UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

AFONSO MATHEUS SOUSA LIMA

ENRIQUECIMENTO DE DICIONÁRIOS PARA APRIMORAMENTO DA CLASSIFICAÇÃO AUTOMÁTICA DE SENTIMENTO EM POSTAGENS RELACIONADAS AO USO DE SISTEMAS

RUSSAS

2018

AFONSO MATHEUS SOUSA LIMA

ENRIQUECIMENTO DE DICIONÁRIOS PARA APRIMORAMENTO DA CLASSIFICAÇÃO AUTOMÁTICA DE SENTIMENTO EM POSTAGENS RELACIONADAS AO USO DE SISTEMAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Marilia Soares Mendes
Co-Orientadora: Prof. Ms. Lívia Almada Cruz Rafael

RUSSAS

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

L696e Lima, Afonso Matheus Sousa.
Enriquecimento de dicionários para aprimoramento da classificação automática de sentimento em postagens relacionadas ao uso de sistemas / Afonso Matheus Sousa Lima. – 2018.
47 f. : il.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2018.
Orientação: Profa. Dra. Marilia Soares Mendes.
Coorientação: Profa. Ma. Lívia Almada Cruz Rafael.

1. Avaliação de Sistemas. 2. Análise de Sentimentos. 3. Classificadores Léxicos. I. Título.
CDD 005

AFONSO MATHEUS SOUSA LIMA

ENRIQUECIMENTO DE DICIONÁRIOS PARA APRIMORAMENTO DA CLASSIFICAÇÃO AUTOMÁTICA DE SENTIMENTO EM POSTAGENS RELACIONADAS AO USO DE SISTEMAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em:

BANCA EXAMINADORA

Profa. Dra. Marilia Soares Mendes (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Ms. Lívia Almada Cruz Rafael (Co-Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Alexandre Matos Arruda
Universidade Federal do Ceará (UFC)

"Nem todos podem se tornar grandes artistas, mas um grande artista pode vir de qualquer lugar." (Ratatouille, 2007)

AGRADECIMENTOS

A Deus, por me proporcionar a experiência de todas as alegrias e tristezas, vitórias e derrotas que fazem parte desta oportunidade chamada vida.

A minha família, que sempre me apoiou e incentivou a continuar. Ao meu pai, Afonso Odério Nogueira Lima, por ser minha maior inspiração e por me mostrar que um grande artista pode vir de qualquer lugar. A minha mãe, Jackselene Maria de Sousa Lima, por sempre me mostrar que a educação é o único caminho a ser seguido. A minha irmã, Rebecca Sousa Lima, por ter me feito ser a pessoa que sou hoje.

A minha mãe de orientação, Lívia Almada Cruz Rafael, por ter acreditado em mim. Por ter me iniciado no mundo da pesquisa científica. Por, mesmo separados por grandes distâncias, nunca ter deixado de me auxiliar, criticar e motivar para eu fosse, não só um pesquisador, mas uma pessoa melhor.

A minha mãe de orientação, Marilia Soares Mendes, por ter me aceitado em seu projeto de pesquisa. Por ter me ensinado que mesmo que algo esteja bom, temos o dever de melhorar. Por me dizer quando eu estava errado. Por ter sempre me incentivado a ir mais longe. Por sempre me receber de braços abertos. E, acima de tudo, por ter confiado em mim e em meu trabalho.

A minha colega de graduação e pesquisa, e minha eterna amiga, Paloma Bispo dos Santos, por ter me ajudado quando mais precisei. Por ter me incentivado a nunca desistir. Por me mostrar o quão bom um ser humano pode ser. E, acima de tudo, por me fazer querer ser uma pessoa melhor.

Aos meus colegas de graduação, meus amigos, Erik Almeida, Carlos Victor, Guilherme Sombra, Hugo Venâncio, Igor Mendes, Isaias Ferreira, Marcos Alencar, Marcos Paulo, Mateus Oliveira, Thomas Dillan, Vinicius Almeida, Alex Frederico, Isaac Rahel, Marilia Cristina, Tágila Lima e José Leandro, por mostrarem que unidos somos mais fortes, e que unidos permaneceremos.

Aos meus professores e professoras, que contribuíram para a minha formação acadêmica e pessoal. Principalmente, aqueles que me fizeram ir mais longe.

Aos meus colegas de pesquisa do Projeto MALTU, Thiago, Douglas, Lavínia e Leon, por terem contribuído direta ou indiretamente com este trabalho. Por termos nos unido por um objetivo maior.

Ao Laboratório INterdisciplinar de Computação e Engenharia de Software – LINCE

e seus membros, por terem me dado a oportunidade de pesquisar e conhecer pesquisadores brilhantes.

A UFC e FUNCAP, pelo apoio dado em minha formação acadêmica em todos os níveis. Aos meus supervisores da FUNCAP, Nelson Costa e Alysson Martins, por terem contribuído para a minha formação profissional.

Aos meu amigos, Paulo Filho, Samuel Moura, Bruno Andrade, Ian Braúna, Antônio Filho, Leonardo Tiraboschi, Henrique Alves, Bruno Mendonça, João Gabriel, Júlio Biasoli e todos os outros, pelos bons momentos que me proporcionaram e pelas preocupações que me fizeram esquecer.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

Por fim, a todos que participaram diretamente ou indiretamente da minha formação nesses quatro anos de graduação. Seja de forma positiva ou negativa, mas que me ajudou a ser a pessoa que sou hoje.

Obrigado a todos!

RESUMO

A avaliação textual consiste em usar narrativas dos usuários a fim de avaliar ou obter alguma percepção sobre o sistema por meio de suas postagens. Tal avaliação é muito importante, pois a crescente popularização e disseminação das aplicações desenvolvidas e o aumento do padrão de qualidade exigido pelos usuários são características cada vez mais presentes atualmente. Torna-se, então, necessário que as desenvolvedoras de softwares aprimorem suas aplicações regularmente, uma vez que a concorrência sempre está presente e a necessidade do usuário pode mudar repentinamente. Por isso, é indispensável que sejam estudadas formas de efetuar a avaliação textual de forma automática, uma vez que classificar manualmente a grande quantidade de informação disponibilizada por usuários em tempo viável é uma tarefa impossível. Ainda há, também, a dificuldade de detectar o sentimento (positivo, neutro ou negativo) do usuário sobre o sistema por meio de postagens escritas em português. Baseado nessa problemática, este trabalho apresenta uma investigação para melhorar a qualidade da classificação de um classificador baseado em léxico, o *SentiStrength*, para detecção automática de sentimento em postagens relacionadas ao uso de sistemas. Para atingir tal objetivo, foi utilizada a métrica TF-IDF para a seleção de palavras que estão dentro do domínio das postagens relacionadas ao uso do sistema, as quais irão enriquecer o dicionário utilizado pelo *SentiStrength* para gerar a polaridade das postagens. Também foi investigada a eficiência de dicionários enriquecidos com palavras em sua raiz (*stemming*) e com palavras lematizadas. A investigação foi realizada com 2108 sentenças extraídas da seção de resenhas da Play Store sobre aplicativos de mobilidade urbana, sendo eles o Waze, Google Maps e GPS Brasil. Um dos resultados obtidos foi um aumento médio de 7,3% na acurácia do classificador quando utilizado os dicionários enriquecidos.

Palavras-chave: Avaliação de Sistemas. Análise de Sentimentos. Classificadores Léxicos.

ABSTRACT

Textual evaluation consists of using user narratives in order to assess or gain some insight into the system through its postings. This evaluation is very important, since the increasing popularization and dissemination of the developed applications and the increase of the quality standard demanded by the users are characteristics more and more present at the moment. It then becomes necessary for software developers to enhance their applications on a regular basis, since competition is always present and the user's need may change suddenly. For this reason, it is indispensable to study ways to perform textual evaluation automatically, since manually classifying the large amount of information made available by users in a viable time is an impossible task. There is also the difficulty of detecting the user's positive, neutral or negative feelings about the system through written Portuguese posts. Based on this problem, this work presents an investigation to improve the classification of a lexical-based classifier, the SentiStrength, for automatic classification of sentiments in postings related to the use of systems. To achieve this goal, the TF-IDF metric was used to select words that are within the domain of the related posts, which will enrich the dictionary used by the SentiStrength to generate the polarity of the posts. We also investigated the efficiency of dictionaries enriched with words in their root (stemming) and with lemmatized words. The research was conducted with 2108 sentences extracted from the reviews section of the Play Store on urban mobility applications, such as Waze, Google Maps and GPS Brasil. One of the results obtained was an average increase of 7.3 % in the accuracy of the classifier when using the enriched dictionaries.

Keywords: Systems Evaluation. Sentiment Analysis. Lexical Classifiers.

LISTA DE TABELAS

Tabela 1 – Classificação de postagens com e sem sentimeto 17
Tabela 2 – Exemplos de postagens para cada polaridade 18
Tabela 3 – Resumo dos trabalhos relacionados 31
Tabela 4 – Parte das palavras adicionadas a cada dicionário 37

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Motivação 13
1.2 Objetivos 14
1.2.1 Objetivo geral 14
1.2.2 Objetivo específicos 14
1.3 Metodologia 14
1.4 Organização do Trabalho 15
2 FUNDAMENTAÇÃO TEÓRICA 16
2.1 Avaliação Textual de Sistemas 16
2.2 Análise de Sentimentos 17
2.3 Recuperação da Informação 19
2.3.1 Classificação usando TF-IDF (Term Frequency - Inverse Document Frequency) 19
2.3.2 Recuperação baseada em semelhança 20
2.4 Pré-Processamento 21
2.5 Algoritmos de Classificação 22
2.6 Ferramentas de Classificação 23
2.6.1 SentiStrength 23
2.6.2 SentiWordNet 24
2.7 Métricas de Avaliação 25
3 TRABALHOS RELACIONADOS 27
3.1 Trabalhos que utilizam classificação automática em postagens 27
3.2 Trabalhos que utilizam o SentiStrength 28
3.3 Trabalhos que utilizam a métrica TF-IDF 29
3.4 Tabela-Resumo 31
4 INVESTIGAÇÃO 32
4.1 Extração dos Dados 32
4.2 Classificação dos Dados 32
4.3 Preparação dos Dados 33
4.4 Enriquecimento do Léxico 33

4.5 **Aplicação no *SentiStrength*** 34
5 **RESULTADOS** 36
5.1 **Métricas utilizadas** 36
5.2 **Palavras Adicionadas** 37
5.3 **Análise de Resultados** 38
6 **DISCUSSÃO** 41
7 **CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS** 43
REFERÊNCIAS 44

1 INTRODUÇÃO

A crescente popularização e disseminação das aplicações desenvolvidas e o aumento do padrão de qualidade exigido pelos usuários são características cada vez mais presentes atualmente. Torna-se, então, necessário que as desenvolvedoras de softwares aprimorem suas aplicações regularmente, uma vez que a concorrência sempre está presente e a necessidade do usuário pode mudar repentinamente. Para isso, a opinião dos usuários é muito importante para a melhoria de qualquer aplicativo, pois é uma fonte de críticas, elogios e dúvidas que são disponibilizadas em seções de resenhas das lojas que disponibilizam essas aplicações.

Dito isso, temos a Avaliação Textual (MENDES, 2015), que consiste em usar narrativas dos usuários a fim de avaliar ou obter alguma percepção sobre o sistema por meio de suas postagens. Esse tipo de avaliação faz uso de Postagens Relacionadas ao Uso (PRUs). Uma PRU é um comentário espontâneo publicado pelos usuários, ao qual se refere ao uso de um sistema. Por exemplo: *"Antes era o melhor aplicativo do mercado, hoje deixa a desejar"*. Tal exemplo foi retirado da seção de resenhas da *Play Store* de um aplicativo de mobilidade urbana. PRUs tem como característica a espontaneidade, uma vez que os usuários não são solicitados a fazerem um relato de seu uso. Durante seu uso, espontaneamente, eles mencionam fatos sobre o sistema, como: suas dúvidas, elogios, dificuldades, críticas, sugestões ou mesmo relatos de experiência. Os autores (MENDES *et al.*, 2015) afirmam que a espontaneidade do relato é importante, uma vez que o questionamento por avaliadores, por meio de entrevistas ou questionários, ou pelo próprio sistema, pode influenciar as respostas fornecidas durante seu uso (FETTER *et al.*, 2011; KORHONEN *et al.*, 2010).

Existe uma técnica chamada MALTU para avaliação de sistemas que considera as PRUs dos usuários para fornecer uma percepção do uso do sistema, como a satisfação ou insatisfação dos usuários (MENDES; FURTADO, 2018). A MALTU possui cinco etapas de avaliação: (1) definição do contexto de avaliação; (2) extração de PRUs; (3) classificação das PRUs; (4) interpretação de resultados e (5) relato dos resultados. Para auxiliar nas etapas 2 e 3 da metodologia MALTU, as quais são extração e classificação de PRUS, respectivamente, existe uma ferramenta chamada UUX-Posts [1] (MENDES, 2015) (MENDES; FURTADO, 2017). A ferramenta extrai e classifica as postagens automaticamente, a fim de fornecer um resultado de avaliação. No entanto essa ferramenta apresenta algumas limitações (MENDES; FURTADO, 2017), por ser usado o modelo de busca booleano, por não disponibilizar técnicas de pré-

[1] UUX-Posts. Disponível em http://uuxposts.russas.ufc.br. Acesso em 25 de novembro de 2018

processamento de texto e por não trabalhar com a classificação automática de sentimentos das postagens. Tal funcionalidade de classificação automática de sentimentos é muito importante para que, a partir do comentário do usuário, possa-se inferir qual o sentimento dele em relação ao sistema, ou seja, se ele está sendo positivo, neutro ou negativo sobre um sistema.

Em um trabalho anterior do autor deste TCC (LIMA *et al.*, 2017), investigou-se a classificação automática de sentimento em postagens de usuários em um sistema social. Utilizou-se dois tipos de classificadores, o primeiro consistiu na utilização de um algoritmo probabilístico, o Naive Bayes, e o segundo consistiu no uso de um classificador léxico, o *SentiStrength* (THELWALL *et al.*, 2011). Embora o classificador probabilístico tenha obtido resultados satisfatórios para a classificação automática de sentimentos, o classificador léxico *SentiStrength* se mostrou pouco eficiente para realizar a mesma tarefa. Tal resultado pôde ser atribuído a falta de palavras, pertencentes ao domínio do sistema, no dicionário utilizado pelo classificador léxico.

Portanto, este trabalho propõe uma investigação para aprimorar a classificação de sentimentos de PRUs usando o classificador léxico *SentiStrength* . Para isso, foram adicionadas no dicionário palavras selecionadas pela métrica matemática *Term Frequency Inverse Document Frequency* (TF-IDF), que se mostra bastante eficiente para categorizar as palavras mais relevantes de um documento (RAMOS *et al.*, 2003). Para verificação de eficiência, o resultado da classificação com o dicionário aprimorado é comparado com o resultado da classificação com dicionário antigo. Também são comparados os resultados de outros dois dicionários enriquecidos criados a partir da mesma base de treinamento, porém foram aplicadas diferentes técnicas de pré-processamento, sendo uma delas a transformação das palavras para sua forma raiz (*stemming*), e a outra para a forma lematizada, ou seja, as palavras serão transformadas para o singular masculino, para substantivos, ou para o infinitivo, no caso de verbos. Os resultados obtidos com este trabalho, além de contribuir para o avanço dos estudos sobre classificadores léxicos aplicados à língua portuguesa, poderão ser usados para aprimorar a ferramenta UUX-Posts, adicionando a funcionalidade de classificação de sentimentos.

1.1 Motivação

Este trabalho faz parte do projeto de pesquisa intitulado: Avaliação da interação em sistemas sociais a partir da linguagem textual do usuário, coordenado pela Profa. Dra. Marília Soares Mendes, iniciado em 2015 e financiado pela Fundação Cearense de Apoio

ao Desenvolvimento Científico e Tecnológico (FUNCAP) no período de 2015 a 2018. Este projeto tem por objetivo estudar e implementar novas técnicas de avaliação da interação em sistemas a partir da linguagem textual do usuário. O autor deste TCC participa do projeto desde Maio de 2016, e durante este período teve experiência com: Mineração de Dados; Extração e Classificação de Postagens; Análise de Sentimentos. O autor deste TCC também teve um artigo aceito e publicado no Human Computer Interaction International 2017 (LIMA *et al.*, 2017).

Com esses anos de projeto e conjunto de experiências, foi possível perceber a importância da avaliação textual de postagens do usuário para a contínua evolução de sistemas, assim como a detecção de sentimentos. Para realizar tais procedimentos, é indispensável a realização de estudos sobre esses assuntos e que sejam feitas investigações a fim de verificar a validade de soluções propostas, principalmente quando se está trabalhando com postagens escritas em português.

1.2 Objetivos

1.2.1 Objetivo geral

Aprimorar a classificação de sentimentos de PRUs usando o classificador léxico *SentiStrength*, enriquecendo o dicionário com palavras obtidas pela métrica TF-IDF que reflitam o domínio do sistema.

1.2.2 Objetivo específicos

- Fornecer um estado da arte sobre análise de sentimentos, recuperação da informação e classificadores;
- Apresentar resultados de um experimento de Análise de Sentimentos para classificação automática de postagens.

1.3 Metodologia

A metodologia aplicada na execução deste trabalho é constituída tanto de estudos teóricos como de estudos práticos.

Na parte de estudos teóricos, são apresentados assuntos e conteúdos necessários para o bom entendimento e boa execução da investigação que este trabalho propõe. Para isso, são

fundamentados os seguintes assuntos: Avaliação textual de sistemas; Análise de sentimentos; Recuperação da informação; Pré-processamento; Algoritmos de classificação; Ferramentas de classificação e métricas de avaliação. Também é feito um estudo e apresentação sobre trabalhos relacionados que abordaram, pelo menos, um dos assuntos listados anteriormente.

Na parte de estudos práticos, é feita uma investigação para verificar a eficiência da métrica TF-IDF para seleção de palavras que irão enriquecer o dicionário do *SentiStrength*. A investigação foi executada seguindo as seguintes etapas: Extração dos dados; Classificação dos dados; Preparação dos dados; Enriquecimento do Léxico; Aplicação no *SentiStrength* e Análise dos resultados. Também foifeita uma discussão sobre o experimento e como o estudo feito neste trabalho impacta na classificação automática de sentimento utilizando o classificador léxico *SentiStrength* em PRUs.

1.4 Organização do Trabalho

Este trabalho se divide em 7 capítulos. O Capítulo 1 mostra, de forma resumida, o problema e suas motivações, bem como o objetivo geral e específicos desta pesquisa e a metodologia utilizada ao longo do desenvolvimento do trabalho. O Capítulo 2 aborda os conhecimentos teóricos necessários para a resolução do problema. Trabalhos relacionados são apresentados no Capítulo 3. No Capítulo 4 é apresentada a investigação realizada, detalhando todos os seus passos. O Capítulo 5 apresenta os resultados obtidos com a investigação. No Capítulo 6, são discutidos os resultados obtidos e seu impacto. Por fim, no Capítulo 7 apresenta-se a conclusão sobre o trabalho, seguida dos trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo é focado em apresentar os conceitos necessários para o entendimento deste trabalho e da investigação apresentada no Capítulo 4.

2.1 Avaliação Textual de Sistemas

A avaliação textual de sistemas consiste na utilização de narrativas de usuários para avaliar ou obter alguma percepção sobre o sistema a ser avaliado (MENDES, 2015). É possível avaliar um ou mais critérios de qualidade de uso com avaliação textual, como usabilidade, experiência do usuário ou suas facetas (satisfação, memorabilidade, aprendizado, eficiência, eficácia, etc.) (HEDEGAARD; SIMONSEN, 2013; MENDES, 2015; MENDES *et al.*, 2015; KORHONEN *et al.*, 2010; PARTALA; KALLINEN, 2011).

Uma metodologia que é utilizada para avaliação de sistemas que utiliza das PRUs para fornecer uma percepção do uso do sistema é chamada MALTU. Ela propõe a coleta das postagens dos usuários no próprio sistema em uso a fim de considerar a espontaneidade do usuário (MENDES, 2015). A metodologia MALTU propõe cinco etapas no processo de avaliação:

- A primeira etapa consiste na definição do contexto da avaliação, no qual envolve o contexto de uso e domínio do sistema, e os objetivos da avaliação;
- A segunda etapa consiste na extração de PRUs, a qual pode ser feita de forma manual ou automática. A forma automática de extração pode ser feita utilizando *Web Crawlers* ou utilizando ferramentas, como o UUX-Posts (MENDES, 2015) (MENDES; FURTADO, 2017), a qual extrai e classifica as postagens automaticamente, a fim de fornecer um resultado de avaliação;
- Na terceira etapa é realizada a classificação de PRUs. A MALTU propõe os seguintes tipos de classificação:
  - classificação por tipo (crítica, elogio, ajuda, dúvida, comparação e sugestão);
  - classificação por intenção (visceral, comportamental e reflexiva);
  - análise de sentimento (positivo, neutro ou negativo);
  - classificação por funcionalidades;
  - classificação por critério de qualidade de uso (usabilidade, experiência do usuário);
  - classificação por artefato (dispositivos utilizados pelos usuários).
- A quarta etapa consiste na interpretação dos resultados da classificação. Ela pode ser feita

utilizando métricas matemáticas para avaliação, como o quantitativo de postagens obtidas por classificação e o relacionamento entre elas, auxiliando na visualização e percepção de usuários;

- A quinta etapa consiste no relato dos resultados.

A investigação feita nesse trabalho é baseada nessa metodologia, sendo que o tipo de classificação de PRUs que é abordado é a análise de sentimentos. Nesse tipo de classificação, é classificada cada PRU como positiva, negativa ou neutra, representando o sentimento do usuário em relação ao sistema. Como exemplo, em um trabalho, os autores (MENDES *et al.*, 2014) fizeram uma investigação em PRUs do Twitter a fim de investigar a presença e a ausência de sentimentos nas PRUs e observaram que tanto as postagens com sentimento, como aquelas ausentes de sentimentos (neutras) são importantes para obtenção de alguma percepção do sistema. Exemplos podem ser vistos na Tabela 1. Neste sentido, a identificação da polaridade das PRUs seria relevante para avaliação da satisfação ou insatisfação (frustração) do usuário no sistema, enquanto as postagens neutras seriam relevantes para identificação de dúvidas em funcionalidades do sistema.

Tabela 1 – Classificação de postagens com e sem sentimeto

Postagem com Sentimento	**Postagem sem Sentimento**
Eu amoooo quando eu quero postar algo no Twitter, ai então o servidor cai!	*O Twitter não deixa eu postar GIFs*
ODEIO o novo Twitter!	*Quando teremos o botão de editar no Twitter?*

Fonte: Adaptado de (MENDES *et al.*, 2014)

No entanto, a detecção automática da polaridade em sentenças é um desafio científico complexo, devido à subjetividade presente na linguagem natural. Segundo Silva (2004), mineração de opinião ou análise de sentimentos é uma área de pesquisa que tem por objetivo definir técnicas automatizadas para extrair informação subjetiva de textos em linguagem natural, como opiniões e sentimentos. Esse assunto é detalhado mais na seção 2.2.

2.2 Análise de Sentimentos

A análise de sentimentos é o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, atitudes e emoções relacionados a produtos, serviços, organizações, pessoas, problemas, eventos etc., expressos em textos (revisões, blogs, discussões, notícias,

comentários, feedback, ou quaisquer outros documentos) (LIU, 2012). Para Wilson *et al.* (2009), a análise de sentimentos é um tipo de análise de subjetividade que foca em identificar opiniões positivas e negativas, emoções e avaliações expressas em linguagem natural.

Em uma postagem, quando se procura classificar seu sentimento, pode ser atribuída uma das seguintes polaridades: positiva, neutra ou negativa. A Tabela 2 mostra exemplos de postagens para cada uma dessas polaridades.

Tabela 2 – Exemplos de postagens para cada polaridade

Positiva	**Neutra**	**Negativa**
Ótimo não me deixa na mão	Sem espaço pra atualizar	Decepção esse app da Google!
Perfeito, estão de parabéns!	Poderia ser offline	Está uma porcaria travando direto!
Ótimo aplicativo, informações precisas	Como adiciona rota?	Burro esse aplicativo!

Fonte: Elaborado pelo Autor

Uma postagem também pode ser subjetiva ou objetiva (LIU, 2012). Uma postagem é subjetiva quando o locutor está expressando o seu ponto de vista ou sua perspectiva sobre um assunto. Uma postagem é objetiva quando fatos estão sendo apresentados, ou seja, são informações comuns a um conjunto de locutores. Em suma, sentenças objetivas apresentam informações sobre fatos do mundo, enquanto uma sentença subjetiva expressa alguns sentimentos pessoais, opiniões ou crenças (LIU, 2012).

De acordo com Medhat *et al.* (2014), as técnicas de análise de sentimentos podem ainda ser classificadas segundo a abordagem que utilizam em: baseadas em léxico, que utilizam um léxico de sentimentos (coleção de itens de sentimentos pré-compilados, como por exemplo, dicionários); baseadas em aprendizado de máquina, que fazem uso dos algoritmos já conhecidos de aprendizado de máquina para classificação de textos; ou híbridas, que combinam as abordagens já mencionadas. No trabalho de Lima *et al.* (2017), foram feitas duas investigações em PRUs espontâneas de Sistemas Sociais, uma utilizando o classificador léxico, e a outra utilizando o classificador probabilístico juntamente com o aprendizado de máquina. Para esse trabalho, o classificador probabilístico obteve resultados melhores que o classificador léxico para classificação de polaridade dessas PRUs.

Neste trabalho, a análise de sentimentos é focada na polaridade: positiva, neutra ou negativa. Para isso, é utilizado um programa, o *SentiStrength*, que identifica se uma determinada

postagem de um usuário é positiva, neutra ou negativa baseado nas palavras que estão presentes naquela postagem. Com isso, aplicando esse programa a um corpus de PRUs, pode-se descobrir o sentimento dos usuários em relação ao sistema de forma automática. Isso possibilita aos avaliadores de sistemas ter um contato imediato com a opinião do usuário e assim, tomar decisões sobre o sistema no qual as PRUs usadas pertenciam.

2.3 Recuperação da Informação

O processo de recuperação de informações consiste em localizar documentos relevantes, com base na entrada do usuário, como palavras-chave ou documentos de exemplo (SILBERSCHATZ *et al.*, 2016). Um usuário de um sistema pode querer recuperar um documento em particular ou uma classe específica de documentos. Os documentos normalmente são descritos por um conjunto de palavras-chave. Essas palavras são especiais, pois com elas pode-se realizar consultas para encontrar documentos em que essas palavras tenham uma grande significância dentro do contexto do documento que se deseja encontrar. Um exemplo seria a realização de uma consulta utilizando as palavras-chave *"ações"* e *"escândalo"* para localizar artigos sobre escândalos no mercado de ações (SILBERSCHATZ *et al.*, 2016)

No entanto, o conjunto de todos os documentos que satisfazem uma expressão de consulta pode ser muito grande. Para contornar esse problema, são utilizadas métricas que definem os termos mais relevantes, ou seja, com avaliações mais altas atribuidas a eles. A seguir, temos algumas técnicas bem aceitas para a classificação de relevância.

2.3.1 Classificação usando TF-IDF (Term Frequency - Inverse Document Frequency)

Como mostrado por Ramos *et al.* (2003), o TF-IDF é uma métrica matemática que determina a frequência relativa de palavras em um determinado documento em comparação com a proporção inversa dessa palavra sobre todos os outros documentos. Ou seja, este cálculo determina a relevância de uma determinada palavra em um documento específico.

Dado uma coleção de documentos D, uma palavra w, e um documento individual $d \varepsilon D$, a formulação matemática do TF-IDF é apresentada a seguir:

$$w_d = f_{w,d} * log(|D|/f_{w,D}) \tag{2.1}$$

Onde $f_{w,d}$ representa a frequência de vezes que w aparece em d, $|D|$ é a quantidade

total de documentos, e $f_{w,D}$ é igual a quantidade de documentos onde w aparece em D. Note que se $f_{w,d}$ for grande e $f_{w,D}$ for pequeno, a palavra w terá um valor bem alto, ou seja, elá terá uma relevância bem maior para o documento d, mas não para a coleção de documentos D. De forma intuitiva, se uma palavra aparecer frequentemente em um texto específico, então ela pode ser boa para classificar esse texto, no entanto, se essa palavra aparecer também nos outros documentos, então ela é uma palavra comum e não uma específica dentro do contexto do conjunto de documentos. Ou seja, é necessário que a palavra apareça muito em um documento específico e não apareça em outros documentos do conjunto.

O TF-IDF é apresentado como uma métrica cuja implementação é simples, mostrando-se eficiente para identificação de palavras relevantes em um conjunto de documentos. Além disso, sua fácil implementação o permite ser usado como base para que melhoramentos sejam feitos. No entanto, suas limitações são aparentes quando se quer classificar palavras semelhantes como um conjunto e não individualmente. Um exemplo seria a palavra "você"e a palavra "vocês", as quais o TF-IDF não reconheceria como sinônimo. Tal situação também ocorre quando se trabalha com palavras abreviadas e ou escritas incorretamente, as quais são muito comuns em ambientes digitais.

2.3.2 Recuperação baseada em semelhança

Na recuperação baseada em semelhança, o usuário pode usar um documento como entrada para um sistema encontrar outros documentos que são semelhantes a ele. A semelhança entre documentos pode ser definida com base nos termos comuns. Para isso, pode-se utilizar os termos com maiores valores TF-IDF como consulta (SILBERSCHATZ *et al.*, 2016). Assim, teremos capturado palavras-chave de um documento utilizando a métrica TF-IDF.

A relevância também pode ser usada para ajudar os usuários a encontrarem documentos relevantes a partir de um grande conjunto de documentos combinando com as palavras-chave de consulta indicadas. Ou seja, se o corpus de documentos retornados for muito extenso, é dada a liberdade ao usuário de escolher entre os documentos encontrados os mais relevantes na visão deles e, assim, realizar uma nova consulta adicionando esses novos documentos para o refinamento da busca. Essa idéia é chamada relevância por *feedback*.

2.4 Pré-Processamento

O pré-processamento é uma das etapas mais importantes no processo de aprendizagem de máquina e extração de conhecimento de bases textuais (IMAMURA, 2001). É nela que se padronizam todos os dados coletados na etapa de extração com a finalidade de facilitar o reconhecimento deles pelos algoritmos, seja com a remoção de valores inúteis, seja com a transformação para outras representações, por exemplo, uma transformação de textual para numérica.

Essa etapa é ainda mais importante quando se trabalha com textos não estruturados da *Web*. Diferentemente de textos estruturados, como questionários, postagens de usuários retiradas de sistemas sociais são sujeitas a ironia, gírias, abreviações e todas particularidades da linguagem que são regularmente usados pelos usuários de sistemas sociais (LIMA *et al.*, 2017). E com isso, ainda existe erros gramaticais, *emoticons*, caracteres especiais, entre outros que precisam ser devidamente tratados na etapa de pré-processamento, adaptando esse processo ao algoritmo que se pretende utilizar. Para isso, pode-se remover ou transformar os dados que foram extraídos.

Uma das técnicas mais utilizadas para pré-processamento de texto consiste na transformação das palavras para a sua forma raiz, chamada de *stemming* (LOVINS, 1968). A idéia central é que toda palavra possa ser representada por um *stem*, que é uma representação mínima não ambígua do termo. Para isso o algoritmo retira algumas letras no final das palavras, assumindo que a cadeia de caracteres restante ainda represente a essência delas (ALVARES, 2005). O *stemming* é utilizado quando é necessário reduzir um conjunto de palavras para uma raiz (*stem*) em comum, objetivando auxiliar nos processos de linguística computacional e recuperação da informação (LOVINS, 1968). Um exemplo seria as palavras *"estou"*, *"estiveram"*, *"estive"* que tem como raiz (*stema*) o termo *"est"*. No entanto, a definição dos *stems* podem variar conforme a implementação do algoritmo de *stemming*.

Outra técnica utilizada quando se trabalha com texto é a lematização. A lematização é o ato de representar as palavras por meio do infinitivo dos verbos e masculino singular dos substantivos e adjetivos (LUCCA; NUNES, 2002). Também se pode definir outros padrões na implementação do lematizador, conforme a necessidade do experimento que se deseja realizar. Tal como o *stemming*, a lematização tem o objetivo de reduzir um conjunto de palavras para um termo em comum. Um exemplo seria as palavras *"melhorou"*, *"melhorando"* e *"melhora"* que podem ser representadas pelo termo *"melhor"*.

2.5 Algoritmos de Classificação

A classificação é uma das tarefas mais utilizada na área de mineração de dados (AMARAL, 2016). Ela também é a mais complexa e a que possui maior diversidade de algoritmos. A classificação consiste em utilizar um atributo (propriedade ou característica de um objeto) entre um conjunto de atributos. Esse atributo especial é chamado de classe e o objetivo é utilizar todos os outros atributos para tentar prever a classe. Para isso, é necessário encontrar relações entre esses outros atributos de forma que sejam encontrados padrões que indiquem o atributo da classe. É possível tratar a análise de sentimentos como um problema de classificação.

Existem algoritmos de classificação muito conhecidos e amplamente aplicados, um deles é o Naive Bayes. Sendo um algoritmo bayesiano, baseado na teoria das probabilidades e na lei de Bayes, supõe também que os atributos vão influenciar a classe de forma independente (AMARAL, 2016). A fórmula da lei de Bayes é a seguinte:

$$P(X|Y) = \frac{P(X) * P(Y|X)}{P(Y)} \tag{2.2}$$

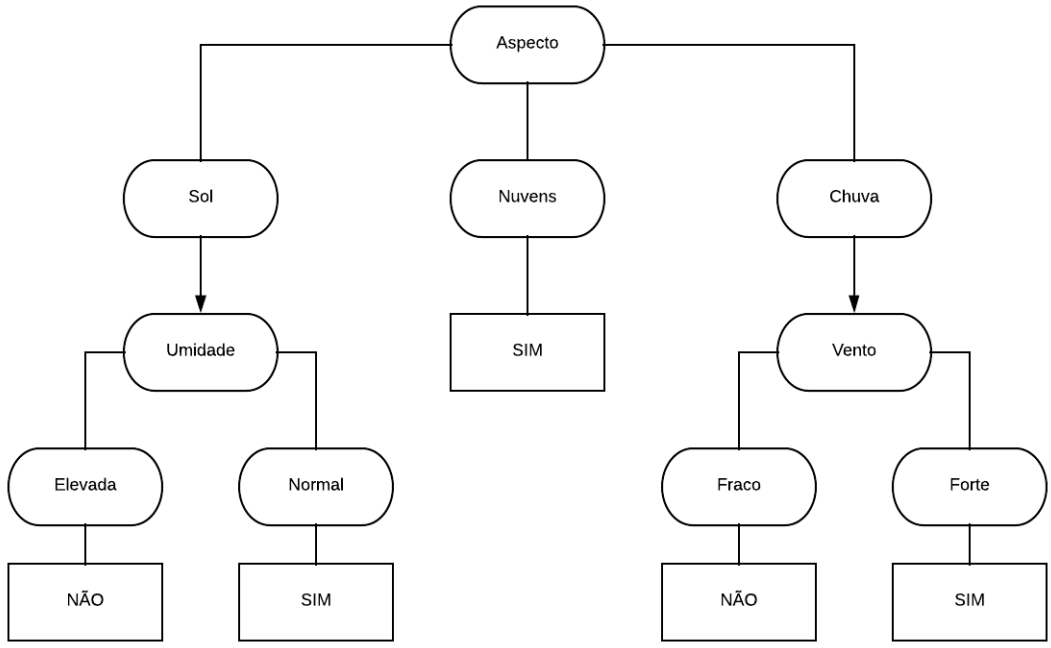
Onde $P(X/Y)$ é a probabilidade posterior da classe X dada o preditor Y, $P(X)$ é a probabilidade original da classe X, $P(Y/X)$ representa a probabilidade de Y dado X e $P(Y)$ é a probabilidade original do preditor Y.

Para realizar a classificação, é construída uma tabela mostrando o quanto cada categoria de cada atributo contribui para cada classe. Após essa construção, na inserção de um novo objeto no conjunto, o classificador vai verificar os pesos presentes na tabela construída anteriormente, somá-los e inferir qual classe teve um peso maior, a qual será atribuída ao novo objeto. Tal algoritmo, apesar de simples, é extremamente eficiente, superando até mesmo algoritmos mais complexos em alguns casos. Além disso, o Naive Bayes tem uma excelente performance do ponto de vista computacional (AMARAL, 2016) e também teve bons resultados para classificar o sentimento de PRUs na investigação feita por Lima *et al.* (2017).

Cita-se, também, as árvores de decisão, as quais são uma das famílias mais populares de classificadores. Seu princípio é simples, é criada uma estrutura de decisão em forma de árvore e, a cada novo objeto inserido, essa árvore será percorrida até chegar a um nó onde está a classe (AMARAL, 2016). Uma árvore de decisão, geralmente, começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, as árvores de decisão podem ser facilmente convertidas em regras

de classificação. Um exemplo pode ser visto na Figura 1, que decide se é possível jogar tênis baseado no clima.

Figura 1 – Exemplo de árvore de decisão



Fonte: Elaborado pelo Autor

2.6 Ferramentas de Classificação

Existem algumas ferramentas de classificação baseada em léxico que são focadas em descobrir a polaridade das postagens. Diferentemente dos algoritmos apresentados anteriormente, os classificadores léxicos são usados principalmente para classificação textual, uma vez que é utilizado da própria estrutura léxica da linguagem para realizar a classificação de sentimentos. Duas das ferramentas mais conhecidas baseadas no léxico, o *SentiStrength* e o *SentiWordNet*, serão apresentadas a seguir:

2.6.1 SentiStrength

O *SentiStrength*[1] é um programa utilizado para análise de sentimentos (THELWALL *et al.*, 2011). Ele estima a força do sentimento positivo e negativo em textos curtos, até mesmo quando se trabalha com linguagem informal. Atesta-se que ele tem boa acurácia para classificar

[1] SentiStrength. Disponível em: http://sentistrength.wlv.ac.uk. Acesso em 25 de Novembro de 2018

pequenos textos em Inglês, como pode ser visto no trabalho de Thelwall (2017). No entanto, ele também se mostrou eficiente para trabalhar com postagens em idiomas diferentes, como o espanhol (VILARES *et al.*, 2015). Em relação ao seu funcionamento, ele utiliza um dicionário que contém várias palavras de um determinado idioma e para cada palavra é atribuída uma força, podendo ser ela somente 1 dos seguintes valores: uma positiva, cuja força varia de 1 (Não positiva) a 5 (Extremamente positiva), ou uma negativa, cuja força varia de -1 (Não negativa) até -5 (Extremamente negativa), ou 0 se a palavra não estiver contida no dicionário, sendo classificada como neutra (THELWALL *et al.*, 2011).

Utilizando as valorações do dicionário, cada palavra de cada sentença do documento de texto fornecido como entrada recebe sua respectiva força positiva ou negativa. Então, as maiores forças positivas e negativas são atribuídas para cada sentença. Com esses valores, pode-se saber qual a polaridade dessa sentença, ou seja, ela será positiva se a força positiva for maior que a negativa, e será negativa se a força negativa for maior que a força positiva. Palavras serão consideradas neutras quando atribuídas com os valores 1 ou -1, ou quando ela não estiver presente no dicionário. Um exemplo seria a postagem: *"Já foi bom, hoje está horrível"*. O *SentiStrength*, utilizando do dicionário, atribuirá as seguintes forças: "Já[0] foi[0] bom[3], hoje[0] está[0] horrível[-5]". Com isso, a maior força positiva será 3 e a maior força negativa será -5, portanto, essa postagem será classificada como negativa, pois é a maior força.

Deve-se ressaltar que os dicionários que o *SentiStrength* utiliza podem ser modificados, seja adicionando ou removendo palavras, assim como, fazendo alterações em suas forças atribuídas, tanto positivas como negativas. Também é possível identificar uma palavra como raiz adicionando posteriormente a ela um asterisco (*), para então todas as palavras com essa raiz sejam valoradas da mesma forma.

2.6.2 SentiWordNet

O *SentiWordNet*[2] (ESULI; SEBASTIANI, 2007) é uma ferramenta muito utilizada em mineração de opinião, e é baseada no dicionário léxico WordNet (MILLER, 1995). Esse dicionário agrupa adjetivos, verbos e outras classes gramaticais em conjuntos chamados *synset*. O *SentiWordNet* associa a cada *synset* do WordNet três valores de pontuação que indicam o sentimento de um texto: positivo, negativo e objetivo (neutralidade).

Cada pontuação é obtida utilizando um método de aprendizagem de máquina semi-

[2] SentiWordNet. Disponível em: https://sentiwordnet.isti.cnr.it. Acesso em 25 de Novembro de 2018

supervisionada, e variam de 0 a 1, com soma igual a 1 (ARAÚJO *et al.*, 2013). Para melhor entender o funcionamento do método, Araújo *et al.* (2013) cita o seguinte exemplo: para um dado *synset* s = [ruim, estranho, terrível] tenha sido extraído de uma postagem. O resultado obtido pelo método é 0,000 para positividade, 0,850 para negatividade e 0,150 para objetividade (Pos(s)=0, Neg(s)=0.85 e Obj(s)=0.15). Ou seja, o valor negativo é superior aos outros dois.

É também possível notar que o *SentiWordNet* tem uma pontuação diferente para diferentes significados de um mesmo termo, variando de acordo com o contexto. Por isso a atribuição de valores numéricos é feita para um *synset*, ao invés de valores direto a um termo, possibilitando de um mesmo termo ter diferentes sentidos e cada um desses sentidos merece uma pontuação diferente (ESULI; SEBASTIANI, 2007). Um exemplo de Sousa (2016) seria a palavra *"broken"*, que está relacionada a dois *synset*, S1 = *wiped out, impoverished, broken* que tem como descrição *"destroyed financially or the broken fortunes of the family"* e pontuações Obj(s)=0.5, Pos(s)=0 e Neg(s)=0.5, e S2=*broken* que tem a descrição *"physically and forcibly separated into pieces or cracked or split"* e pontuações Obj(s)=0.875, Pos(s)=0 e Neg(s)=0.125

Infelizmente, seria necessário a criação de novos *synsets* próprios para a língua portuguesa e para o contexto de avaliação de sistemas. Tal atividade necessitaria de mais estudos, tanto sobre a ferramenta e suas características, como sobre própria estrutura do idioma que iria ser aplicada, o português.

2.7 Métricas de Avaliação

Umas das maneiras de avaliar a corretude de uma classificação automática é utilizar métricas matemáticas que auxiliam na representação da relação entre os resultados da classificação automática com a classificação verdadeira que é feita anteriormente, como uma classificação manual, por exemplo. Baseado na descrição de Powers (2011), as métricas que serão utilizadas neste trabalho são:

- a Acurácia, a qual consiste na medida mais básica de avaliação, sendo a fração de instâncias corretamente classificadas. Por exemplo: se o classificador classifica como positiva 10 de 20 postagens que realmente são positivas, como negativa 15 de 30 postagens que são realmente negativas e como neutra 5 de 10 postagens que são realmente neutras, então teremos 30 acertos de 60, ou seja, 50% de acurácia;
- a Precisão, a qual determina quantas instâncias de uma classe foram corretamente previstas dado todas instâncias previstas dessa classe. Por exemplo: se o classificador classifica

corretamente como positiva 25 de 50 postagens que ele previu como positiva, então teremos 50% de precisão para postagens positivas;

- 0 *Recall* fornece quantas instâncias de uma classe foram corretamente previstas dado todas as instâncias que verdadeiramente pertencem a esta classe. Por exemplo: se o classificador classifica corretamente como positiva 25 de 100 postagens que realmente são positivas, então teremos 25% de *recall* para postagens positivas;
- *F-Measure* é a média harmônica entre a precisão e o *recall*. O cáculo é obtido com a seguinte fórmula: $F = 2 * \frac{precisao * recall}{precisao + recall}$. Por exemplo: se a classificação do classificador teve precisão de 75% e *recall* de 50%, então o *F-measure* dessa classificação será de 60% ($F = 2 * \frac{0,75 * 0,50}{0,75 + 0,50} = 0,60$).

3 TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos que utilizaram conceitos e técnicas semelhantes as propostas por este.

3.1 Trabalhos que utilizam classificação automática em postagens

No trabalho feito por Rolim *et al.* (2016), é proposta uma solução computacional para análise de identificação de dúvidas em fóruns educacionais. O sistema proposto realiza a classificação de postagens em três categorias: dúvida, postagem neutra e resposta. Com essa informação o professor pode responder as dúvidas e avaliar a participação dos alunos, criar grupos de estudo, recomendar material complementar para o estudo, entre outros. Para avaliar o sistema, um *dataset* contendo postagens de fóruns educacionais foi criado e foram testadas várias técnicas de classificação de texto, sendo que as técnicas de rede bayesiana, árvore de decisão e a rede neural artificial foram as que alcançaram melhores resultados chegando a obter uma taxa de acerto de 97%.

A metodologia utilizada por esse trabalho (ROLIM *et al.*, 2016) foi muito semelhante a este, onde houve etapas de coletas de dados e pré-processamento. Também foi feita uma etapa de extração de características, onde foram extraídas as características que se tornariam os atributos das técnicas de classificação. Para isso, foi recuperada a frequência das palavras para cada classe utilizada: dúvida, postagem neutra e resposta. Outra característica utilizada como atributo foi a somatória do valor TF-IDF das palavras pertencentes de cada classe. Após obtidos os atributos, foi realizada a etapa de classificação e assim, aplicadas as métricas de avaliação.

No entanto, esse trabalho não realizou estudos voltados a análise de sentimentos das postagens, ou seja, não se sabe se os usuários estão sendo positivos ou negativos nos fóruns educacionais. Também não foi analisada a eficiência de classificadores léxicos para solucionar o problema proposto, além de que não estarem trabalhando no contexto de avaliação de sistemas.

Em um trabalho anterior do autor deste TCC (LIMA *et al.*, 2017), foi investigada a classificação automática da polaridade de opinião nas PRUs. Nesse trabalho foram coletadas 1.345 PRUs de um sistema acadêmico com características sociais (por exemplo, comunidades, fóruns, bate-papos, etc.) desde o momento da implantação do sistema (agosto de 2010) até janeiro de 2014. Foram realizadas duas investigações com as PRUs. A primeira usando o classificador léxico *SentiStrength* e segunda aplicando o algoritmo de aprendizado de máquina, o

Naive Bayes.

Avaliando os resultados obtidos na primeira investigação, nota-se que as métricas de avaliação utilizadas apresentaram baixos resultados para as taxas de acerto. Pode-se atribuir este resultado ao fato do dicionário disponibilizado para a língua portuguesa não estar adaptado a reconhecer precisamente a polaridade em PRUs e do seu domínio do sistema (LIMA *et al.*, 2017). Um exemplo que mostra a necessidade de inserir-se um contexto no dicionário seria a sentença "O Google Chrome é bem melhor". Neste exemplo, pelo contexto do sistema, o qual funciona somente no navegador Firefox, essa sentença seria classificada como sentimento negativo, fazendo uma crítica a limitação do sistema de não ser suportado por outros navegadores. Entretanto, a ferramenta por falta de informação, classificaria como um sentimento positivo.

Na segunda investigação pode-se observar que o classificador Naive Bayes obteve um melhor resultado para todas as métricas avaliadas quando comparado aos resultados obtidos pelo classificação do *SentiStrength*. Todavia, para utilizar o Naive Bayes, é necessário que se tenha uma base de treinamento previamente rotulada para a criação do modelo que será utilizado na classificação, portanto é necessário uma grande base de postagens classificadas para assegurar a confiabilidade do modelo.

Por fim, esse trabalho não fez mais investigações sobre os classificadores utilizados, deixando como trabalhos futuros a possibilidade de novos experimentos serem feitos. Principalmente, é necessário verificar se a adição de palavras que representem o domínio do sistema e o contexto de avaliação de sistemas será relevante para o aprimoramento da classificação do *SentiStrength*.

3.2 Trabalhos que utilizam o SentiStrength

Chalothorn e Ellman (2012) apresentaram um trabalho que compara duas técnicas para analisar as postagens com conteúdo radical em fóruns na Internet. Uma das técnicas analisadas foi o *SentiStrength*. Como conclusão, atestou-se a viabilidade do *SentiStrength* para classificação de sentimento de postagens no ambiente proposto. Todavia, alguns erros de classificação foram cometidos, os quais foram atribuídos pela falta de palavras que representassem o contexto do domínio que estava sendo trabalhado.

No entanto, o *SentiStrength* é reconhecido por se tornar mais limitado quando se trabalha com postagens em idiomas diferentes do inglês, como mostrado por Shalunts *et al.* (2014). Nesse trabalho, foi feita uma expansão de uma metodologia para a análise de

sentimentos de dados retirados de mídias sociais da Alemanha sobre desastres naturais. Ao final dos experimentos, foi desenvolvida uma ferramenta derivada do *SentiStrength* que é própria para resolver o problema de classificação proposto pelo trabalho. Entre algumas das adições feitas, pode-se citar: especificação do domínio para notícias gerais e desastres naturais, melhoramento do suporte para outras línguas e alteração no esquema de classificação. No entanto, tal ferramenta não tem suporte para a língua portuguesa, tornando inviável seu uso para este trabalho.

O resultado de Chalothorn e Ellman (2012) motiva a utilização do *SentiStrength* para a classificação de PRUs. No entanto, é necessário confeccionar um dicionário próprio para aquele domínio para que a ferramenta possa entender que palavras são positivas e negativas baseadas no contexto que aquelas postagens estão inseridas. Ou seja, é afirmada a necessidade de enriquecer o dicionário com novas palavras relevantes.

Em Shalunts *et al.* (2014), também se percebe a importância do domínio para uma melhor classificação de sentimentos, sendo necessária a inserção de modificações em certas etapas envolvidas no processo do *SentiStrength*. No caso deste trabalho, a modificação será feita no próprio dicionário utilizado, enriquecendo-o com novas palavras pertencentes ao domínio do sistema.

Deve-se notar, também, a importância do idioma para a eficiência da classificação. Como este trabalho está sendo feito com postagens em português (idioma ao qual não foi utilizado em nenhum dos trabalhos descritos anteriormente), um cuidado deve ser tomado quanto ao dicionário, o qual é o principal responsável para classificar uma palavra como positiva ou negativa.

3.3 Trabalhos que utilizam a métrica TF-IDF

Como descrito na seção 2.3.1, o TF-IDF é uma métrica matemática que pode ser facilmente implementada dentro de um algoritmo, sendo possível aprimorar conforme a necessidade do trabalho. Tal afirmação pode ser vista em Neto *et al.* (2000), que utilizaram uma variação da métrica TF-IDF, que auxiliou na implementação de um algoritmo de mineração de dados baseado em clusterização para confeccionar de resumo de textos. Essa variação foi feita trocando a quantidade de documentos pela quantidade de sentenças na fórmula apresentada na seção 2.3.1. Essa implementação provou-se, além de rápida, ser aplicável em textos cujo o idioma não é o inglês, uma vez que a métrica aplicada não necessita de informação semântica das palavras.

Em um outro trabalho Trstenjak *et al.* (2014) apresentaram um *framework* para classificação de texto utilizando o algoritmo *K-Nearest Neighbor*, juntamente com a métrica TF-IDF. Ambas implementações são simples para aprendizado de máquina e classificação textual, respectivamente. Esse *framework* teve como objetivo classificar a categoria de um documento, sendo elas: "Esportes, Politica, Finanças e Notícias Diárias", levando em consideração também a velocidade dessa classificação. Os testes feitos indicaram que esse *framework* se mostrou estável e confiável, obtendo bons resultados independente do valor de *K*.

O trabalho de Neto *et al.* (2000) motiva a utilização do TF-IDF para aquisição de palavras mais relevantes para identificação de conjuntos textuais. Essa capacidade da métrica de trabalhar com textos e obter resultados satisfatórios motivam sua aplicação para o problema de enriquecimento de dicionários, graças a versatilidade de implementação da métrica. Além disso, como as postagens estão escritas em português, é interessante utilizar uma métrica que não necessita de uma avaliação semântica, pois quando se trabalha em ambientes digitais, os erros gramaticais e os vícios de linguagem são muito presentes, assim dificultando essa avaliação.

Com esses trabalhos (NETO *et al.*, 2000; TRSTENJAK *et al.*, 2014), percebe-se a versatilidade que o TF-IDF tem para ser aplicado juntamente com outras técnicas de classificação. Isso motiva a investigação para aprimorar os dicionários utilizados pelo *SentiStrength* utilizando a métrica TF-IDF, já demonstrada eficiente para classificação textual. Nota-se, também, a importância de um bom pré-processamento para a qualidade final da classificação, uma vez que a quantidade de palavras não usadas era um fator determinante para o resultado da classificação. Por fim, em ambos os trabalhos é atestada a facilidade de modificar o algoritmo que implementa o TF-IDF, graças a sua simplicidade, característica indispensável para adaptação ao problema apresentado neste trabalho.

No entanto, esses trabalhos (NETO *et al.*, 2000; TRSTENJAK *et al.*, 2014) não investigaram o contexto de avaliação de sistemas e de análise de sentimentos propostos por este trabalho. As etapas de coleta de dados e de pré-processamento também serão diferentes das aplicadas por eles, uma vez que serão coletadas postagens de usuários de um sistema específico e será necessário adaptar essas postagens para o programa *SentiStrength*, bem como para utilizá-las em conjunto com métrica TF-IDF para realizar a investigação proposta.

3.4 Tabela-Resumo

A Tabela 3 representa as principais características deste trabalho e dos trabalhos relacionados abordados anteriormente. O principal diferencial deste trabalho é proporcionar um resultado mais preciso no contexto de avaliação de sistemas utilizando postagens espontâneas. Para isso, é utilizando palavras relacionadas ao seu próprio domínio, retiradas com a métrica TF-IDF, para aprimorar os resultados do classificador léxico *SentiStrength*, ao mesmo tempo que investiga-se a eficiência dessa ferramenta para classificar postagens escritas em português.

Tabela 3 – Resumo dos trabalhos relacionados

Trabalho	**Técnica utilizada**	**Base de Dados Utilizada**	**Análise de sentimentos**	**Algoritmos/ Ferramentas**
Rolim et al. (2016)	Aprendizado de máquina	Fóruns educacionais	Não	Rede bayesiana Árvore de decisão Rede Neural
Lima et al. (2017)	Aprendizado de máquina Classificação léxica	Sistemas sociais	Sim	Naive Bayes SentiStrength
Chalothorn e Ellman (2012)	Classificação léxica	Fóruns de discussão	Sim	SentiStrength
Shalunts et al. (2014)	Classificação léxica	Mídias sociais	Sim	SentiStrength
Neto et al. (2000)	Aprendizado de máquina	Textos diversos	Não	Algoritmos de clusterização
Trstenjak et al. (2014)	Aprendizado de máquina	Textos diversos	Não	KNN
Este trabalho	Classificação léxica	Resenhas de aplicativos de mobilidade urbana	Sim	SentiStrength

Fonte: Elaborado pelo Autor

4 INVESTIGAÇÃO

Este capítulo descreve todas as etapas da investigação seguidas para verificar a eficiência da métrica TF-IDF para seleção de palavras que irão enriquecer o dicionário do *SentiStrength*, objetivando melhorar a classificação de sentimentos. A investigação foi executada seguindo as seguintes etapas: 1) extração dos dados; 2) classificação dos dados; 3) preparação dos dados; 4) aplicação do TF-IDF; 5) aplicação no *SentiStrength* e no Capítulo 5 é feita a análise dos resultados.

4.1 Extração dos Dados

A investigação foi realizada utilizando PRUs escritas em português, extraídas da seção de resenhas da Play Store, de três aplicativos de navegação por satélite: Google Maps, Waze e GPS Brasil. Neles são publicadas novas resenhas diariamente. As postagens foram extraídas durante o período de 1 mês, iniciando o período de coleta em 11 de Dezembro de 2017 e finalizando-o em 9 de janeiro de 2018, utilizando um *framework* desenvolvido em Python para extrair dados estruturados de páginas Web chamado Scrapy [1]. Ao fim deste período, obteve-se 1286 postagens.

No entanto, 587 postagens repetidas foram retiradas, para não enviesar o resultado desta investigação. As postagens também foram seccionadas em sentenças, pois em uma postagem pode haver mais de um sentimento, como no exemplo a seguir: *"Está bom, mas está pior do que antes!"*. Com isso, além de aumentar a quantidade e diversidade da base de postagens, também nos garante uma amostragem mais fiel para a análise de sentimentos, uma vez que em uma postagem o usuário pode ser tanto positiva como negativa. Após a segmentação, foram obtidas 1450 sentenças. Também foram utilizadas 658 sentenças coletadas por um outro grupo de pesquisadores (SILVA *et al.*, 2017) do mesmo projeto de pesquisa, as quais pertencem ao mesmo domínio. Ao final, foram utilizadas 2108 sentenças para a realização deste experimento.

4.2 Classificação dos Dados

Obtida a base de sentenças, realizou-se uma classificação manual de sentimento, com o objetivo de comparar a classificação realizada neste artigo com os resultados que serão obtidos pelo *SentiStrength*, assim, são utilizadas métricas para medir a eficiência entre o dicionário

[1] Scrapy. Disponível em: https://docs.scrapy.org/en/latest/. Acesso em 25 de Novembro de 2018

original e o dicionário enriquecido com novas palavras. A classificação manual foi feita por duas pessoas e validada por uma terceira, respectivamente, o autor deste trabalho, uma participante do projeto de pesquisa e a orientadora deste trabalho. Com isso, aumenta-se a confiabilidade do sentimento atribuído a cada sentença durante o período de 1 mês. Com essa classificação, das 2108 sentenças obtidas, constatou-se a presença de 880 sentenças positivas, 525 sentenças neutras e 703 sentenças negativas.

4.3 Preparação dos Dados

Com a base devidamente classificada, iniciou-se a etapa de pré-processamento. Para realização desta etapa, foi implementado um programa escrito em Python, utilizando uma biblioteca para processamento da linguagem natural, a Natural Language Toolkit (NLTK) (BIRD; LOPER, 2004). Foi feita a remoção de caracteres especiais, *emoticons*, sinais de pontuação e de acentuação. Além disso, toda a base foi transformada para caixa-baixa, facilitando as futuras etapas deste trabalho, uma vez que o programa *SentiStrength* necessita que todas as sentenças estejam nesse formato.

Também foram feitos dois pré-processamento separados em outras duas bases de dados iguais, um pré-processamento para cada uma dessas bases. No primeiro foi aplicado o processo de *stemming*, transformando as palavras desta base de dados para a sua forma raiz, utilizando o algoritmo da biblioteca NLTK (ORENGO; HUYCK, 2001). O segundo foi aplicado o processo de lematização, para isso, um algoritmo de lematização criado por outro membro do projeto foi utilizado nesta base de dados. Então, foi possível verificar se a aplicação dessas técnicas irão influenciar nas palavras mais relevantes, valoradas pelo TF-IDF. Portanto, são utilizadas três bases de dados no experimento: base com palavras originais (BO); base com *stemming* (BS) e Base lematizada (BL).

4.4 Enriquecimento do Léxico

Para o cálculo da medida estatística TF-IDF, são necessários múltiplos documentos, como explicado na seção 2.3.1, então foram criados três documentos, onde o primeiro contém somente sentenças positivas, o segundo somente sentenças neutras e o terceiro somente sentenças negativas. Com isso, dividiu-se cada documento, onde 75% (1581 sentenças) foi utilizado como treinamento e 25% (525 sentenças) foi utilizado como treino, a escolha de sentenças foi feita

de forma aleatória, utilizando um algoritmo desenvolvido pelo autor. Fazendo isso, além de evitar a tendenciosidade do resultado com a divisão da base em treinamento e teste, garantiu uma quantidade proporcional de sentenças positivas, negativas e neutras.

Assim, utilizando a base de treinamento separada em três documentos, e com o auxílio de um *script* desenvolvido em Python, utilizando a biblioteca para tarefas de processamento da linguagem natural chamada *TextBlob* (LORIA *et al.*, 2014), calculou-se as palavras mais relevantes segundo a métrica TF-IDF. Gerando duas listas, uma formada por 40 palavras positivas e outra formada por 40 palavras negativas, com o valor atribuído pelo TF-IDF ainda atrelado a cada palavra. Essa quantidade foi escolhida baseada no intervalo de valor TF-IDF estipulado pelo autor. Essa situação ocorreu para as três bases utilizadas no experimento (BO, BS e BL).

4.5 Aplicação no *SentiStrength*

Como explicado nas seção 2.6.1, o *SentiStrength* retorna as maiores forças positivas e negativas de uma sentença e, a partir desta comparação, pode-se inferir se uma sentença é positiva, negativa ou neutra. Por exemplo, na frase: *"O sistema era bom, mas agora está horrível!"*, supondo que *"bom"* tenha força 3 e que *"horrível"* tenha força -5, assim as forças positivas e negativas dessa sentença serão (3,-5), portanto essa sentença será classificada como negativa. A relação pode ser vista na equação 4.1, que reflete essa comparação entre a força positiva e a força negativa para detectar a polaridade de uma sentença.

$$\begin{cases} f.positiva > f.negativa & Positivo \\ f.positiva < f.negativa & Negativo \\ f.positiva = f.negativa & Neutro \end{cases} \tag{4.1}$$

Sabe-se que o programa *SentiStrength* valoriza suas palavras com valores entre 1 e 5, para palavras positivas, e entre -1 e -5, para palavras negativas, portanto é necessário atribuir um valor às novas palavras seguindo essa regra. Para isso, foi utilizado o valor atribuído em cada palavra pelo TF-IDF. Com ele, pode-se definir um padrão onde para a palavra com o maior valor TF-IDF será atribuida a nota 5 ou -5, dependendo da polaridade da palavra, a qual é definida pelo documento ao qual ela foi retirada. Com isso, podemos definir uma escala onde o maior valor X atribuído pelo TF-IDF terá força 5 ou -5, e os demais valores Y menores do que X terão forças proporcionais ao valor de X. Um exemplo seria o maior valor TF-IDF de um conjunto

de palavras ser 2, então essa palavra com valor 2 será atribuída a força 5, e uma outra palavra com valor TF-IDF igual a 1, será atribuída a força 3, proporcionalmente. A medida que o valor TF-IDF for diminuindo comparado ao maior valor dentro do conjunto de valores, a atribuição também seguirá esse padrão até o mínimo de 2 ou -2, uma vez que os valores 1 e -1, para o *SentiStrength*, significa a ausência de positividade e a ausência de negatividade, respectivamente. Assim, criando essa relação palavra/força, foi possível adicionar as novas palavras ao dicionário usado pelo programa. Para isso, copiou-se o documento de texto com o conjunto de palavras (dicionário) utilizado pelo *SentiStrength* três vezes, uma para cada base (BO, BS, BL), e em cada um deles foi adicionado seus respectivos conjuntos de 80 palavras (40 positivas e 40 negativas). Com isso, teremos três dicionários diferentes enriquecidos com palavras novas.

Todo esse procedimento descrito anteriormente foi realizado 10 vezes, sendo geradas bases de treinamento e de teste diferentes, mas que ainda se originaram da base de dados obtidas na seção 4.3. Com isso, pode-se garantir a validade dos resultados, constatando a existência de um padrão nos valores obtidos.

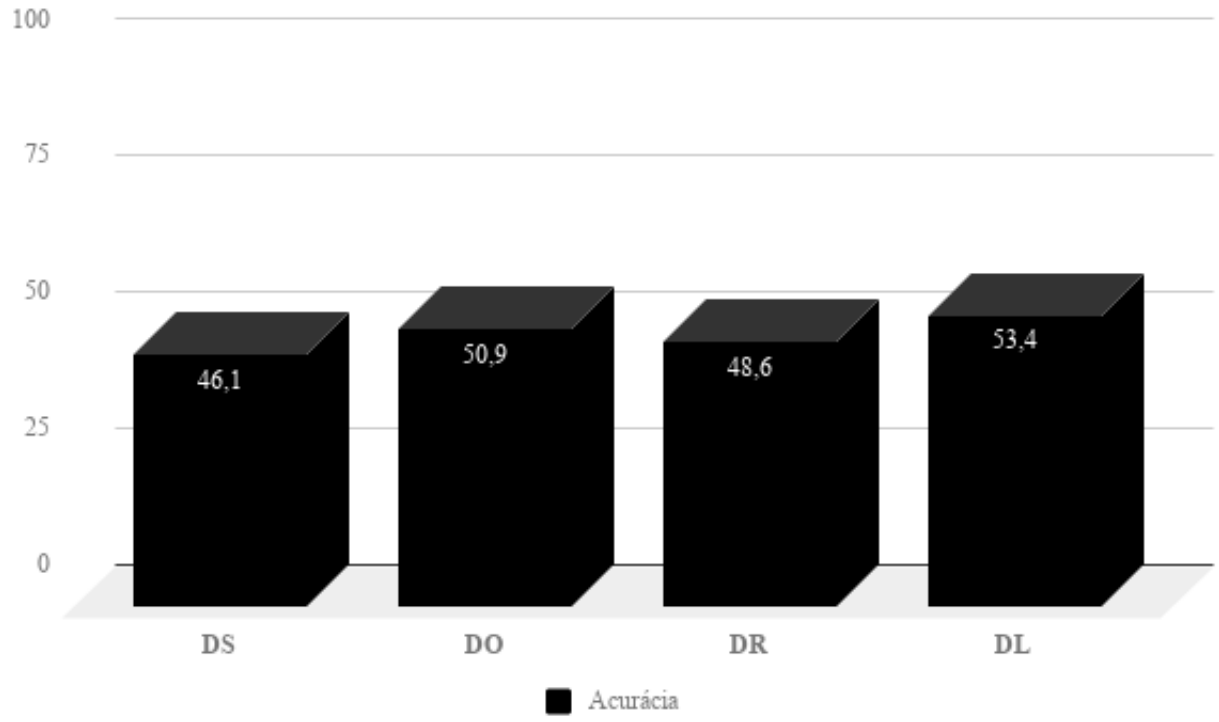
5 RESULTADOS

Este capítulo é focado em apresentar os resultados obtidos com a investigação descrita no Capítulo 4. São apresentados resultados obtidos com quatro dicionários: o dicionário do *SentiStrength* sem adição de palavras (DS) , o dicionário enriquecido com palavras na sua forma original (DO), o dicionário enriquecido com palavras na forma raiz (DR) e o dicionário enriquecido com palavras lematizadas (DL). São mostrados os valores obtidos utilizando a média das 10 vezes que a investigação foi realizada. Também são apresentadas algumas palavras que foram adicionadas em cada um dos dicionários.

5.1 Métricas utilizadas

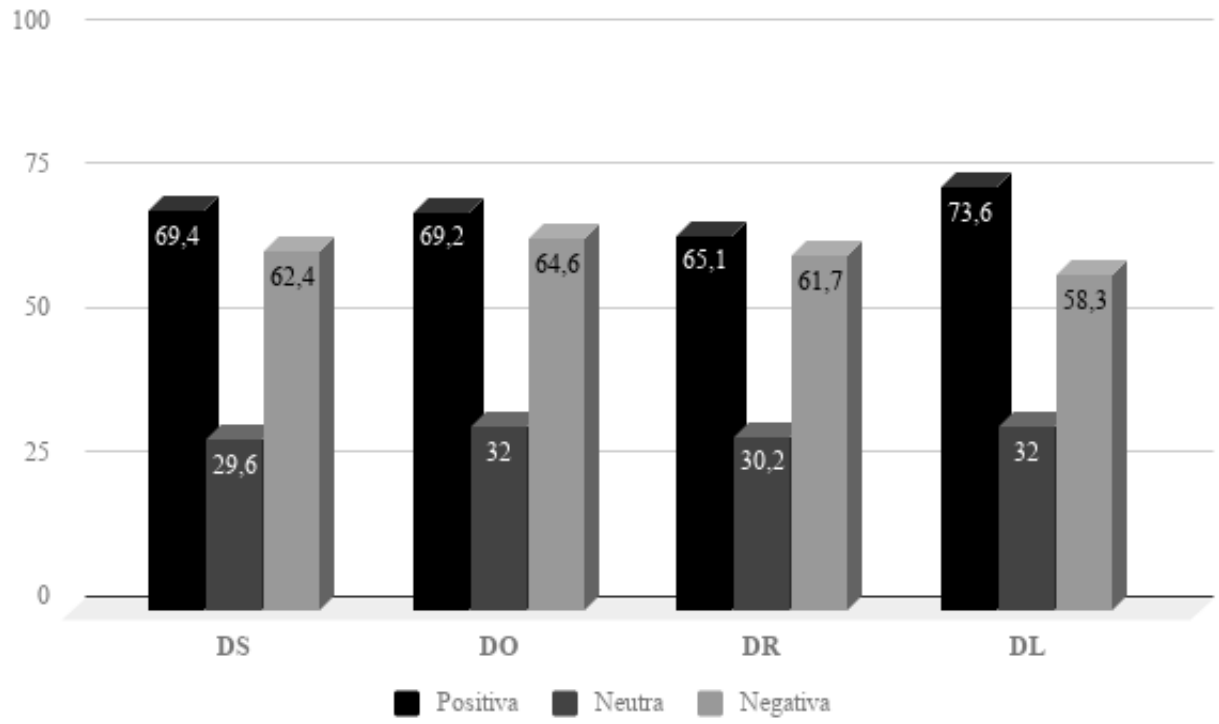
O Gráfico 1 representa a acuŕacia, o Gráfico 2 representa a precisão, o Gráfico 3 representa o *recall* e o Gráfico 4 representa o *F-measure*.

Gráfico 1 – Média da acurácia(%) obtida para cada um dos dicionários



Fonte: Elaborado pelo Autor

Gráfico 2 – Média de precisão (%) obtida para cada um dos dicionários



Fonte: Elaborado pelo Autor

5.2 Palavras Adicionadas

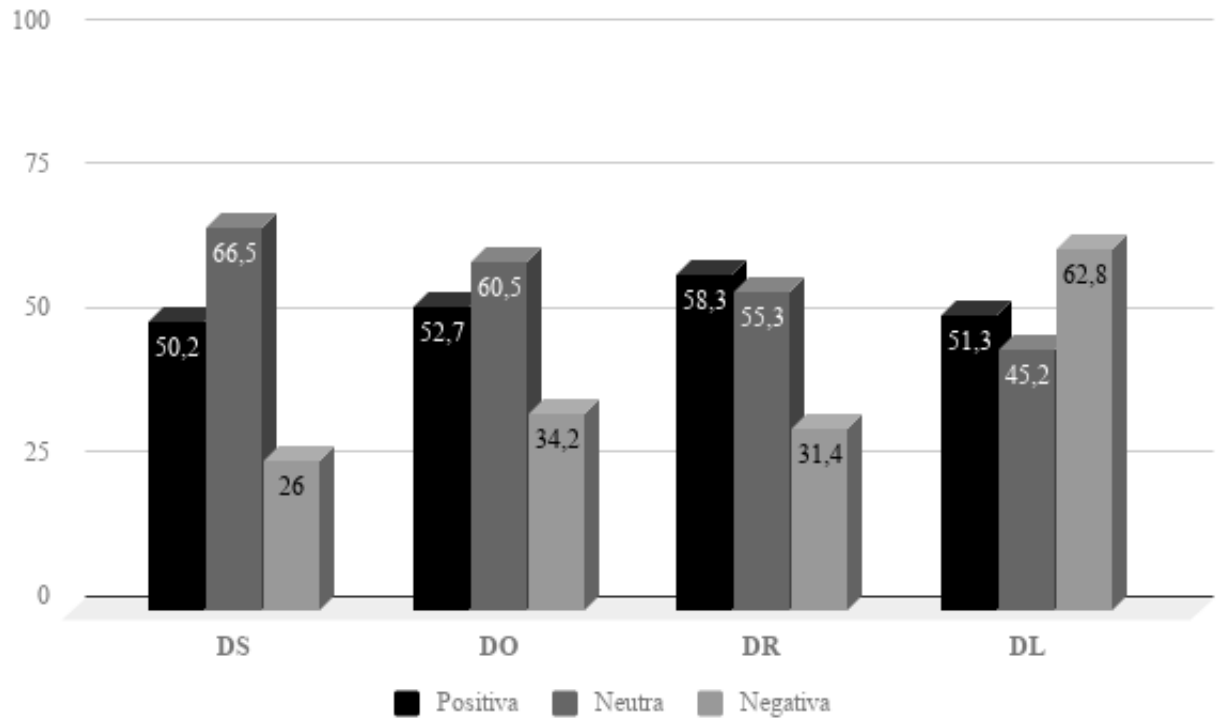
A Tabela 4 mostra algumas das palavras, tanto negativas como positivas, que foram adicionadas aos dicionários utilizando a métrica TF-IDF. Como o *SentiStrength* não atribui valores para palavras neutras, não foi adicionada nenhuma palavra neutra nos dicionários. Uma discussão mais detalhada sobre elas será feita no Capítulo 6.

Tabela 4 – Parte das palavras adicionadas a cada dicionário

Dicionário	Positiva	Negativa
DO	maravilhoso, sensacional, cumpre, amei, merece, amo, adoro, diariamente, certinho, inteligente, incomparavel	horrivel, pessimo, porcaria, lixo, travando, cai, bateria, pior, perdeu, arrumem, favela, troquei, pessima
DR	maravilh*, cumpr*, fantas*, intelig*, viaj*, sensac*, aprov*, recomend*, ame*, amo*, imbat*, exel*	horri*, porc*, atras*, lix*, bat*, cai*, infeliz*, prejudic*, consom*, irrit*, pior*, troq*, demor*
DL	maravilhoso, amar, cumprir, sensacional, exelente, diariamente, imbativel, satisfacao, fantastico, eficaz	horrivel, porcaria, prejudicar, lixar, piorar, consumir, decepcao, bugado, dificultar, chato, contramao

Fonte: Elaborado pelo Autor

Gráfico 3 – Média de *recall* (%) obtida para cada um dos dicionários



Fonte: Elaborado pelo Autor

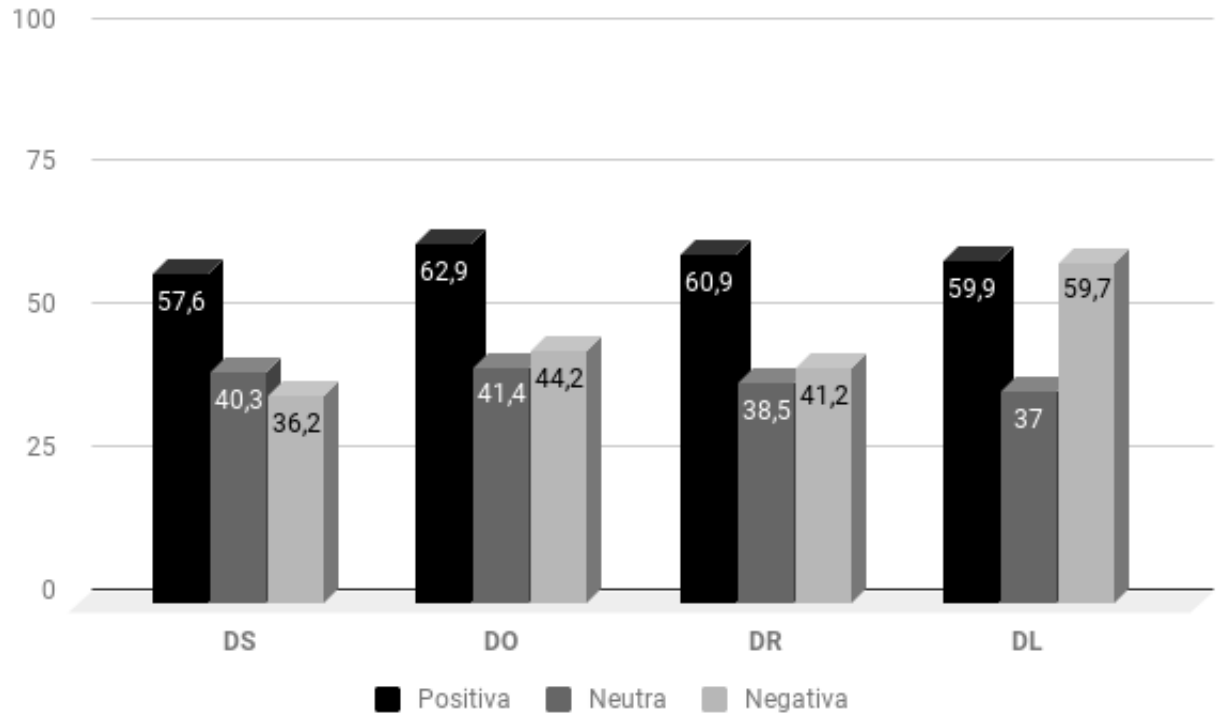
5.3 Análise de Resultados

A primeira métrica utilizada para comparar a eficiência entre os dicionários foi a acurácia. Como pode ser visto no Gráfico 1, a acurácia de todos os dicionários enriquecidos foi melhor quando comparado ao dicionário original. Destaca-se, no entanto, o experimento feito com o dicionário enriquecido com palavras lematizadas, obtendo um aumento na acurácia de 7,3%, comparando-o ao dicionário original.

Todavia, somente a acurácia não é suficiente para analisar os pontos fortes e fracos dos experimentos feitos. É necessário saber qual é a capacidade de classificar corretamente cada polaridade em cada um dos dicionários. Com isso, será possível saber qual dicionário é melhor para classificar cada uma das 3 polaridades possíveis: positiva, neutra e negativa.

Observa-se, no Gráfico 2, que a precisão para classificar sentenças positivas se manteve similar entre todos os dicionários, onde o crescimento mais significativo ocorreu com a utilização do dicionário lematizado. Para a classificação de sentenças neutras, percebe-se um

Gráfico 4 – Média de *F-Measure* (%) obtida para cada um dos dicionários



Fonte: Elaborado pelo Autor

aumento nessa precisão ao utilizar os dicionários enriquecidos, porém os valores permaneceram inferiores a 35%. Por fim, a precisão para as sentenças negativas também se manteve similar entre todos os experimentos. Dessa vez, o dicionário com o melhor resultado foi o enriquecido com palavras na sua forma original, e o resultado menos satisfatório, sendo inferior até mesmo comparado-o com o dicionário sem enriquecimento, foi o dicionário que utilizou palavras lematizadas.

Observando o Gráfico 3, percebe-se alguns dos resultados mais interessante dos experimentos feitos. Embora todos os dicionários enriquecidos superaram o dicionário não enriquecido do *SentiStrength*, o melhor *recall* para sentenças positivas foi obtido utilizando o dicionário enriquecido com palavras em sua raiz, obtendo um aumento de 8% comparado ao dicionário original. Já o *recall* para sentenças neutras, percebe-se uma tendência entre todos os experimentos. Enquanto o *recall* das sentenças positivas e negativas aumentam para cada experimento, o *recall* para as sentenças neutras diminuem. Essa situação pode ser atribuída, mais uma vez, a natureza do *SentiStrength*, focada em classificar as polaridades positivas e negativas, havendo perdas para classificar sentenças neutras. Finalmente, ao observar o *recall* para sentenças negativas, percebe-se o resultado observável mais significativo deste trabalho,

onde o dicionário enriquecido com palavras lematizadas teve um aumento de 36,8% comparado ao dicionário original. Além disso, comparando-o aos outros dicionários enriquecidos, houve um aumento superior a 28%. Ou seja, para classificação de sentenças realmente negativas, o dicionário com palavras lematizadas é obviamente o mais indicado.

Analisando o Gráfico 4, é observada a relação entre a precisão e o *recall* para cada polaridade. O maior *F-measure* para a polaridade positiva foi obtida pelo dicionário enriquecido por palavras na sua forma normal, assim como para a polaridade neutra. Para a polaridade negativa, o dicionário enriquecido com palavras lematizadas superou significativamente os resultados dos outros dicionários. A métrica *F-measure* pode ser a mais confiável para decidir que técnica de enriquecimento é a mais indicada para aplicar em cada polaridade, já que ela representa uma forma de média entre os resultados obtidos com a precisão e o *recall*. No caso deste trabalho, enriquecer o dicionário com palavras na sua forma original irá melhorar a classificação de sentenças positivas, enquanto que para sentenças negativas, é mais interessante o uso de palavras lematizadas.

6 DISCUSSÃO

Neste capítulo, apresenta-se uma discussão dos resultados obtidos no experimento e de fatores que poderiam, de alguma forma, enviesar ou melhorar os resultados. Liste primeiramente os pontos a serem discutidos: 1) quanto às bases utilizadas; 2) quanto às polaridades; 3) quanto ao tamanho da base; 4) quanto ao domínio das postagens; 5) quanto ao desempenho do classificador.

Quanto às bases utilizadas, observando a Tabela 4, pode-se obter informações interessantes sobre elas. Primeiramente, pode-se notar algumas palavras que são comuns em todos os dicionários, como: maravilhoso, sensacional, horrível, porcaria, lixo, entre outros. A repetição dessas palavras em todos os dicionários refletem nos resultados mostrados pelos gráficos da seção 5, onde todos os dicionários enriquecidos apresentaram resultados semelhantes, principalmente para a classificação de sentenças positivas. Percebe-se também, observando o dicionário enriquecido com palavras em sua raiz, que a implementação utilizada se provou ineficaz em alguns casos, como por exemplo as palavras: ame* e amo*. Tal erro pode ser atribuído a dificuldade de implementação de algoritmos que fazem a transformação de palavras para a sua forma raiz, principalmente quando se trabalha com uma linguagem tão complexa como o português. No entanto, pode-se observar que tal erro não ocorreu no dicionário lematizado, onde a palavra "*amar*"passou a representar propriamente todas as suas palavras derivadas, como: amei, amo, entre outras.

Quanto às polaridades, enriquecer o dicionário com palavras na forma original irá melhorar a classificação de sentenças positivas, enquanto que para sentenças negativas, é mais interessante o uso de palavras lematizadas. Isso mostra que uma solução híbrida de técnicas também é viável para aprimorar a eficiência do *SentiStrength*. Para a polaridade neutra, nota-se um declínio no número de acertos para cada dicionário enriquecido. Isso pode ter ocorrido pelo foco do *SentiStrength* em classificar as sentenças baseadas em palavras positivas e negativas, sendo classificada como neutra somente quando não há nas sentenças palavras presentes no dicionário, ou quando a forças positivas e negativas são iguais. Sobre as palavras neutras, elas são essenciais para descobrir justamente quais palavras não tem influencia nenhuma para a classificação de polaridade. Sem as sentenças neutras, não seria possível aplicar o TF-IDF de forma satisfatória, uma vez que é preciso conhecer quais palavras são comuns entre as sentenças neutras e as sentenças positivas e negativas. Pode-se também utilizar as palavras neutras capturadas com o TF-IDF para outras finalidades além da classificação de sentimentos.

Um exemplo seria o reconhecimento de dúvidas, que também serviria para o aprimoramento do sistema.

Quanto ao domínio das postagens, percebe-se que foram incluídas palavras que retratam o domínio das sentenças utilizadas, ao qual se refere a aplicativos de navegação por satélite. Palavras como "*bugado, contramao, favela, travando, bateria*"entre outras refletem tanto a relação delas com um sistema (bugado, travando, bateria), quanto a relação com os aplicativos de navegação por satélite, demonstrado pelas palavras negativas "*favela*"e "*contramao*", indicando erros cometidos pela aplicação. Vale ressaltar, também, que o conjunto de palavras utilizadas para enriquecer cada dicionário será diferente para cada domínio. Por isso, é estimulado que a investigação seja refeita utilizando sentenças de outros domínios para averiguar quais palavras serão diferentes e quais palavras serão semelhantes para então definir um padrão entre os dicionários enriquecidos.

Quanto ao tamanho da base, acredita-se que o uso de uma base maior poderia modificar as palavras encontradas, uma vez que haveria mais sentenças e com isso o valor TF-IDF seria alterado para cada palavra. Com isso, novas palavras poderiam ser encontradas e adicionadas ao dicionário, além de haver uma garantia ainda maior da significância dessas palavras para representar as polaridades positivas e negativas. Além disso, realizar novos experimentos com bases maiores também poderá averiguar se palavras na forma original realmente são as melhores para a classificação de sentenças positivas e se as palavras lematizadas são as melhores para sentenças negativas, como foi mostrada na investigação descrita neste artigo.

Finalmente, quanto ao desempenho do classificador, foi observada uma pequena melhoria para a classificação de sentimentos, porém o TF-IDF já se apresenta como uma solução viável para a seleção de palavras de cada polaridade. No entanto, é necessário que seja feito outros estudos que busquem refinar mais ainda essa seleção de palavras, seja aprimorando a métrica TF-IDF utilizando-a em conjunto com outros conceitos de valoração das palavras, seja alterando a técnica de transformação do valor TF-IDF para as forças do *SentiStrength*, seja refazendo a investigação com bases de dados maiores utilizando algoritmos de pré-processamento mais eficientes.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Ao final deste trabalho perceber-se que, embora não tão expressiva, as palavras adicionadas pela métrica TF-IDF melhoraram a classificação de sentimentos feita pelo *SentiStrength*. Para todos os dicionários enriquecidos, foi mostrado um aumento mais significante para a classificação de sentenças positivas e negativas, principalmente para o dicionário enriquecido com palavras lematizadas. Inicialmente conclui-se que, para este experimento, palavras lematizadas são mais eficientes para a classificação de sentimentos. Todavia, observando mais criteriosamente, uma abordagem híbrida também demonstra-se interessante, uma vez que as palavras inalteradas mostraram-se um pouco mais eficientes para classificar sentenças positivas, mas as palavras lematizadas foram melhores para classificar sentenças negativas. No entanto, a classificação de sentenças neutras foram prejudicadas conforme o aumento na eficiência da classificação de sentenças positivas e negativas.

Como trabalhos futuros, ainda é necessário melhorar a eficiência do *SentiStrength* para classificar sentenças em português. Tal objetivo pode ser alcançado estudando outras métricas para obtenção de palavras, podendo até mesmo utilizá-las em conjunto com o TF-IDF. O problema para classificar sentenças neutras também deve ser abordado, uma vez que elas são úteis para avaliação de sistemas. Para isso, podem ser abordadas novas maneiras de valorar as forças atribuídas às palavras utilizadas pelo *SentiStrength*. Por fim, trabalhar com outros domínios e com quantidades cada vez maiores de sentenças, fazendo com que os resultados sejam mais concretos e confiáveis.

REFERÊNCIAS

ALVARES, R. V. **Investigação do processo de Stemming na língua portuguesa**. Tese (Doutorado) — Universidade Federal Fluminense, 2005.

AMARAL, F. **Aprenda Mineração de dados:** teoria e prática. Rio de Janeiro: Alta Books Editora, 2016. v. 1.

ARAÚJO, M.; GONÇALVES, P.; BENEVENUTO, F.; CHA, M. Métodos para análise de sentimentos no twitter. In: **Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)**. Salvador: ACM New York, 2013.

BIRD, S.; LOPER, E. Nltk: the natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL 2004 on Interactive poster and demonstration sessions**. Barcelona, 2004. p. 31.

CHALOTHORN, T.; ELLMAN, J. Sentiment analysis of web forums: Comparison between sentiwordnet and sentistrength. In: THE 4TH INTERNATIONAL CONFERENCE ON COMPUTER TECHNOLOGY AND DEVELOPMENT (ICCTD 2012). 24-25 NOVEMBER 2012. Bangkok, 2012.

ESULI, A.; SEBASTIANI, F. Sentiwordnet: a high-coverage lexical resource for opinion mining. **Evaluation**, v. 17, p. 1–26, 2007.

FETTER, M.; SCHIRMER, M.; GROSS, T. Caessa: visual authoring of context-aware experience sampling studies. In: ACM. **CHI'11 Extended Abstracts on Human Factors in Computing Systems**. Vancouver, 2011. p. 2341–2346.

HEDEGAARD, S.; SIMONSEN, J. G. Extracting usability and user experience information from online user reviews. In: ACM. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. Paris, 2013. p. 2089–2098.

IMAMURA, C. Y.-M. **Pré-processamento para extraçao de conhecimento de bases textuais**. Tese (Doutorado) — Universidade de São Paulo, 2001.

KORHONEN, H.; ARRASVUORI, J.; VÄÄNÄNEN-VAINIO-MATTILA, K. Let users tell the story: evaluating user experience with experience reports. In: ACM. **CHI'10 Extended Abstracts on Human Factors in Computing Systems**. Atlanta, 2010. p. 4051–4056.

LIMA, A. M.; SILVA, P. B.; CRUZ, L. A.; MENDES, M. S. Investigating the polarity of user postings in a social system. In: SPRINGER. **International Conference on Social Computing and Social Media**. Vancouver, 2017. p. 246–257.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LORIA, S.; KEEN, P.; HONNIBAL, M.; YANKOVSKY, R.; KARESH, D.; DEMPSEY, E. *et al.* Textblob: simplified text processing. **Secondary TextBlob: Simplified Text Processing**, 2014.

LOVINS, J. B. Development of a stemming algorithm. **Mech. Translat. & Comp. Linguistics**, v. 11, n. 1-2, p. 22–31, 1968.

LUCCA, J. D.; NUNES, M. d. G. V. Lematização versus stemming. **USP, UFSCar, UNESP, São Carlos, São Paulo**, 2002.

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.

MENDES, M. S. **Model for evaluation of interaction in social systems from the Users Textual Language**. Tese (Doutorado) — Universidade Federal do Ceará, Fortaleza, CE - Brasil, 2015.

MENDES, M. S.; FURTADO, E. An experience of textual evaluation using the maltu methodology. In: SPRINGER. **International Conference on Social Computing and Social Media**. Las Vegas, 2018. p. 236–246.

MENDES, M. S.; FURTADO, E.; FURTADO, V.; CASTRO, M. F. de. How do users express their emotions regarding the social system in use? a classification of their postings by using the emotional analysis of norman. In: SPRINGER. **International Conference on Social Computing and Social Media**. Heraklion, 2014. p. 229–241.

MENDES, M. S.; FURTADO, E.; FURTADO, V.; CASTRO, M. F. de. Investigating usability and user experience from the user postings in social systems. In: SPRINGER. **International Conference on Social Computing and Social Media**. Los Angeles, 2015. p. 216–228.

MENDES, M. S.; FURTADO, E. S. Uux-posts: a tool for extracting and classifying postings related to the use of a system. In: ACM. **Proceedings of the 8th Latin American Conference on Human-Computer Interaction**. Guatemala, 2017. p. 2.

MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995.

NETO, J. L.; SANTOS, A. D.; KAESTNER, C. A.; ALEXANDRE, N.; SANTOS, D. *et al.* Document clustering and text summarization. Citeseer, 2000.

ORENGO, V.; HUYCK, C. A stemming algorithmm for the portuguese language. In: IEEE. **spire**. Laguna de San Rafael, 2001. p. 0186.

PARTALA, T.; KALLINEN, A. Understanding the most satisfying and unsatisfying user experiences: Emotions, psychological needs, and context. **Interacting with computers**, Oxford University Press Oxford, UK, v. 24, n. 1, p. 25–34, 2011.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011.

RAMOS, J. *et al.* Using tf-idf to determine word relevance in document queries. In: **Proceedings of the first instructional conference on machine learning**. Piscataway: CS Rutgers, 2003. v. 242, p. 133–142.

ROLIM, V.; FERREIRA, R.; COSTA, E. Identificação automática de dúvidas em fóruns educacionais. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. Uberlândia: Sociedade Brasileira de Computação, 2016. v. 27, n. 1, p. 936.

SHALUNTS, G.; BACKFRIED, G.; PRINZ, P. Sentiment analysis of german social media data for natural disasters. In: **ISCRAM**. Pennsylvania: The Pennsylvania State University, 2014.

SILBERSCHATZ, A.; SUNDARSHAN, S.; KORTH, H. F. **Sistema de banco de dados**. Rio de Janeiro: Elsevier Brasil, 2016.

SILVA, C. F. da. **Grupos gramaticais e sintáticos em categorização automática com Support Vector Machines**. Tese (Doutorado) — Universidade do Vale do Rio dos Sinos, São Leopoldo, RS - Brasil, 2004.

SILVA, T. H. O. da; FREITAS, L. M.; MENDES, M. S. Beyond traditional evaluations: user's view in app stores. In: ACM. **Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems**. Joinville, 2017. p. 15.

SOUSA, R. C. C. d. **Identificando sentimentos de texto em português com o Sentiwordnet traduzido**. 2016. Monografia (Bacharel em Ciéncia da Computação), UFC (Universidade Federal do Ceará), Quixadá, Brasil.

THELWALL, M. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In: **Cyberemotions**. Warsaw: Springer, 2017. p. 119–134.

THELWALL, M.; BUCKLEY, K.; PALTOGLOU, G. **SentiStrength. Available online**. 2011.

TRSTENJAK, B.; MIKAC, S.; DONKO, D. Knn with tf-idf based framework for text categorization. **Procedia Engineering**, Elsevier, v. 69, p. 1356–1364, 2014.

VILARES, D.; THELWALL, M.; ALONSO, M. A. The megaphone of the people? spanish sentistrength for real-time analysis of political tweets. **Journal of Information Science**, Sage Publications Sage UK: London, England, v. 41, n. 6, p. 799–813, 2015.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. **Computational linguistics**, MIT Press, v. 35, n. 3, p. 399–433, 2009.