 0.85 um valor muito utilizado (DING, 2011).

O fator de amortecimento também previne um dos problemas mais comuns do *PageRank*, o ciclo fechado (*rank sink*) no grafo de páginas, onde nenhuma das páginas do ciclo tem ligação para uma página fora do ciclo, veja Figura 11. Nessa situação o cálculo do *PageRank* fica ‘preso’ no ciclo infinito e, em cada iteração, o valor de *PageRank* é transmitido de um vértice para outro do ciclo, nunca distribuindo o valor para vértices fora do ciclo.

Figura 11 – Ciclo fechado durante o cálculo do *PageRank* das páginas A, B, C e D.

Finalmente, podemos representar o *PageRank* através da equação:

$$\text{PR}(v_i) = d \sum_{j=1}^n \frac{\text{PR}(v_j)}{L(v_j)} + \frac{1-d}{n}, \quad (1.3)$$

em que $L(v_j)$ representa o número de arestas saindo de v_j , n o número de nós da rede e d o fator de amortecimento.

Agora que compreendemos melhor os principais conceitos envolvendo a Teoria dos Grafos e Redes Complexas, partiremos para o próximo capítulo onde analisamos alguns trabalhos relacionados que utilizaram a estrutura de grafo para a classificação de dados.

Trabalhos Relacionados

Embora as abordagens léxicas e as baseadas em AM sejam as mais utilizadas no desenvolvimento de trabalhos sobre análise de sentimento por apresentarem um bom desempenho, alguns autores apostam na abordagem que utiliza a teoria de grafos para determinação de sentimento, principalmente para textos curtos, tais como os provenientes de redes sociais como o Twitter. Um facilitador para estas abordagens de grafos foi o surgimento do método *Label Propagation* – LP (ZHU; GHAMRANI, 2002), um método semi-supervisionado que possui o objetivo de reduzir a demanda de dados rotulados. Demanda essa essencial nas técnicas de aprendizado de máquina.

Speriosu e seus colaboradores (SPERIOSU et al., 2011) foram os primeiros a aplicar o LP para análise de sentimento no Twitter. Eles propuseram explorar o grafo de seguidores de um conjunto de perfis do Twitter a fim de se produzir uma classificação de sentimento eficiente. A partir desses relacionamentos eles construíram um grafo que tinha usuários, *tweets*, unigramas, bigramas, *hashtags* e *emoticons* como seus vértices. Estes vértices são conectados baseados numa relação existente entre eles, por exemplo: usuários são conectados com *tweets* criados por eles; *tweets* são ligados com palavras unigramas e bigramas contidas por eles e assim por diante. Um pequeno conjunto de vértices desse grafo, também conhecidos como ‘nós sementes’, foi rotulado e sobre eles foi aplicado o *Label Propagation*. A informação contida nos nós semente foi propagada ao longo de todo o grafo construído e como resultado os autores alegaram que obtiveram uma taxa de acerto de 87.4% sobre o conjunto de dados utilizado.

Cui e seus colegas (CUI et al., 2011) também utilizaram o método de LP baseado em *tokens* de emoção para análise de sentimento no Twitter. A primeira etapa consistiu na extração dos *tokens* de emoção a partir de um conjunto de *tweets*. Em seguida um algoritmo de propagação em grafos foi utilizado para atribuir uma polaridade para cada token extraído. Por último, um algoritmo de análise de sentimento classificou os *tokens* de emoção *emoticons*. Os resultados mostraram que o uso de *tokens* de emoção é uma abordagem bastante interessante, pois eles obtiveram bom desempenho na análise de

sentimento de mensagens escritas em qualquer linguagem natural.

Tanto as pesquisas de Speriosu et al. (2011) quanto de Cui et al. (2011) foram desenvolvidas utilizando a língua inglesa para a detecção do sentimento. Essa é uma prática bastante comum na comunidade de AS por conta da popularidade do idioma. Esse foco dos trabalhos científicos na língua inglesa levou outros pesquisadores a trabalharem em projetos sobre análise de sentimento em textos não sensíveis a língua, ou desenvolveram métodos eficazes para textos multilíngues, como veremos a seguir.

Quase na mesma época dos dois trabalhos citados anteriormente, Aisopos et al. (2012) fizeram um estudo sobre alguns dos maiores desafios em AS associados ao conteúdo de microblogs. Dentre estes desafios estavam a dificuldade na aplicação de métodos tradicionais de análise e a classificação de sentimentos, devido às algumas características inerentes a este tipo de texto, tais como, erros de ortografia, siglas, gírias, entre outros. Para solucionar esses problemas, os autores elaboraram um método que depende de duas fontes ortogonais e complementares de evidência que são: a) as características baseadas em conteúdo, capturadas através de grafos de n-gramas; e, b) as características baseadas em contexto que são adquiridas a partir da topologia da rede de relacionamentos do usuários. Ambos os métodos foram projetados para serem independentes da língua usada do texto e tolerantes a ruídos (erros de digitação, gramática, abreviação, etc) garantindo alta robustez e eficácia, segundo os autores. Os grafos foram construídos com base na sequência de n-gramas dos textos e associam todos os vizinhos de caracteres com arestas que denotam sua taxa de co-ocorrência (média) dentro de um *tweet* único, ou uma coleção de tweets. A fim de assegurar que essa abordagem possa ser aplicada em grandes quantidades de dados, teve-se que aumentar a sua eficiência em termos de tempo. Assim, os autores propuseram conjuntos alternativos de características com baixo custo computacional de extração, exploraram técnicas de redução de dimensionalidade e discretização, além de experimentar múltiplos e diferentes métodos de classificação, tendo alcançado resultados com valor de acurácia acima dos 80%.

Acreditando que a orientação semântica poderia ser transferida entre diferentes idiomas através de alinhamento de grafos, Scheible et al. (2010) construíram grafos monolíngues baseados em dois tipos de relação: a) afinidade entre adjetivos (por exemplo: ‘saudável’ e ‘saboroso’ remetem a um sentimento positivo); e b) adjetivos modificadores de substantivos (por exemplo: ‘comida saudável’). Estes pesquisadores utilizaram um léxico de sentimentos para determinar a polaridade das palavras presentes nos vértices do grafo Inglês, e também usaram um léxico bilíngue para determinar quais seriam os vértices sementes responsáveis pela ligação entre o grafo inglês e o grafo de língua estrangeira. O algoritmo *SimRank*, que determina a similaridade entre dois vértices de grafos diferentes, foi utilizado pelos autores para expandir a relação entre o idioma inglês e o idioma estrangeiro.

Gao e colegas (GAO et al., 2015) apresentaram uma abordagem semelhante à

anterior de Scheible et al. (2010) ao construírem um grafo composto de dois sub-grafos monolíngues representando dois idiomas distintos, um de origem e outro de destino, respectivamente. A ligação entre esses dois grafos é feita por um subgrafo inter-idiomático baseado na informação de alinhamento de palavras num corpus paralelo. As arestas do grafo monolíngue podem ter pesos positivos ou negativos, correspondendo às relações de sinonímia ou antonímia entre as palavras. Por fim, o algoritmo de *Label Propagation* foi utilizado para propagar rótulos de polaridade de sentimento partindo do inglês, língua de origem, para o chinês, idioma destino.

Focando mais na utilização de grafos para detecção de sentimentos e utilizando algoritmos de similaridade entre grafos, Violos et al. (2016) propuseram o método chamado Análise de Sentimento com Grafos de Palavras que foi aplicado sobre *tweets* considerando sua sequência de palavras para a representação em forma de grafo. Um conjunto já anotado de *tweets* foi dividido em três categorias: positivo, negativo e neutro. Cada uma dessas categorias foi representada por um grafo que era o resultado da união de vários *tweets* de sua respectiva classe. Em seguida, o grafo de um *tweet* não rotulado era comparado com cada um dos grafos de classes e o resultado dessa comparação é um vetor de características com valores que correspondiam ao grau de similaridade que aquele *tweet* tinham com cada classe de sentimentos. Após esta etapa, foram utilizados classificadores automáticos para fazer a predição do sentimento de cada texto baseado nos valores do seu vetor de características. Os testes foram realizados variando-se a forma de construção dos grafos e usando diferentes classificadores automáticos. O melhor experimento obteve acurácia de 75.05%.

Por último, citamos o trabalho de Cupertino e Zhao (2012), que juntamente com o trabalho de Violos et al. (2016) foram os inspiradores desta dissertação. Nesse trabalho os autores introduziram uma técnica de classificação supervisionada de múltiplos conjuntos de padrões através da centralidade de *Katz* em redes complexas. Foram criadas redes rotuladas (classificadas) através do método de agrupamento k-NN, seguido da inserção de padrões não rotulados em cada uma dessas redes, ou seja, um determinado dado foi inserido em diversas redes diferentes. Assim, novas redes foram formadas, as quais tiveram sua centralidade de *Katz* calculada. A rede que obteve o maior valor de *Katz* após a inserção do dado não rotulado foi escolhida como a classe desse dado. Os resultados da técnica foram satisfatórios para coleções de dígitos manuscritos e múltiplos objetos, superando vários métodos de classificação presentes no estado-da-arte como o *Label Propagation* e *Manifold-Based Smoothing under Constraints*.

Materiais e Método

Recordamos que no trabalho recente de Violos et al. (2016) estes pesquisadores utilizaram o conceito de similaridade entre grafos com algoritmos baseados em Semelhança de Contenção e Máximo Subgrafo Comum para detecção de polaridade de sentimentos em textos curtos. Fazendo uso de grafo de palavras para modelagem dos dados rotulados, nesta dissertação inovaremos no processo de análise de sentimento através da utilização de uma importante característica das redes, a centralidade de vértices.

Em resumo, o método proposto funciona como um sistema de eleição majoritária, em que cada elemento votante corresponde a uma medida de centralidade, entre elas, a centralidade de autovetor, *Katz* e *PageRank*. Cada palavra do documento a ser classificado tem sua centralidade calculada quando este está inserido nos grafos que correspondem às classes positiva e negativa. Por fim, a classe que obtiver o maior número de votos por partes das medidas corresponderá ao resultado da classificação.

3.1 Conjuntos de Dados

Os experimentos deste estudo foram conduzidos utilizando-se cinco conjuntos de dados em inglês, pré-classificados como positivo ou negativo, formados por *tweets* e separados em duas classes distintas. A classe neutra não foi considerada aqui, devido ao fato que alguns dos conjuntos rotulados não possuíam dados pertencentes a esta classe, portanto utilizou-se apenas *tweets* positivos e *tweets* negativos. Além disso, para os experimentos foram utilizados 80% de cada conjunto de dados para a etapa de construção das redes de palavras e 20% restantes foram destinados à etapa de testes.

A seguir, uma breve descrição de cada um dos cinco *datasets* adquiridos para esta pesquisa. Os quatro primeiros conjuntos encontram-se disponíveis no site Figure-Eight¹,

¹ <https://www.figure-eight.com/data-for-everyone/>

enquanto o último *dataset* foi obtido através de um *script* que faz o *download* de *tweets* de forma automatizada, disponibilizada pelo evento SemEval 2017 através do GitHub².

1. **US Airline 2015:** Conjunto de *tweets* coletados em fevereiro de 2015 que opinam a respeito das companhias aéreas dos Estados Unidos. Essas opiniões variam desde elogios até a reclamações sobre algum aspecto de uma companhia. Os dados foram classificados em positivos, negativos e neutros; após o pré-processamento dos dados e a desconsideração dos 3.103 *tweets* neutros, o conjunto de dados passou a ser composto por 9.171 documentos negativos e 2.362 positivos.
2. **Coachella 2015:** Um trabalho de análise de sentimento sobre o *lineup* do evento Coachella no ano de 2015. Foram coletados as postagens que falavam a respeito dos melhores shows do evento e também do evento como um todo. Nesta coleção temos 2.362 documentos marcados como positivos e 635 marcado como negativos.
3. **Deflategate 2015:** *Dataset* contendo *tweets* que opinaram a respeito do caso conhecido como *Deflategate* ocorrido no início de 2015 envolvendo duas equipes da National Football League, Patriots e Colts. Resumidamente, a equipe Patriots foi acusada de levar 11 bolas murchas para o jogo contra os Colts, fato que poderia explicar a sua vitória esmagadora por 45 a 7. Para este assunto foram adquiridos 1.236 postagens de sentimento positivo e 5.033 postagens negativas.
4. **GOP Debate 2016:** Análise de *tweets* sobre o debate do Partido Republicado dos Estados Unidos realizado no início de agosto de 2015 em Ohio. Basicamente os dados coletados estava relacionados a candidatos e a assuntos que estavam sendo debatidos. Este *dataset* possui no total 2.236 *tweets* classificados como positivo e 8.492 considerados negativos.
5. **SemEval-2017:** Adquirido a partir da Tarefa 4 da conferência SemEval 2017 (ROSENTHAL; FARRA; NAKOV, 2017). Diferentemente dos dois conjuntos de dados anteriores, este *dataset* é formado a partir de *tweets* sobre tópicos bastante variados, incluindo pessoas (por exemplo, Gadafi, Steve Jobs), produtos (por exemplo, Kindle, telefones Android) e eventos (por exemplo, terremoto no Japão, playoffs da NHL). Trata-se também de um grupo desbalanceado de textos sendo constituído por 7.059 *tweets* positivos e 3.231 *tweets* negativos.

A Tabela 1 sumariza o número de *tweets* divididos por categoria, pertencentes aos seus respectivos *datasets*.

² https://github.com/seirasto/twitter_download

Tabela 1 – Conjuntos de dados adquiridos para esta pesquisa.

Dataset	Positivo	Negativo
US Airline 2015	2.361	9.171
Coachella 2015	2.362	635
Deflategate 2015	1.236	5.033
GOP Debate 2016	2.236	8.492
SemEval 2017	7.059	3.231

3.2 Desbalanceamento do Conjunto de Dados

Ao analisarmos os dados da Tabela 1 é fácil notar uma distribuição não igualitária do número de exemplos por categoria em todos os conjuntos de dados. Esse desbalanceamento de dados é uma situação bastante frequente em *datasets* encontrados na *web*, em que poucas vezes encontramos conjuntos com uma diferença ínfima entre o número de exemplos de cada uma das classe e em raras situações essa diferença é zero.

Mas qual problema esse alto desequilíbrio dos dados pode trazer para a classificação de sentimento? O desequilíbrio dos dados é um dos grandes desafios da mineração de dados e afeta principalmente sistemas preditivos baseados em aprendizado de máquina supervisionado. Imagine que você está trabalhando em um sistema detector de fraudes em cartões de crédito. Para construir um modelo de classificação desse tipo de dado é preciso ter um conjunto de dados que servirá de treinamento para o seu classificador. Uma vez que você está lidando com fraude, que é um acontecimento raro se comparado com o número de transações diárias de uma empresa de cartão de crédito, podendo ter um poder de representatividade de até em 5% nas observações coletadas, consequentemente você terá um grande desequilíbrio na quantidade de exemplos por categoria. Ao construir um modelo preditivo a partir de uma base de dados tão desbalanceada, o sistema de predição terá uma alta tendência de classificar um dado como não-fraudulento, pois ele foi treinado com mais dados não-fraudulentos do que fraudulentos.

Além disso, a base de testes também pode estar desbalanceada e poderá passar a falsa ilusão que o sistema estará com uma acurácia bastante alta, errando apenas para os raros exemplos de fraude que poderão aparecer. Esse tipo de situação na classificação de dados é conhecida como ‘paradoxo da acurácia’ (AKOSA, 2017). Por fim, podemos ainda ter o *overfitting* na classificação que ocorre quando um sistema preditivo pode ter um bom desempenho em um conjunto de testes mas ao processar um novo conjunto de testes, ele se mostra ineficaz na predição de novos resultados, situação que também é explicada por Akosa (2017). Na situação de detecção de fraudes, podemos considerar esse novo conjunto de testes uma base contendo apenas dados fraudulentos que provavelmente irá ser classificado como não-fraudulentos em quase sua totalidade.

Tabela 2 – Conjuntos de dados balanceados que serão usados para construção de grafos de palavras e testes, e seus respectivos número de *tweets* positivos e negativos.

Dataset	Positivo	Negativo
US Airline 2015	2.361	2.361
Coachella 2015	635	635
Deflategate 2015	1.236	1.236
GOP Debate 2016	2.236	2.236
SemEval 2017	3.231	3.231

Atualmente existem algumas abordagens para lidar com situações em que há o desbalanceamento na base dados. Neste trabalho optamos por utilizar uma estratégia simples denominada *Random Under-Sampling* que consiste em reduzir de forma aleatória o número de exemplos da classe majoritária até que ela fique com a mesma quantidade de exemplos da classe menor (CHAWLA et al., 2002). Desse modo, foram removidos de forma aleatória *tweets* das classes majoritárias de cada um dos cinco conjunto de dados apresentados na secção anterior. Ver Tabela 2 com os dados já balanceados.

Apesar da grande perda de informação, este processo é muito importante para a construção de um analisador de sentimento com um nível de viés reduzido. Reforçando o motivo pelo qual termos escolhido o processo de *Random Under-Sampling*, como em nossa análise utilizamos grafos de palavras com nós únicos, ou seja, não há repetição de uma mesma palavra ao longo do grafo, se tivéssemos feito o processo contrário, *Random Over-Sampling*, o qual aumenta de forma aleatória o número de exemplos da classe minoritária até que alcance a mesma quantidade da classe majoritária; esta técnica não surtiria efeito.

Finalmente, os *datasets* no estado da Tabela 2 foram divididos em dois subgrupos, o primeiro totalizando em 80% do conjunto de dados foi destinado para a construção de redes de palavras e 20% restantes para o conjunto de testes do classificador. O resultado final da divisão pode ser visto na Tabela 3.

3.3 Pré-processamento de Textos

Uma vez que iremos trabalhar com grafos de palavras, precisamos filtrar o conteúdo dos documentos que irão servir de insumo para a construção dessas estruturas, como forma de redução de dimensionalidade. Devemos remover termos que pouco contribuem para os sentimentos nos textos e que estão propensos a estarem presentes, com frequência, em textos de ambas as classes. Como consequência disso, os grafos terão um número menor de nós e arestas o que irá contribuir para a redução do tempo computacional dos cálculos de centralidade do método de análise de sentimento.

Tabela 3 – *Datasets* divididos em conjuntos de treino e teste, e seus respectivos número de *tweets* positivos e negativos.

Dataset	Conjunto	Positivo	Negativo
US Airline 2015	Treino	1.889	1.889
	Teste	472	472
Coachella 2015	Treino	508	508
	Teste	127	127
Deflategate 2015	Treino	989	989
	Teste	247	247
GOP Debate 2016	Treino	1.789	1.789
	Teste	447	447
SemEval 2017	Treino	2.585	2.585
	Teste	646	646

O pré-processamento dos dados coletados consistiu na execução de algumas técnicas apresentadas na Secção 1.5 como: remoção de *stop words*, códigos HTML, *emoticons* e *emojis*, URLs, caracteres especiais, números, sinais de pontuação e elementos reservados do Twitter como a palavra ‘RT’, menções de usuários e *hashtags*; além da expansão de contrações da língua inglesa, normalização de palavras enfatizadas e stematização. A seguir, um exemplo de normalização de um *tweet*.

Tweet original: ‘@AmericanAir I tried already and I am on the waiting list for a call back #2hourwaitsucks’

Tweet pré-processado: ‘i tri already i wait list call back’

3.4 Grafos de Palavras: Construção

Relembramos que um grafo de palavras é uma estrutura de nós e arestas, sem *self-loops* e com nós únicos, em que cada nó representa uma palavra do documento e uma aresta conecta duas palavras. Neste trabalho o conceito de aresta está ligada ao de vizinhança entre duas palavras. A construção do grafo depende diretamente do valor de seu *word frame* (wf) que pode variar de 2 a n , sendo n o total de palavras do texto. O *word frame* remete à quantidade de palavras subsequentes que um determinado termo irá se ligar somada ao termo em questão. Para um melhor entendimento, considere uma frase como sendo $S = w_1w_2w_3 \dots w_n$, composta por uma sequência de palavras w_i que seja utilizada para construir um grafo G_2 representado pelo seguinte conjunto de arestas $E_2 = \{w_1w_2, w_2w_3, w_3w_4, \dots, w_{n-1}w_n\}$. Neste caso, G_2 é construído com *word frame* de tamanho igual a 2, $|wf| = 2$, ou seja uma palavra é ligada à sua próxima palavra consecutiva, como é mostrado na Figura 12.

Figura 12 – Grafo de palavra com *word frame* igual a 2.

Também pode-se construir um grafo G_3 considerando uma palavra conectada às suas duas palavras subsequentes. Nesta situação,

$$E_3 = \{w_1w_2, w_1w_3, w_2w_3, w_2w_4, \dots, \\ w_{n-3}w_{n-2}, w_{n-3}w_{n-1}, \\ w_{n-2}w_{n-1}, w_{n-2}w_n\}$$

para $|wf| = 3$, e assim por diante. Na Figura 13 mostramos alguns passos dessa construção, a palavra corrente e as novas arestas adicionadas ao grafo são destacadas em negrito, enquanto o *word frame* é ilustrado através de um retângulo envolvendo suas palavras.

Figura 13 – Grafo de palavra com *word frame* igual a 3.

Como pode ser visto no Passo 7, se o número de elementos subsequentes à palavra corrente for menor que o tamanho da janela, o tamanho do *word frame* será igual ao número de palavras restantes.

3.5 Método

Uma vez compreendido o processo de construção de uma rede de palavras, nesta secção apresentaremos o método proposto dividido em três etapas. Resumidamente temos:

Etapa 1 Esta etapa compreende o estado onde os dados de um conjunto de *tweets* são convertidos em grafos de palavras formando redes de sentimento, enquanto na;

Etapa 2 o grafo T de um *tweet* de polaridade desconhecida é inserido em cada uma das redes de sentimento tendo seus vértices e arestas unidas a elas; e finalmente, na;

Etapa 3 a centralidade dos vértices pertencentes a T tem suas centralidades calculadas, as quais irão impactar diretamente no resultado final do analisador de sentimento.

Etapa 1: Redes de Sentimento

Figura 14 – Transformação dos textos e união dos grafos de palavras de cada texto contido no conjunto de dados positivos.

Como mencionado na Secção 3.1, 80% de cada conjunto de dados é destinado para a criação das redes de sentimento. Primeiramente, cada um dos textos rotulados são transformados em um grafo de palavras tendo suas ligações controladas pelo valor do *word*

frame passado como parâmetro. Em seguida, é feita uma operação de união entre os grafos de palavras de mesma polaridade, ou seja, um *tweet* positivo se une à outro *tweet* positivo para então formar um novo grafo, que por sua vez será unido a um outro texto positivo e, assim por diante, até que todos os textos estejam agrupados por classes de sentimento. Esse procedimento é aplicado igualmente sobre os documentos negativos e no final desta etapa temos como resultado duas redes complexas de palavras P e N , representando os sentimentos positivo e negativo respectivamente.

A operação de união entre dois grafos mantém as arestas e vértices comuns para ambos, porém é feito uma reorganização dos vértices e de suas arestas com a finalidade de se eliminar as redundâncias, quando houver. Podemos exemplificar esse processo com uma simples situação: considere que em um grafo G_1 existe um vértice v_1 ligado a v_2 e v_3 , enquanto em G_2 o mesmo v_1 está ligado a v_2 e v_4 , então o grafo resultante da união entre G_1 e G_2 terá apenas um vértice v_1 ligado a v_2 , v_3 e v_4 . A Figura 14 ilustra o processo de transformação dos textos e união de seus grafos de palavras formando a rede P que representa o agrupamento de todos os *tweets* positivos.

Etapa 2: Inserção de Documento nas Redes de Sentimento

Figura 15 – Representação das redes de sentimento P e N antes da união temporária com o grafo de palavras T .

Um *tweet* pertencente ao conjunto de testes é também convertido em um grafo de

palavras T e, em seguida, é unido temporariamente a cada uma das redes de sentimento, assim formando duas novas redes, $P \cup T$ e $N \cup T$. Essa adição pode alterar a estrutura de P e N se um novo vértice (palavra) for adicionado ou se uma nova aresta for necessária para ligar duas palavras anteriormente não conectadas no grafo original.

Acreditamos que os nós de um *tweet* ainda não classificado harmonizam-se de forma mais adequada no grafo de sentimento para as quais as suas palavras apresentam os mais altos graus de conectividade; localizando-se em regiões distantes dos limites da rede. A Figura 16 mostra o resultado do processo de união de T em ambas as redes de sentimento. Os novos nós e arestas foram destacados com cores diferentes para uma melhor visualização da rede final.

Etapa 3: Cálculo das Centralidades

Figura 16 – Representação das redes de sentimento P e N após a adição de T em ambos.

Nesta etapa são computadas as centralidades de autovetor, *Katz* e *PageRank*, para cada um dos nós de T presentes em $P \cup T$ e $N \cup T$. Após todos os cálculos, os vértices pertencentes a T são removidos de ambas as redes retornando P e N , e assim, asseguramos que P e N estão em suas formas de origem, atuando como estruturas de referência, para o cálculo das centralidades de um novo documento não classificado.

3.6 SentiElection

Neste trabalho, apresentamos a abordagem chamada SentiElection que adota a estrutura de grafos de palavras e o valor de conectividade de seus nós para realizar a análise de sentimento de um texto. Seu funcionamento é similar a de um *ensemble* de algoritmos de aprendizado de máquina, o qual usa múltiplos classificadores supervisionados para formar um modelo preditivo de alta eficiência. Entretanto, ao invés de algoritmos de aprendizado de máquina, o SentiElection faz uso de algoritmos de centralidade de vértices em grafos.

Na maioria das implementações os *ensembles* são baseados em um sistema de voto majoritário, em que cada algoritmo de aprendizado atua como um indivíduo votante podendo haver um peso atrelado ao seu voto. Ao final da votação, a classe que receber o maior número de votos corresponderá ao resultado final da classificação. No SentiElection cada métrica de centralidade indica a provável classe que um determinado texto pertence, assim funcionando como um voto para o sistema. Consideramos cada medida de forma igualitária atribuindo o peso 1 para cada um dos votos.

A Figura 17 ilustra o funcionamento geral do classificador proposto.

Figura 17 – Representação geral do funcionamento do SentiElection.

O SentiElection foi desenvolvido considerando dois níveis para o cálculo da importância de um documento inserido nas redes de sentimento.

- 1) **SentiElection a nível de frase:** Considere Cent como uma representação genérica de uma das três medidas de centralidades utilizadas no SentiElection (autovetor, Katz e PageRank) e uma mensagem T com n palavras, cada uma representada por w_i . O valor de centralidade CentM de T é calculado pela soma dos valores de centralidade de todo w_i presente em $P \cup T$ e $N \cup T$.

Dessa forma, pode-se representar o CentM(T) para $P \cup T$ através da seguinte equação:

$$\text{CentM}_P(T) = \sum_{i=1}^n \text{Cent}(w_i), \quad (3.1)$$

em que $w_i \in (P \cup T) \cap T$. De forma análoga, pode-se calcular $\text{CentM}_N(T)$ para cada $w_i \in (N \cup T) \cap T$.

Se $\text{CentM}_P(T)$ for maior que $\text{CentM}_N(T)$, então T recebe um voto como documento de polaridade positiva, caso contrário T recebe um voto como documento negativo.

Lembramos que o cálculo do $\text{CentM}(T)$ é feito para cada uma das métricas de centralidade.

Uma vez que o SentiElection indica a classe do texto por meio de uma votação majoritária, a classe de sentimento que receber o maior número de votos vence. Matematicamente temos:

$$\text{SentiElection} = \sum V_E V_K V_{PR}, \quad (3.2)$$

em que V_E , V_K e V_{PR} correspondem aos votos das centralidades de autovetor, *Katz* e *PageRank* repectivamente, podendo assumir o valor de -1 para negativo e 1 para positivo. Essa versão do SentiElection pode ser representada pela Figura 18. Veja também o Pseudocodigo 1 para mais detalhes.

Figura 18 – Cálculo a nível de sentença da centralidade de uma frase inserida nas duas redes de sentimento.

- 2) **SentiElection a nível de palavra:** Ligeiramente diferente da versão anterior, um termo w_i do documento T tem seu valor de Cent calculado em ambas as redes de sentimento e em seguida, os valores obtidos são comparados entre si e a rede onde w_i possuir o maior valor de centralidade corresponderá a polaridade de w_i , o seu WordPol. Esse processo é feito a cada palavra de T e a cada iteração o número de palavras positivas e negativas presentes na frase é atualizado. No fim, o classificador prediz o sentimento indicando a classe com maior representatividade no texto.

$$\text{WordPol}_P(T) = \text{Cent}(w_i), \text{ para } w_i \in (P \cup T) \cap T$$

Algorithm 2 SentiElection a nível de palavra

```

totalPos  $\leftarrow$  0
totalNeg  $\leftarrow$  0
P  $\leftarrow$  createGraph(positiveDataset, frameSize)
N  $\leftarrow$  createGraph(negativeDataset, frameSize)
T  $\leftarrow$  createGraph(tweet, frameSize)
for all node  $\in$  T do
    WordPolP(T)  $\leftarrow$  Cent(P  $\cup$  T, node)
    WordPolN(T)  $\leftarrow$  Cent(N  $\cup$  T, node)
    if WordPolP(T) > WordPolN(T) then
        totalPos  $\leftarrow$  totalPos + 1
    else
        totalNeg  $\leftarrow$  totalNeg + 1
    end if
end for
if totalPos > totalNeg then
    return 1  $\triangleright$  positivo
else
    return -1  $\triangleright$  negativo
end if

```

3.7 Resumo do SentiElection

Finalmente, resumizamos todo o *pipeline* de processos da análise de sentimento com o SentiElection na Figura 20.

Figura 20 – Sequência de tarefas da análise de sentimento usando o método SentiElection.

Métodos do Estado-da-Arte

No capítulo anterior detalhamos todas as etapas que compõem o processo de análise de sentimento utilizando o SentiElection. Neste capítulo apresentamos as técnicas de AS já existentes no estado-da-arte e que foram utilizadas em paralelo ao SentiElection sobre os cinco *datasets* apresentados nesta dissertação, assim formando nossa *baseline* de resultados. Para competir com o SentiElection fizemos uso de métodos de classificação de sentimento baseados nas abordagens léxicas, aprendizado de máquina supervisionado e grafos.

4.1 Semantic Orientation Calculator (SO-CAL)

A abordagem léxica é representada neste trabalho pelo uso da técnica conhecida como Semantic Orientation Calculator ou simplesmente SO-CAL apresentada por Taboada et al. (2011). Como já mencionado, a abordagem léxica faz uso de dicionários de palavras pré-rotulados com base em sua orientação de sentimento e então utiliza-se um algoritmo para prever o sentimento de um texto de acordo com a ocorrência dessas palavras pré-classificadas. Ao longo dos anos diversas melhorias foram implementadas para esse tipo de abordagem e como grande destaque tivemos o surgimento do SO-CAL.

Taboada e seus colegas (TABOADA et al., 2011) criaram manualmente dicionários de palavras, em inglês, classificadas em neutro, positivo e negativo com base em uma escala variando de -5 a 5, sendo -5 a valência das palavras com sentimento extremamente negativos (e.g. *monstrosity*), 0 para palavras neutras (e.g. *sky*), e 5 o valor atribuído às palavras com forte sentimento positivo (e.g. *marvelous*). Os dicionários são compostos de substantivos, verbos, adjetivos e advérbios extraídos a partir de 500 textos pertencentes a diferentes base de dados. Ao todo foram gerados cinco dicionários com 2.257 adjetivos, 1.142 substantivos, 903 verbos, 745 advérbios e por último um dicionário contendo 177 intensificadores de palavras e expressões da língua inglesa (BROOKE; TOFILOSKI; TABOADA, 2009). O conjunto de termos intensificadores e expressões podem aumentar ou diminuir a força do sentimento da palavra subsequente, por exemplo, a negação de uma palavra extremamente

negativa, como *terrible*, amenizará o valor da sua polaridade final que ao invés de ser -5, poderá ter uma valência igual a -1 ($not\ terrible + 4 - 5 = -1$). De forma análoga a negação de um termo positivo diminuirá a valência final do termo.

Além disso, o SO-CAL se mostra um analisador de sentimento que vai além de uma simples soma da polaridades das palavras de uma frase. Cada termo de seu dicionário de expressões e intensificadores está associado a um fator de multiplicação. Segundo Brooke, Tofiloski e Taboada (2009) a palavra *very* no dicionário do SO-CAL tem um fator multiplicantes de 0.25 o que significa que o valor da polaridade de qualquer adjetivo precedido de *very* será aumentado em 25%.

4.2 Algoritmos de Aprendizado

Aprendizado de máquina (AM) é o mecanismo pelo qual o ser humano tenta fazer com que máquinas consigam aprender e tomar decisões (induzir) a partir de um conjunto particular de exemplos. Algoritmos de AM costumam ser bem sucedidos na tarefa de análise de sentimentos, por conta disso, para fins de comparações com o SentiElection adotamos três algoritmos dessa abordagem nos experimentos desta pesquisa; Naïve Bayes, *Support Vector Machines* (SVM) e *Random Forest*. Esses três classificadores fazem parte do grupo de algoritmos de aprendizado de máquina supervisionado, os quais podem ser descritos como uma função f que recebe uma entrada x e retorna uma predição y .

O modelo preditivo utilizado nos classificadores foi gerado a partir da extração de características dos textos utilizando-se a técnica chamada TF-IDF (*Term frequency-inverse document frequency* ou, em português, Frequência do termo-inverso da frequência do documento). Essa técnica é uma medida estatística que indica a relevância, não a frequência, de uma palavra presente em um determinado texto em relação a um todo conjunto de textos (RAMOS et al., 2003). Através da identificação dessas palavras-chaves, pode-se tentar inferir a que tópico um certo texto está ligado.

O TF-IDF, como o próprio nome sugere, é a junção de duas tarefas. A primeira se chama frequência do termo (TF) também conhecida por *bag-of-words*, um modelo de representação que recebe como entrada um conjunto de textos e tem como saída uma tabela contendo as contagens de frequência de cada palavra em cada documento, ignorando a ordem e a gramática dos termos. Em outras palavras, o modelo se preocupa apenas com o fato de palavras conhecidas ocorrerem no documento e não onde elas ocorrem de fato no texto. Um exemplo de sua aplicação pode ser visto a seguir. Considere as seguintes frases:

Tabela 4 – Tabela de frequência das palavras contabilizadas por documento.

	I	love	software	development	and	testing	don't	like	prefer	than
Texto 1	1	1	2	1	1	1	0	0	0	0
Texto 2	1	0	1	1	0	0	1	1	0	0
Texto 3	1	0	2	1	0	1	0	0	1	1

Tabela 5 – Tabela do inverso da frequência das palavras contabilizadas por documento.

	I	love	software	development	and	testing	don't	like	prefer	than
Texto 1	-0.041	0.058	-0.082	-0.041	0.058	0.0				
Texto 2	-0.058		-0.058	-0.058			0.081	0.081		
Texto 3	-0.041		-0.082	-0.041		0.0			0.058	0.058

Texto 1: *I love software development and software testing*

Texto 2: *I don't like software development*

Texto 3: *I prefer software testing than software development*

A Tabela 4 ilustra o resultado da aplicação do TF sobre as sentenças exemplos. Cada palavra teve seu número de aparecimento registrados em cada uma das frases.

Embora o TF seja um modelo satisfatório para a conversão de textos brutos, numa representação numérica ele pode não ser suficiente para uma boa análise de sentimento. Quando trabalhamos com muitos textos sobre um determinado assunto, algumas palavras se tornam bastante presentes em documentos de diferentes classes. Essas palavras geralmente acabam se tornando irrelevantes para o que se deseja buscar, sendo necessário descartá-las. É nesta etapa que entra em cena o inverso da frequência (IDF) que atribui a cada palavra de um documento um número proporcional à sua frequência no texto e inversamente proporcional ao número de documentos nos quais ela ocorre. Para que um termo seja considerado uma palavra-chave de um documento, ele não deve aparecer com frequência em outros textos. O IDF é geralmente calculado como:

$$\text{idf}(w_i) = \log \left(\frac{N}{\text{df}_i} \right), \quad (4.1)$$

enquanto o TF-IDF é o produto dessas duas frequências que resulta em:

$$\text{tf-idf}(w_i) = \text{tf}(w_i, d) \times \log \left(\frac{N}{\text{df}_i} \right), \quad (4.2)$$

para $\text{tf}(w_i, d)$ o termo da frequência de w_i em d , ou seja, o número de vezes que a palavra w_i aparece no documento d , df_i o número de documentos contendo a palavra w_i e N o total de documentos. Concluindo, para uma palavra ter alto TF-IDF (importância) em um documento, ela deve aparecer várias vezes em um documento e deve estar ausente nos outros documentos, caracterizando-se como uma palavra-chave do texto (AIZAWA, 2003).

Para cada texto é gerado um vetor de características que contém os valores TF-IDF de cada um de seus termos, como é mostrado na Tabela 5 ao considerarmos o exemplo

anterior. Os vetores são os dados de entrada para algoritmos de aprendizado estabelecerem uma função matemática (modelo preditivo) que, ao ser aplicada a uma massa de dados, é capaz de identificar padrões e inferir a classe a qual pertence cada uma das observações.

Ao analisarmos os dados da Tabela 5, podemos notar que as palavras *love* e *and* foram as palavras-chaves do Texto 1, enquanto para o Texto 2 os principais termos foram *don't* e *like* e por fim, *prefer* e *than* foram as palavras mais importantes do Texto 3. As palavras-chaves são as características extraídas das três sentenças e então podemos associar uma ou mais palavras-chave a uma classe.

Uma vez compreendido a extração de característica para geração do modelo preditivo, chegou a hora de falarmos brevemente sobre cada um dos algoritmos de AM supervisionado utilizados em nossa análise de sentimento.

Naïve Bayes

Naïve Bayes (NB) é um simples e eficiente classificador probabilístico que estima a probabilidade de um dado pertencer a determinada categoria, baseado em um conhecimento prévio. Ele é bastante aplicado na classificação de textos e na solução de problemas clássicos como a detecção de *spam* (ANDROUTSOPOULOS et al., 2000).

Esse classificador é considerado ‘ingênuo’ (*naïve*) pois assume que todos os atributos de um dado não dependem de outros dados. Por conta dessa suposta independência de atributos, as características dos dados podem ser aprendidas separadamente, o que facilita bastante a complexidade do aprendizado, principalmente quando se tem uma grande quantidade de atributos (KIM et al., 2006).

Relativamente à análise de sentimentos, a probabilidade de um documento d pertencer a uma classe de sentimento c é calculada seguindo o teorema da probabilidade condicional proposto por Bayes e representada pela equação:

$$P(c|d) = \text{ArgMax} \frac{P(d|c) P(c)}{P(d)}, \quad (4.3)$$

a qual $P(d|c)$ representa a probabilidade estimada de d pertencer a c , $P(c)$ a frequência em que c aparece no conjunto de treinamento, $P(d)$ a frequência ponderada dos termos que aparecem no documento, enquanto o ArgMax corresponde a classe com maior probabilidade de estar correta.

SVM (Support Vector Machines)

Support Vector Machines é um conjunto de métodos de aprendizagem supervisionada comumente utilizada em problemas de regressão, classificação e detecção de *outliers*. O

SVM faz uso do espaço vetorial para a representação de documentos e, em sua implementação mais simples, tem como objetivo encontrar o melhor hiperplano de separação, capaz de dividir um conjunto de dados em duas classes linearmente separáveis, com uma margem máxima. Essa margem remete à menor distância do hiperplano ao elemento mais próximo de cada uma das classes (WU et al., 2008). A margem máxima define o hiperplano ótimo, visto que há uma infinidade de hiperplanos que podem servir como classificadores. Portanto, o hiperplano ótimo se encontra o mais longe possível dos dados de ambas as classes. A Figura 21 mostra a separação simples de duas classes linearmente separáveis.

Em problemas de separação linear entre duas classes existem diversos algoritmos de aprendizado supervisionado capazes de encontrar hiperplanos separadores. Com o *perceptron*, por exemplo, é possível de se encontrar um hiperplano divisor, porém sem a obrigação que este seja o divisor ótimo (FRIAS-MARTINEZ; SANCHEZ; VELEZ, 2006). O principal diferencial do SVM é a sua capacidade de encontrar o separador com a maior margem de distância entre os objetos mais próximos do conjunto, sem considerar os demais elementos. Essa distância é crucial para separar com maior precisão elementos mais próximos, ou seja, aqueles que se tem um baixo grau de confiança quanto às suas classificações.

Para problemas não linearmente separáveis utiliza-se o SVM juntamente com funções matemáticas auxiliares, denominadas *kernels*, dedicadas a transformar dados em outra dimensão que possua uma margem de divisão clara entre as classes de dados.

Figura 21 – Versão mais simples do SVM denominado SVM Linear. A linha em negrito representa o hiperplano que divide um conjunto de dados e as linhas tracejadas correspondem à margem máxima entre os dados mais próximos de ambas as classes.

Random Forest

O *Random Forest* (RF) parte do princípio ‘dividir para conquistar’ sendo um *ensemble* de árvores de decisão criadas aleatoriamente a partir de amostras do conjunto de testes. Assim como o SentiElection, o *Random Forest* encontra a melhor solução baseado em um sistema de votação majoritária, onde cada voto é fornecido por uma árvore de decisão. Esse algoritmo é muito utilizado na literatura por ser bastante flexível e fácil de ser usado, além de produzir grandes resultados mesmo sem o *tuning*¹ de hiperparâmetros².

Conforme a estrutura do RF cresce, o algoritmo aumenta a aleatoriedade no momento da criação de novas árvores. Cada nó da árvore de decisão representa uma característica dos dados e após uma decisão ter sido tomada, o nó em questão expande-se em dois novos ramos e assim também se torna uma característica. O RF tem como diferencial o fato de não escolher a melhor característica para o próximo nó da árvore, mas sim a melhor característica dentre um subconjunto randômico de características. Isso resulta em uma larga diversidade em seu processamento aumentando as chances de obter um modelo satisfatório.

Vamos compreender melhor o funcionamento *Random Forest* fazendo uma analogia com uma situação do mundo real. Suponha que você queira adquirir um novo telefone celular e pergunta para um amigo qual modelo ele o recomendaria. Seu amigo então começa a fazer uma série de perguntas relacionadas às características dos modelos de celulares que você já possuiu anteriormente. Com base nas suas respostas ele poderá recomendar o modelo mais adequado a seus gostos. Em seguida, você ainda indeciso, decide consultar outros amigos pedindo para que eles façam perguntas a você de forma análogo a que o amigo anterior fez. Novamente baseado em suas respostas, cada um dos amigos irá recomendar um modelo. No contexto do RF, cada amigo seria uma árvore de decisão e cada pergunta feita a você seria equivalente a um nó da árvore, ou seja, uma regra que terá sua consequência diretamente impactada pela resposta dada. A recomendação final de cada amigo corresponde a um voto do sistema de votação e assim, o celular que obtiver o maior número de recomendações será sua escolha final. Ilustramos o funcionamento do *Random Forest* na Figura 22.

4.3 Análise de Sentimento com Grafos de Palavras (ASGP)

A Análise de Sentimento com Grafos de Palavras (ASGP) apresentada por Violos et al. (2016) possui etapas semelhantes ao SentiElection como a transformação de textos em

¹ Processo de escolha do conjunto de parâmetros que maximizam o poder preditivo do modelo.

² Parâmetros iniciais do classificador.

Figura 22 – Representação geral do funcionamento do *Random Forest*.

grafos de palavras e a construção das redes de sentimento. A mineração de opinião é realizada através do uso de classificadores lineares como o SVM e Naïve Bayes que, por sua vez, recebem como entrada vetores de características contendo os valores de similaridade de um grafo de um texto ainda não rotulado com as diferentes redes de sentimento, positiva, negativa e neutra.

Segundo os autores, a similaridade entre o grafo de um documento e um grafo de sentimento pode refletir o nível de correspondência entre ambos em relação ao sentimento, ou seja, quanto maior a semelhança entre grafos maior será a chance que os dois compartilhem o mesmo sentimento. Para calcular essa similaridade, Violos e seus colaboradores, Violos et al. (2016), necessitaram de técnicas que fizessem uso da palavra contida em cada nó do grafo, da orientação de suas arestas e também do número de ligações compartilhadas entre dois grafos. As medidas escolhidas por eles foram Semelhança de Contenção e algoritmos que calculam a semelhança entre grafos através do máximo subgrafo comum.

Semelhança de Contenção

Introduzida por Aiosopos em (AISOPOS et al., 2012), essa medida tem sido usada no cálculo da similaridade entre grafos que possuem vértices formados por n-gramas. Basicamente, representa a divisão entre o número de arestas em comum de um grafo

não rotulado T e um grafo de sentimento G , pelo número de arestas do menor grafo. Matematicamente temos:

$$CS(T, G) = \frac{\sum_{e \in T} \mu(e, G)}{\min(|T|, |G|)}, \quad (4.4)$$

tais que $|T|$ e $|G|$ denotam o número de ligações de T e G respectivamente, e é uma aresta do grafo de palavras, enquanto a função $\mu(e, G)$ assume o valor 1, se e somente si, e pertence ao mesmo a T e G , caso contrário $\mu(e, G)$ tem valor 0.

Algoritmos baseados em Máximo Subgrafo Comum

Como mostrado no Capítulo 1, página 36, o máximo subgrafo comum (MSC) de dois grafos, G_1 e G_2 , é um grafo H com o maior número de vértices e que ao mesmo tempo é isomorfo a G_1 e G_2 . A busca por um máximo subgrafo comum é um problema NP-Completo, mas uma vez que os vértices de um grafo de palavras são identificados pela própria palavra correspondente, esse problema passa a ter uma complexidade linear. Violos e seus colegas propuseram três variações de métricas de similaridade que utilizam o MSC entre o grafo de um documento não classificado e o grafo de sentimento. Essas variações estão descritas nas equações a seguir:

$$MCSNS = \frac{MCSN(|T|, |G|)}{\min(|T|, |G|)}, \quad (4.5)$$

tal que $MCSN(|T|, |G|)$ é o número total de vértices que estão presentes no MSC de T e G enquanto o $\min(|T|, |G|)$ representa o menor grafo, em termos de arestas, entre T e G , assim como ocorre no cálculo da Semelhança de Contenção;

$$MCSUES = \frac{MCSUE(|T|, |G|)}{\min(|T|, |G|)}, \quad (4.6)$$

em que $MCSUE(|T|, |G|)$ equivale ao total de arestas pertencentes ao MSC de T e G , desconsiderando suas direções nos grafos T e G . E finalmente,

$$MCSDES = \frac{MCSDE(|T|, |G|)}{\min(|T|, |G|)}, \quad (4.7)$$

sendo que $MCSDE(|T|, |G|)$ denota o total de ligações contidas no MSC de T e G e que possuem a mesma orientação nos grafos originais T e G .

Lembramos que os autores utilizam três grafos de sentimento (positivo, negativo e neutro) e cada documento tem sua similaridade calculada em relação a cada um desses grafos. O resultado de cada comparação é disposto em um vetor de características que logo em seguida tem seu sentimento definido por algoritmos de aprendizado de máquina supervisionado.

O uso de grafos para a extração de valores de similaridades e aprendizagem de máquina supervisionada tornam a ASGP uma abordagem híbrida. Neste trabalho, decidimos modificar a ASGP desconsiderando o uso de classificadores automáticos a fim de deixarmos o método puramente baseado em grafos para uma melhor comparação com o SentiElection. Ao invés de passarmos os valores de similaridade em um vetor de características para que, então, classificadores automáticos façam a análise, calculamos a similaridade do grafo de teste com cada uma das redes e aquela que possuir o maior grau de similaridade será a classe do *tweet*.

Avaliação dos Métodos

5.1 Métricas de Desempenho

Para avaliar o desempenho do SentiElection e das demais técnicas implementadas, usaremos as métricas acurácia, precisão, revocação e medida-F média. Para entender como essas medidas são calculadas, considere a matriz de confusão da Tabela 6, a seguir.

Tabela 6 – Matriz de confusão.

		Classe Prevista	
		Positiva	Negativa
Classe Real	Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Os termos Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) são valores utilizados para comparar os resultados da classificação feita pela metodologia seguida com resultados considerados verdadeiros. VP representa o número de *tweets* que foram considerados positivos pelo classificador e que verdadeiramente eram, enquanto FP é a quantidade de documentos incorretamente previstos como positivos. FN e VN têm significado correspondente para a classe negativa.

A acurácia mede a frequência com que o classificador analisa corretamente os textos quanto aos seus sentimentos. Em outras palavras, a acurácia reflete uma precisão geral do classificador, a qual é representada pela razão da quantidade de acertos sobre o total de documentos analisados:

$$\text{Acurácia (Acc)} = \frac{\text{Acertos}}{\text{Total}} = \frac{(\text{VP} + \text{VN})}{(\text{VP} + \text{VN} + \text{FP} + \text{FN})} \quad (5.1)$$

Enquanto a acurácia representa a taxa de acerto geral do sistema, a precisão de um classificador remete à exatidão do método, ou seja, o número de textos classificados

corretamente como ‘X’ dividido pelo número total de instâncias previstas como ‘X’. Dessa forma, teremos a precisão da classe positiva e também a precisão da classe negativa representadas pelas seguintes equações:

$$\text{Precisão}_{Pos} (\text{Prec}) = \frac{\text{Acertos}_{Pos}}{\text{Total classificado como positivo}} = \frac{VP}{(VP + FP)} \quad (5.2)$$

$$\text{Precisão}_{Neg} (\text{Prec}) = \frac{\text{Acertos}_{Neg}}{\text{Total classificado como negativo}} = \frac{VN}{(VN + FN)} \quad (5.3)$$

c Revocação ou *recall* refere-se à frequência com que o classificador prediz corretamente os exemplos de uma classe. Em outros termos, a revocação é fração do total de *tweets* corretamente classificados como ‘X’ pelo total de *tweets* que são, de fato, ‘X’.

$$\text{Revocação}_{Pos} (\text{Rev}) = \frac{\text{Acertos}_{Pos}}{\text{Total positivo}} = \frac{VP}{(VP + FN)} \quad (5.4)$$

$$\text{Revocação}_{Neg} (\text{Rev}) = \frac{\text{Acertos}_{Neg}}{\text{Total negativo}} = \frac{VN}{(VN + FP)} \quad (5.5)$$

A medida-F combina a precisão e a revocação em uma média harmônica ponderada, a qual determina com robustez a qualidade geral da classificação de uma classe. Essa métrica é calculada da seguinte forma:

$$\text{Medida-F} (\text{Med-F}) = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \quad (5.6)$$

Finalmente, a medida-F corresponde a qualidade geral do classificador considerando todas as classes do problema. Em nosso caso, a medida-F média é a média dos resultados de medida-F das classes positiva e negativa.

$$\text{Medida-F}_{Média} (\text{Med-F}_{Méd}) = \frac{\text{Medida-F}_{Pos} + \text{Medida-F}_{Neg}}{2} \quad (5.7)$$

5.2 Validação Cruzada K-Fold

A performance de um classificador é dada pelo desempenho da classificação do conjunto de teste. Devido ao fato do conjunto de teste ser menor do que o conjunto de treinamento, existe a chance que ele não seja suficientemente representativo em relação a todo o conjunto de *tweets*, o que pode ter como consequência uma avaliação ruidosa.

Para evitar esse tipo de avaliação não desejada, geralmente, utiliza-se a estratégia validação cruzada *K-Fold*. Essa técnica consiste em dividir o conjunto de dados em K subconjuntos distintos reservando $K - 1$ subconjuntos para treinamento e o restante para teste. O processo de divisão é feito K vezes resultando em K combinações diferentes de subconjuntos de treinamento e teste (SAMMUT; WEBB, 2011). Cada combinação é chamada de *Fold*, o que justifica o nome da técnica se chamar *K-Fold*. Um exemplo de divisão em *Folds* pode ser visto na Figura 23.

Por fim, todas as métricas de desempenho são calculadas para cada *Fold* e as médias dos resultados são dispostas nas tabelas do próximo capítulo.

Figura 23 – Exemplo de validação cruzada K-Fold utilizando 5 *Folds*.

Resultados e Discussão

Neste capítulo apresentamos uma breve descrição da implementação de cada um dos diferentes métodos aplicados nesta pesquisa, comparando-os e discutindo seus resultados na seção 6.5.

6.1 SO-CAL

Disponível no site¹ oficial do projeto o SO-CAL é um analisador de sentimento de abordagem léxica, implementado em *Python* e de complexidade computacional $O(n)$. Esse método possui como dependência o *Stanford CoreNLP*², que é uma ferramenta que provê diversas funcionalidades para se trabalhar com a linguagem humana, incluindo os dicionários de léxicos utilizados pelo SO-CAL.

O SO-CAL recebe um texto como entrada, consulta os dicionários localizados no *Stanford CoreNLP* e assim obtém o polaridade de cada um dos termos do texto. Por fim, ele realiza a computação do sentimento e retorna a classe a qual o texto pertence. Por ser uma técnica de abordagem léxica, o SO-CAL não utiliza o conjunto de treinamento separado em cada *dataset*, avaliando apenas o conjunto de teste em conjunto com a técnica de validação cruzada *K-Fold*.

Segundo a Tabela 7 podemos notar a robutez da técnica SO-CAL, a qual teve seu início de desenvolvimento em 2004 e recebendo melhorias a cada ano. Mesmo não tendo sido treinada especificamente com dados de mesmo domínio que o conjunto de teste, essa técnica mostrou possuir recursos tão generalizados e ajustados a ponto de alcançar um ótimo desempenho levando em consideração todos os *datasets*, tendo acurácia mínima de 70,56% e medida-F média a partir de 69,21% os quais já são bons resultados na área de mineração de opinião, principalmente ao se analisar textos ruidosos como *tweets*.

¹ <https://github.com/sfu-discourse-lab/SO-CAL>

² <https://stanfordnlp.github.io/CoreNLP>

Tabela 7 – Resultados do classificador SO-CAL em porcentagem (%)

SO-CAL								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	83.4	83.17	93.65	71.67	81.2	77.05	95.14	85.15
Coachella 2015	84.3	83.25	91.67	69.47	79.04	80.79	95.31	87.46
GOP Debate 2016	80.47	80.24	88.89	69.64	78.1	75.05	91.29	82.38
Deflategate 2015	70.56	69.21	85.42	49.6	62.76	64.49	91.53	75.67
SemEval 2017	71.25	69.66	89.17	48.38	62.73	64.58	94.13	76.6

6.2 Aprendizado de Máquina Supervisionado

Utilizamos as implementações dos algoritmos de AM supervisionado disponíveis no pacote *scikit-learn*³, uma biblioteca popular em estudos envolvendo AM e Python. A fim de chegarmos aos melhores resultados, combinamos o uso do TF-IDF aplicado a unigramas juntamente com o *fine tuning* de hiperparâmetros dos classificadores através do módulo *GridSearchCV*⁴.

Tabela 8 – Resultados do classificador Naïve Bayes em porcentagem (%)

Naïve Bayes								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	82.66	82.62	86.18	77.8	81.78	79.99	87.73	83.47
Coachella 2015	67.19	66.9	64.47	76.56	70.0	71.15	57.81	63.79
GOP Debate 2016	71.37	71.27	69.06	77.42	73.0	74.31	65.32	69.53
Deflategate 2015	66.52	65.91	72.56	53.12	61.34	63.03	79.91	70.47
SemEval 2017	72.41	70.72	65.14	96.45	77.76	93.15	48.38	63.68

Tabela 9 – Resultados do classificador SVM em porcentagem (%)

Support Vector Machine								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	83.4	83.38	86.07	79.7	82.77	81.1	87.1	84.0
Coachella 2015	65.62	65.37	63.33	74.22	68.35	68.87	57.03	62.39
GOP Debate 2016	68.3	67.92	73.43	57.37	64.41	65.02	79.24	71.43
Deflategate 2015	69.76	69.95	72.48	63.71	67.81	67.63	75.81	71.48
SemEval 2017	72.41	70.74	66.17	96.29	77.73	92.9	48.53	63.76

Dentre os três classificadores o Naïve Bayes foi o que obteve maior êxito geral na classificação dos *datasets*, especialmente se considerarmos aqueles com maior número de exemplos (US Airline 2015, GOP Debate 2016 e SemEval 2017) tendo desempenho de

³ <https://scikit-learn.org/stable/index.html>

⁴ https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Tabela 10 – Resultados do classificador *Random Forest* em porcentagem (%)

Random Forest								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	81.71	81.66	85.55	76.32	80.67	78.63	87.1	82.65
Coachella 2015	66.8	66.77	65.93	69.53	67.68	67.77	64.06	65.86
GOP Debate 2016	67.75	67.3	73.18	56.03	63.46	64.38	79.46	71.13
Deflategate 2015	70.56	70.46	73.39	64.52	68.67	68.35	76.61	72.24
SemEval 2017	67.54	64.98	61.38	94.59	74.45	88.22	40.49	55.51

acurácia e media-F média superiores a 70%, chegando até mesmo a superarem os 80% em ambas as medidas no conjunto US Airline 2015. Além dos melhores resultados, o Naïve Bayes tem como vantagem a sua complexidade computacional; enquanto classificadores como o *Random Forest* e SVM com complexidades $O(n \log(n))$ e $O(n^3)$ respectivamente, o Naïve Bayes possui a complexidade de $O(n)$ o que o torna mais ágil no processo de análise de sentimento.

6.3 Análise de Sentimento com Grafo de Palavras

Para a transformação de sentenças pré-processadas em grafos de palavras, implementamos um código *Python* que recebe uma sentença e um *word frame* como parâmetros e retorna um grafo de palavras direcionado e não-valorado com um conjunto de arestas baseado na estrutura do texto e no tamanho do *word frame*. A criação e manipulação dos grafos de palavras foram feitas com auxílio do pacote *NetworkX*⁵ em sua versão 2.1.

Os experimentos foram conduzidos em cada um dos diferentes algoritmos da ASGP variando-se o tamanho do *word frame* em valores de 2 a 4. É válido ressaltar que para o MCSUES as redes tiveram suas direções desconsideradas, uma vez que essa variação do *Maximum Common Subgraph* não leva em conta a direção das ligações nos grafos.

Tabela 11 – Resultados do classificador CS (*word frame* = 2) em porcentagem (%)

CS - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	70.93	70.51	66.89	82.88	74.03	77.5	58.99	66.99
Coachella 2015	62.11	58.34	57.56	92.19	70.87	80.39	32.03	45.81
GOP Debate 2016	62.28	61.25	59.26	78.57	67.56	68.21	45.98	54.93
Deflategate 2015	54.44	50.05	52.88	81.45	64.13	59.65	27.42	37.57
SemEval 2017	61.44	56.32	56.79	95.67	71.27	86.27	27.2	41.36

⁵ <https://networkx.github.io/documentation/stable/index.html>

Tabela 12 – Resultados do classificador CS (*word frame* = 3) em porcentagem (%)

CS - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	74.52	74.47	72.39	79.28	75.68	77.1	69.77	73.25
Coachella 2015	62.89	60.16	58.46	89.06	70.59	77.05	36.72	49.74
GOP Debate 2016	66.18	65.81	63.4	76.56	69.36	70.42	55.8	62.27
Deflategate 2015	54.04	51.93	53.24	79.44	63.75	59.52	30.24	40.11
SemEval 2017	64.91	61.13	59.18	96.14	73.26	89.71	33.69	48.99

Tabela 13 – Resultados do classificador CS (*word frame* = 4) em porcentagem (%)

CS - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	75.05	75.02	73.28	78.86	75.97	77.12	71.25	74.07
Coachella 2015	63.67	61.49	59.26	87.5	70.66	76.12	39.84	52.31
GOP Debate 2016	67.19	66.99	64.86	75.0	69.57	70.37	59.38	64.41
Deflategate 2015	55.85	52.75	53.87	81.45	64.85	61.98	30.24	40.65
SemEval 2017	65.77	62.63	59.98	94.74	73.46	87.5	36.79	51.8

Tabela 14 – Resultados do classificador MCSNS (*word frame* = 2) em porcentagem (%)

MCSNS - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	65.82	64.35	61.3	86.05	71.59	76.51	45.55	57.1
Coachella 2015	62.11	57.66	57.35	94.53	71.39	84.44	29.69	43.93
GOP Debate 2016	61.83	60.03	58.31	83.04	68.51	70.54	40.62	51.56
Deflategate 2015	54.44	50.49	52.84	82.66	64.47	60.19	26.21	36.52
SemEval 2017	57.34	49.85	54.14	95.98	69.23	82.31	18.7	30.48

Tabela 15 – Resultados do classificador MCSNS (*word frame* = 3) em porcentagem (%)

MCSNS - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	69.34	69.03	66.08	79.49	72.17	74.27	59.2	65.88
Coachella 2015	61.72	58.63	57.58	89.06	69.94	75.86	34.38	47.31
GOP Debate 2016	65.4	64.51	61.69	81.25	70.13	72.55	49.55	58.89
Deflategate 2015	54.69	51.26	53.49	80.24	64.19	58.47	28.51	38.33
SemEval 2017	60.97	56.4	56.66	93.35	70.52	81.14	28.59	42.29

Como destaque nos experimentos tivemos o algoritmo *Containment Similarity* com *word frame* igual a 4, superando as variações do *Maximum Common Subgraph* em quatro dos cinco conjuntos de dados, perdendo apenas para o MCSUES com *wf* igual a 2 na avaliação dos *tweets* do conjunto Deflategate 2015. Entretanto mesmo sendo o melhor

Tabela 16 – Resultados do classificador MCSNS (*word frame* = 4) em porcentagem (%)

MCSNS - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	71.25	70.98	67.85	80.76	73.75	76.24	61.73	68.22
Coachella 2015	44.38	42.01	57.5	50.44	53.74	27.56	33.59	30.28
GOP Debate 2016	65.4	64.87	62.37	77.68	69.18	70.41	53.12	60.56
Deflategate 2015	55.44	52.21	53.58	81.45	64.64	61.34	29.44	39.78
SemEval 2017	63.14	59.35	58.16	93.66	71.76	83.73	32.61	46.94

Tabela 17 – Resultados do classificador MCSUES (*word frame* = 2) em porcentagem (%)

MCSUES - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	66.6	65.45	62.17	84.78	71.74	76.08	48.41	59.17
Coachella 2015	61.72	56.85	57.01	95.31	71.35	85.71	28.12	42.35
GOP Debate 2016	61.05	59.47	57.92	80.8	67.47	68.27	41.29	51.46
Deflategate 2015	61.46	54.73	61.28	87.8	72.18	62.26	26.61	37.29
SemEval 2017	58.19	50.96	54.63	96.6	69.79	85.33	19.78	32.12

Tabela 18 – Resultados do classificador MCSUES (*word frame* = 3) em porcentagem (%)

MCSUES - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	69.88	69.72	67.53	76.96	71.94	73.02	62.77	67.51
Coachella 2015	63.67	60.46	58.71	92.19	71.73	81.82	35.16	49.18
GOP Debate 2016	64.4	63.76	61.38	77.68	68.57	69.6	51.12	58.94
Deflategate 2015	53.02	49.38	51.97	79.84	62.96	56.52	26.21	35.81
SemEval 2017	61.36	56.91	56.91	93.51	70.76	81.82	29.21	43.05

Tabela 19 – Resultados do classificador MCSUES (*word frame* = 4) em porcentagem (%)

MCSUES - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	72.41	72.25	69.49	79.92	74.34	76.37	64.9	70.17
Coachella 2015	58.59	54.59	55.39	88.28	68.07	71.15	28.91	41.11
GOP Debate 2016	65.29	65.02	63.0	74.11	68.1	68.56	56.47	61.93
Deflategate 2015	56.45	53.24	54.23	82.66	65.5	63.56	30.24	40.98
SemEval 2017	57.34	49.85	54.14	95.98	69.23	82.31	18.7	30.48

classificador da abordagem, seu desempenho foi abaixo de 70% em praticamente todos os *datasets*, com exceção do *dataset* US Airline 2015. Como pontos positivos, podemos destacar a sua implementação simples e complexidade computacional de $O(n)$.

Tabela 20 – Resultados do classificador MCSDES (*word frame* = 2) em porcentagem (%)

MCSDES - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	65.96	61.39	61.39	86.05	71.65	76.68	45.88	57.41
Coachella 2015	62.5	57.98	57.55	95.31	71.76	86.36	29.69	44.19
GOP Debate 2016	61.72	59.94	58.24	82.81	68.39	70.27	40.62	51.49
Deflategate 2015	54.44	50.49	52.84	82.66	64.47	60.19	26.21	36.52
SemEval 2017	57.19	49.6	82.07	18.39	30.05	54.05	95.98	69.15

Tabela 21 – Resultados do classificador MCSDES (*word frame* = 3) em porcentagem (%)

MCSDES - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	70.82	70.72	68.62	76.74	72.46	73.62	64.9	68.99
Coachella 2015	62.5	59.7	58.08	89.84	70.55	77.59	35.16	48.39
GOP Debate 2016	65.62	64.91	62.15	79.91	69.92	71.88	51.34	59.9
Deflategate 2015	54.44	51.29	52.94	79.84	63.67	59.02	29.03	38.92
SemEval 2017	61.51	57.36	57.09	92.74	70.67	80.66	30.29	44.04

Tabela 22 – Resultados do classificador MCSDES (*word frame* = 4) em porcentagem (%)

MCSDES - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	73.57	73.53	71.91	77.38	74.54	75.51	69.77	72.53
Coachella 2015	64.84	62.5	59.9	89.84	71.88	79.69	39.84	53.12
GOP Debate 2016	66.18	65.81	63.4	76.56	69.36	70.42	55.8	62.27
Deflategate 2015	55.85	53.05	53.93	80.24	64.51	61.42	31.45	41.6
SemEval 2017	62.91	59.28	58.08	92.74	71.43	81.99	33.08	47.14

6.4 SentiElection

Relembramos que SentiElection é uma solução baseada em redes complexas e utiliza medidas de centralidade de vértices para inferir o sentimento de um texto. Para sua implementação utilizamos os algoritmos de centralidade de autovetor e *Katz* contidos no pacote *NetworkX*, enquanto o *PageRank* foi aproveitado da biblioteca *python-igraph*⁶ versão 0.7 pelo fato de sua implementação estar melhor otimizada nesse pacote do que a que está presente na versão 2.1 do *NetworkX*.

Assim como na abordagem ASGP os experimentos com o SentiElection em suas duas versões foram executados com alternância no valor do *word frame* das redes, indo de 2 a 4.

⁶ <https://igraph.org/python/>

Tabela 23 – Resultados do classificador SentiElection a nível de sentença (*word frame* = 2) em percentagem (%)

SentiElection (Sentença) - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	79.49	79.46	82.22	75.26	78.59	77.19	83.72	80.32
Coachella 2015	60.16	59.44	58.02	73.44	64.83	63.83	46.88	54.05
GOP Debate 2016	64.51	63.84	61.4	78.12	68.76	69.94	50.89	58.91
Deflategate 2015	64.11	63.68	68.04	53.23	59.73	61.59	75.0	67.64
SemEval 2017	66.85	64.25	60.94	93.82	73.89	86.58	39.88	54.6

Tabela 24 – Resultados do classificador SentiElection a nível de sentença (*word frame* = 3) em percentagem (%)

SentiElection (Sentença) - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	79.49	79.48	78.07	82.03	80.0	81.07	76.96	78.96
Coachella 2015	63.67	62.17	59.78	83.59	69.71	72.73	43.75	54.63
GOP Debate 2016	67.86	67.31	64.18	80.8	71.54	74.1	54.91	63.08
Deflategate 2015	69.76	69.73	70.94	66.94	68.88	68.7	72.58	70.59
SemEval 2017	71.41	69.96	64.88	93.35	76.55	88.15	49.46	63.37

Tabela 25 – Resultados do classificador SentiElection a nível de sentença (*word frame* = 4) em percentagem (%)

SentiElection (Sentença) - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	82.24	82.23	81.06	84.14	82.57	83.52	80.34	81.9
Coachella 2015	65.23	63.55	60.66	86.72	71.38	76.71	43.75	55.72
GOP Debate 2016	70.09	69.38	65.41	85.27	74.03	78.85	54.91	64.74
Deflategate 2015	71.21	70.62	66.55	85.27	74.76	79.5	57.14	66.49
SemEval 2017	74.46	73.53	67.41	93.97	78.5	90.3	55.25	68.56

Tabela 26 – Resultados do classificador SentiElection a nível de palavra (*word frame* = 2) em percentagem (%)

SentiElection (Palavra) - WF = 2								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	74.0	74.0	74.0	74.0	74.0	74.0	74.0	74.0
Coachella 2015	61.72	59.49	57.98	85.16	68.99	72.06	38.28	50.0
GOP Debate 2016	59.93	58.97	57.61	75.22	65.25	64.31	44.64	52.7
Deflategate 2015	57.26	55.45	62.16	37.1	46.46	55.17	77.42	64.43
SemEval 2017	61.59	56.17	56.81	96.75	71.58	89.06	26.43	40.76

Tabela 27 – Resultados do classificador SentiElection a nível de palavra (*word frame* = 3) em porcentagem (%)

SentiElection (Palavra) - WF = 3								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	74.84	74.75	72.13	80.97	76.29	78.31	68.71	73.2
Coachella 2015	58.59	54.82	55.45	87.5	67.88	70.37	29.69	41.76
GOP Debate 2016	61.16	58.81	57.55	85.04	68.65	71.37	37.28	48.97
Deflategate 2015	61.9	61.81	63.11	57.26	60.04	60.89	66.53	63.58
SemEval 2017	67.62	65.52	61.8	92.27	74.02	84.76	42.97	57.03

Tabela 28 – Resultados do classificador SentiElection a nível de palavra (*word frame* = 4) em porcentagem (%)

SentiElection (Palavra) - WF = 4								
Dataset	Acc	Med-F _{Méd}	Classe Positiva			Classe Negativa		
			Prec	Rev	Med-F	Prec	Rev	Med-F
US Airline 2015	75.26	75.19	72.76	80.76	76.55	78.38	69.77	73.83
Coachella 2015	60.94	58.97	57.61	82.81	67.95	69.44	39.06	50.0
GOP Debate 2016	61.16	58.81	57.55	85.04	68.65	71.37	37.28	48.97
Deflategate 2015	63.1	63.01	61.9	68.15	64.88	64.57	58.06	61.15
SemEval 2017	68.24	65.93	61.99	94.28	74.8	88.06	42.19	57.05

Observando os resultados o SentiElection apresentou melhor desempenho em sua versão a nível de sentença com *word frame* igual a 4, alcançando resultados satisfatórios para um método recente e ainda sem grandes refinamentos. Podemos notar que conforme a número de arestas aumenta sua acurácia e medida-F também aumentam, o que nos leva a crer que a estrutura de um grafo é capaz de refletir a estrutura sintática de um texto.

Além disso, é claramente vista uma certa distância entre as performances das duas versões do SentiElection. Essa diferença de desempenho se deve ao fato de que o SentiElection a nível de sentença é capaz de preservar a intensidade positiva ou negativa de uma palavra em um *tweet* como um todo, ou seja, se em um texto tivermos uma palavra com alta centralidade positiva ou negativa, certamente ela terá um grande peso ao calcularmos a importância do texto para cada uma das redes de sentimento. Enquanto isso no SentiElection a nível de palavra, o sentimento é definido baseado na quantidade de termos positivos e negativos no texto, assim cada palavra terá o mesmo peso na predição de sentimento do texto. Em outras palavras, no SentiElection a nível de palavra um termo fortemente positivo ou negativo será contabilizado da mesma forma de uma palavra com baixa intensidade de sentimento, sem um peso atrelado as palavras ambas terão o mesmo grau de representatividade no documento.

O ponto negativo para o SentiElection fica por conta do seu custo computacional, visto que se trata de um *ensemble* com três medidas globais de centralidade. Enquanto

as centralidades de autovetor e *Katz* tem complexidade $O(n^3)$ cada uma, estima-se que o *PageRank* possua o *big O* de $O(n + m)$, sendo n o número de vértices e m o número de arestas. A soma dessas complexidades faz com que o sistema necessite de um bom tempo para processar os exemplos do conjunto de testes, podendo levar horas caso as redes de sentimento tenham uma alta magnitude.

6.5 Comparação entre Métodos

Nesta secção fazemos a comparação do desempenho entre os métodos de análise de sentimento sobre os conjuntos de dados coletados. Para as abordagens ASGP e SentiElection selecionamos suas configurações que tiveram os maiores valores de acurácia e medida-F média, sendo assim, a comparação final ficou:

- SO-CAL
- Naïve Bayes
- SVM
- Random Forest
- ASGP (*Containment Similarity* - $wf = 4$)
- SentiElection (Nível de sentença - $wf = 4$)

No primeiro grupo de teste, US Airline 2015, tivemos uma disputa acirrada entre o SO-CAL e o SVM, com a acurácia de ambos empatadas, o segundo acabou sendo levemente superior em relação ao valor da medida-F média. Observamos que para este conjunto de exemplos todos os classificadores tiveram médias de acurácia e medida-F acima dos 75%, o que mostra que esse conjunto possui *tweets* bem definidos quanto ao sentimento e facilmente classificáveis. Destacamos também o fato do SentiElection ter sido bastante competitivo com as técnicas presentes no estado-da-arte, tendo uma diferença em torno de 1% em relação ao melhor classificador.

Ao contrário do US Airline 2015, o Coachella 2015 foi o *dataset* com menores valores de desempenho dos classificadores, com exceção do SO-CAL que surpreendentemente conseguiu superar os 83% nas métricas de qualidade. Enquanto isso, as outras cinco técnicas não foram capazes de alcançar sequer 67%.

No GOP Debate 2016 mais uma vez o SO-CAL mostrou-se bem superior em relação as demais estratégias. Com uma margem por volta de 10% de distância, vemos o SentiElection com performance muito próxima ao Naïve Bayes o qual ficou em segundo lugar em desempenho na classificação dos dados.

Figura 24 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados US Airline 2015.

Figura 25 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados Coachella 2015.

No penúltimo conjunto de exemplos, Deflagate 2015, o SentiElection conseguiu ter a melhor performance entre as técnicas, tendo acurácia e medida-F média rondando os 70% de acerto. Random Forest e SO-CAL também tiveram um bom aproveitamento com números bastante parecidos ao classificador vencedor. Destoando dos demais, o método ASGP teve um baixo desempenho tendo uma diferença superior a 10% em relação aos

Figura 26 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados GOP Debate 2016.

Figura 27 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados Deflategate 2015.

outros classificadores.

Por último comparamos as estratégias na avaliação dos dados do SemEval 2017. Mais uma vez o SentiElection conseguiu ser a técnica de classificação com melhores valores de acerto, sendo cerca de 10% superior ao seu concorrente direto, ASGP.

Em um contexto geral, podemos afirmar que o SO-CAL foi o classificador com

Figura 28 – Valores de acurácia e medida-F média dos classificadores sobre o conjunto de dados SemEval 2017.

melhor performance dentre os seis analisados. Levamos em conta o fato de que no segundo e terceiro *datasets* ele obteve um aproveitamento bastante superior aos demais e em nos conjuntos de teste restantes, mesmo não sendo a melhor estratégia, o SO-CAL sempre esteve com valores de acurácia e medida-F média bem próximas ao primeiro colocado. Essa técnica vem sendo aprimorada ao longo de mais de 10 anos, por isso não é uma surpresa que ela consiga ser superior aos outros métodos principalmente os mais recentes como ASGP e SentiElection.

Os algoritmos de aprendizado supervisionado tiveram como característica um desempenho estável e regular durante os experimentos, sempre atingindo valores de performance entre 65% e 73%, exceto no primeiro *dataset* onde os algoritmos obtiveram valores de acerto superiores a 80%. O fato mais interessante é que o uso do TF-IDF, para as quais as palavras com menor frequência em documentos diferentes são determinantes para a classificação, funcionou quase tão bem quanto o uso de medidas globais de centralidade em grafos, na qual o sentimento é determinado principalmente por palavras com alta centralidade, ou seja, palavras bastante presentes em diferentes textos.

Delimitando o escopo de comparação para a abordagem de grafos, vimos que o SentiElection conseguiu ser razoavelmente superior ao ASGP, que por sua vez foi o classificador com os menores índices de performance em todos conjuntos de dados. Podemos afirmar então, que o uso da centralidade foi uma estratégia bem mais interessante do que a busca pelo máximo subgrafo comum ou pelo número máximo de arestas comuns entre dois grafos.

Conclusão

Em nosso trabalho tivemos como objetivo apresentar e executar um novo método, baseado em grafos, para a análise de sentimento em *tweets* pré-rotulados. Além disso, na introdução desta dissertação levantamos três questionamentos que agora, após a execução dos experimentos e análise dos resultados, somos capazes de responder.

1. É possível e viável realizar uma classificação de texto baseado na centralidade de seus termos?

Resposta: Sim, felizmente foi possível realizar a análise de sentimento baseado na centralidade global de vértices em grafos de palavras, seguindo a metodologia proposta. Ela também é viável pois, assim como o TF-IDF, o valor de centralidade das palavras de um texto mostrou ser uma característica eficaz para avaliação de sentimento, sendo capaz de preservar a ordem da disposição dos termos e a semântica do documento.

2. O quão eficiente é a classificação feita utilizando o método proposto em comparação à outras abordagens existentes como a léxica, aprendizado de máquina e ASGP?

Resposta: Com resultados bastante competitivos, em dois dos cinco experimentos o SentiElection a nível de sentença foi capaz de superar tradicionais soluções para a tarefa de classificação de sentimento.

Embora o SO-CAL tenha sido o método com melhor desempenho geral nos experimentos, demonstramos que o SentiElection pode ser comparável com algoritmos de AM supervisionados, além de ter sido sempre superior à outra abordagem de grafos, ASGP, o que nos fez concluir que, para a mineração de opinião, a importância dos vértices de um grafo de palavras é uma característica mais adequada a ser usada do que a similaridade entre grafos de palavras.

3. Quais as limitações deste projeto e seus cenários de aplicação?

Resposta: Embora tendo um desempenho otimista, o SentiElection enfrenta um problema quanto a sua complexidade computacional. Como mencionado na Secção 6.4, o método proposto possui como *big O* a soma de três complexidades referentes às medidas de centralidade que compõem o *ensemble* de classificação, medidas as quais demandam um alto gasto computacional no processamento das centralidades: $O(n^3)$ (autovetor), $O(n^3)$ (*Katz*) e $O(n + m)$ (*PageRank*). Em termos custo computacional, o SentiElection fica em último lugar dentre as técnicas utilizadas aqui.

Infelizmente isso faz com que o método necessite de uma máquina com alto poder de processamento para ser capaz de fazer a análise de sentimento em conjuntos de dados de alta magnitude, ou que seja necessário realizar uma redução na quantidade de dados para compor as redes de sentimento, visto que ao estarmos trabalhando com centralidades globais em grafos, quanto maior o número de nós e arestas dessas redes, maior será o tempo para o cálculo da importância de cada termo. Esse contratempo nos estimula a acreditar que, no estado atual, o SentiElection é uma ferramenta viável para cenários em que o processamento seja aplicado em textos curtos como *tweets* ou para línguas nos quais léxicos de sentimento não estejam disponíveis.

É válido mencionar também que na AS, a qualidade dos resultados depende do *corpus* a ser analisado. *Tweets* são considerados conjuntos de dados textuais sobrecarregados de informalidades, ou seja, cheios de termos sem função sintática que prejudicam a precisão do classificador, mesmo recorrendo ao uso de mecanismos de PLN como pré-processamento de dados. Outro fator que pode contribuir para um viés de classificação é o desequilíbrio no número de *tweets* positivos e negativos entre os conjuntos de dados, o qual tratamos aplicando a técnica *Random Under-Sampling*. Para situações em que a perda de dados para formar os conjuntos de treinamento e teste não seja desejável, é necessário recorrer a outras técnicas que lidam com *datasets* desbalanceados e também a outras métricas de desempenho para medir a qualidade dos classificadores. Para isso recomendamos a leitura dos trabalhos (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017) e (AKOSA, 2017).

7.1 Trabalhos Futuros

Por ser a primeira versão do método, o SentiElection acabou nos surpreendo positivamente ao apresentar resultados que nos encorajaram e seguir apostando nessa técnica e em ir em busca de melhorias. A primeira delas é tentar amenizar o alto gasto computacional demandado pela técnica através da combinação das medidas centralidade com características extraídas do texto, como a frequência de termos (unigramas, bigramas e trigramas) ou de arestas, para que a construção das redes de

sentimento seja baseada em ligações ou palavras mais relevantes para a classe. Além disso, podemos recorrer ao processo de representação de cada palavra por meio de um número inteiro e assim possibilitar um ganho de desempenho nas buscas por vértices nos grafos.

Também pretendemos aplicar a técnica de aprendizado contínuo que consiste em sempre estar sempre renovando o conhecimento dos grafos de referência, enquanto novos vértices e aresta são adicionadas, os pares mais antigos são removidos da rede. Dessa forma, o modelo poderá ser capaz de se adaptar conforme o contexto do tópico analisado se modifica.

Por fim, propomos a execução de experimentos utilizando outras medidas de centralidade baseada em vizinhanças, bem como a busca por outras características da rede que podem influenciar diretamente na importância de um nó para o grafo. Também será de grande importância para a validação do método a sua comparação com outras abordagens emergentes na área de análise de sentimento como o *Deep Learning*.

Referências

- AGGARWAL, C. C. An introduction to social network data analytics. *Social Network Data Analytics*, Springer, p. 1–15, 2011.
- AISOPOS, F. et al. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In: ACM. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. [S.l.], 2012. p. 187–196.
- AIZAWA, A. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, Elsevier, v. 39, n. 1, p. 45–65, 2003.
- AKOSA, J. Predictive accuracy: A misleading performance measure for highly imbalanced data. In: *Proceedings of the SAS Global Forum*. [S.l.: s.n.], 2017.
- ALMANSA, L. F. et al. Information learned from monitoring microblogs during the 2014 seasonal flu vaccination in brazil. In: IEEE. *2014 IEEE 10th International Conference on e-Science*. [S.l.], 2014. v. 2, p. 65–66.
- ALVARES, R. V. *Investigação do processo de Stemming na língua portuguesa*. Tese (Doutorado) — Universidade Federal Fluminense, 2005.
- ANDROUTSOPOULOS, I. et al. An evaluation of naive bayesian anti-spam filtering. *arXiv preprint cs/0006013*, 2000.
- ASLAM, S. Merriam-Webster online dictionary. 2018. Disponível em: <<https://www.omniconereagency.com/twitter-statistics/>>.
- BAEZA, Y.; RIBEIRO-NETO, B. *Modern information retrieval- the concepts and technology behind search*. [S.l.]: Pearson, 2011.
- BONACICH, P. Power and centrality: A family of measures. *American journal of sociology*, University of Chicago Press, v. 92, n. 5, p. 1170–1182, 1987.
- BONACICH, P. Some unique properties of eigenvector centrality. *Social Networks*, Elsevier, v. 29, n. 4, p. 555–564, 2007.
- BONACICH, P.; LLOYD, P. Eigenvector-like measures of centrality for asymmetric relations. *Social networks*, Elsevier, v. 23, n. 3, p. 191–201, 2001.
- BROOKE, J.; TOFILOSKI, M.; TABOADA, M. Cross-linguistic sentiment analysis: From english to spanish. In: *Proceedings of the international conference RANLP-2009*. [S.l.: s.n.], 2009. p. 50–54.
- CAMBRIDGE. 2017. <<http://dictionary.cambridge.org/grammar/british-grammar/writing/contractions>>. Acesso em: 18 de junho de 2017.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CHEW, C.; EYSENBAACH, G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, Public Library of Science, v. 5, n. 11, p. e14118, 2010.

CUI, A. et al. Emotion tokens: Bridging the gap among multilingual Twitter sentiment analysis. In: SPRINGER. *Asia Information Retrieval Symposium*. [S.l.], 2011. p. 238–249.

CUPERTINO, T. H.; ZHAO, L. Using katz centrality to classify multiple pattern transformations. In: IEEE. *Brazilian Symposium on Neural Networks*. [S.l.], 2012. p. 232–237.

DENNY, M. J.; SPIRLING, A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Unpublished manuscript, Dep. Polit. Sci, Stanford Univ and Inst. Quant. Soc. Sci., Harvard Univ.* <https://ssrn.com/abstract>, v. 2849145, 2017.

DING, Y. Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 62, n. 2, p. 236–245, 2011.

ESTRADA, E. Graph and network theory in physics. *arXiv preprint arXiv:1302.4378*, 2013.

FAHAD, A. et al. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, IEEE, v. 2, n. 3, p. 267–279, 2014.

FANG, X.; ZHAN, J. Sentiment analysis using product review data. *Journal of Big Data*, Nature Publishing Group, v. 2, n. 1, p. 5, 2015.

FERREIRA, J. Z. et al. *Análise de sentimento em documentos financeiros com múltiplas entidades*. [S.l.]: Universidade Federal do Amazonas, 2014.

FRIAS-MARTINEZ, E.; SANCHEZ, A.; VELEZ, J. Support vector machines versus multi-layer perceptrons for efficient off-line signature recognition. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 19, n. 6, p. 693–704, 2006.

GAO, D. et al. Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*, MIT Press, 2015.

GIACHANOU, A.; CRESTANI, F. Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, ACM, v. 49, n. 2, p. 28, 2016.

HATZIVASSILOGLOU, V.; MCKEOWN, K. R. Predicting the semantic orientation of adjectives. In: Association for Computational Linguistics. *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. [S.l.], 1997. p. 174–181.

JINDAL, N.; LIU, B. Review spam detection. In: ACM. *Proceedings of the 16th International Conference on World Wide Web*. [S.l.], 2007. p. 1189–1190.

- JOEL, P.; NADA, L.; DUNJA, M. A rule based approach to word lemmatization. In: *Proceedings C of the 7th International Multi-Conference Information Society IS*. [S.l.: s.n.], 2004.
- KATARIYA, M. N. P. et al. Text preprocessing for text mining using side information. *International Journal of Computer Science and Mobile Applications*, v. 3, n. 1, p. 01–05, 2015.
- KATZ, L. A new status index derived from sociometric analysis. *Psychometrika*, Springer, v. 18, n. 1, p. 39–43, 1953.
- KAUR, H.; MANN, E. J. S. Text normalization using statistical machine approach. *International Research Journal of Engineering and Technology*, 2016.
- KHAN, M. T.; KHALID, S. Sentiment analysis for health care. In: *Big Data: Concepts, Methodologies, Tools, and Applications*. [S.l.]: IGI Global, 2016. p. 676–689.
- KIM, S.-B. et al. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, IEEE, v. 18, n. 11, p. 1457–1466, 2006.
- KNÖBEL, C. *From random graphs to complex networks: a modelling approach*. Dissertação (Mestrado) — Vienna University Technology, 2009.
- KOLKUR, S.; DANTAL, G.; MAHE, R. Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, 2015.
- KOPPEL, M.; SCHLER, J. The importance of neutral examples for learning sentiment. *Computational Intelligence*, Wiley Online Library, v. 22, n. 2, p. 100–109, 2006.
- LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, JMLR. org, v. 18, n. 1, p. 559–563, 2017.
- LIU, B. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, v. 2, p. 627–666, 2010.
- LIU, B. Sentiment analysis and opinion mining: synthesis lectures on human language technologies. *Morgan and Claypool Publishers, San Rafael*, v. 5, n. 1, p. 1–167, 2012.
- LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: *Mining text data*. [S.l.]: Springer, 2012. p. 415–463.
- MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, Elsevier, v. 5, n. 4, p. 1093–1113, 2014.
- MUDINAS, A.; ZHANG, D.; LEVENE, M. Combining lexicon and learning based approaches for concept-level sentiment analysis. In: *ACM. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. [S.l.], 2012. p. 5.
- MUKHERJEE, A. et al. What yelp fake review filter might be doing? In: *ICWSM*. [S.l.: s.n.], 2013. p. 409–418.

- OPSAHL, T.; AGNEESSENS, F.; SKVORETZ, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, Elsevier, v. 32, n. 3, p. 245–251, 2010.
- PAGE, L. et al. *The PageRank citation ranking: Bringing order to the web*. [S.l.], 1999.
- PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing-Volume 10*. [S.l.], 2002. p. 79–86.
- PORTER, M. F. An algorithm for suffix stripping. *Program*, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980.
- RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*. [S.l.: s.n.], 2003. v. 242, p. 133–142.
- RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, Elsevier, v. 89, p. 14–46, 2015.
- ROSENTHAL, S.; FARRA, N.; NAKOV, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In: *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada: Association for Computational Linguistics, 2017. (SemEval '17).
- SAMMUT, C.; WEBB, G. I. *Encyclopedia of machine learning*. [S.l.]: Springer Science & Business Media, 2011.
- SCHEIBLE, C. et al. Sentiment translation through multi-edge graphs. In: Association for Computational Linguistics. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. [S.l.], 2010. p. 1104–1112.
- SILVA, T. C.; ZHAO, L. *Machine learning in complex networks*. [S.l.]: Springer, 2016. v. 2016.
- SPERIOSU, M. et al. Twitter polarity classification with label propagation over lexical links and the follower graph. In: Association for Computational Linguistics. *Proceedings of the First workshop on Unsupervised Learning in NLP*. [S.l.], 2011. p. 53–63.
- STEEN, M. *Graph Theory and Complex Networks: An Introduction*, Maarten van Steen. [S.l.]: Maarten van Steen, 2010.
- TABOADA, M. et al. Lexicon-based methods for sentiment analysis. *Computational linguistics*, MIT Press, v. 37, n. 2, p. 267–307, 2011.
- THAKKAR, M. H. V. *Twitter sentiment analysis using Hybrid Naïve Bayes*. 2013.
- THET, T. T.; NA, J.-C.; KHOO, C. S. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, Sage Publications Sage UK: London, England, v. 36, n. 6, p. 823–848, 2010.
- UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. *Information Processing & Management*, Elsevier, v. 50, n. 1, p. 104–112, 2014.

VIOLOS, J. et al. Sentiment analysis using word-graphs. In: *WIMS*. [S.l.: s.n.], 2016. p. 22.

VOHRA, M. S.; TERAIYA, J. Applications and challenges for sentiment analysis: A survey. In: ESRSA Publications. *International Journal of Engineering Research and Technology*. [S.l.], 2013. v. 2.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and information systems*, Springer, v. 14, n. 1, p. 1–37, 2008.

YU, H.; HATZIVASSILOGLU, V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2003. p. 129–136.

ZHU, X.; GHAHRAMANI, Z. *Learning from labeled and unlabeled data with label propagation*. [S.l.], 2002. Technical Report CMU-CALD-02-107.