UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM ENGENHARIA DE SOFTWARE

RANDERSON LESSA MELO

AVALIANDO O DICIONÁRIO EM PORTUGUÊS DO MÉTODO DE ANÁLISE DE SENTIMENTOS SENTISTRENGTH

QUIXADÁ

2017

RANDERSON LESSA MELO

AVALIANDO O DICIONÁRIO EM PORTUGUÊS DO MÉTODO DE ANÁLISE DE SENTIMENTOS SENTISTRENGTH

Monografia apresentada ao curso de Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia de Software. Área de concentração: Computação.

Orientadora: Prof[a]. Dra. Paulyne Matthews Jucá.

QUIXADÁ

2017

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)

M486a Melo, Randerson Lessa.
Avaliando o dicionário em português do método de análise de sentimentos SentiStrength / Randerson Lessa Melo. – 2017.
53 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2017.
Orientação: Profa. Dra. Paulyne Matthews Jucá.

1. SentiStrength. 2. Twitter (Redes Sociais on-line). I. Título.

CDD 005.1

RANDERSON LESSA MELO

AVALIANDO O DICIONÁRIO EM PORTUGUÊS DO MÉTODO DE ANÁLISE DE SENTIMENTOS SENTISTRENGTH

Monografia apresentada ao curso de Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia de Software. Área de concentração: Computação.

Aprovada em: ___/ ___ / ___.

BANCA EXAMINADORA

_____
Prof[a]. Dr. Paulyne Matthews Jucá (Orientadora)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Arthur de Castro Callado
Universidade Federal do Ceará (UFC)

_____
Prof. Me. Bruno Góis Mateus
Universidade Federal do Ceará (UFC)

A Deus.

Aos meus pais, Paulo Cézar e Rosilande.

A minha namorada Késia Fernandes.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por toda a sabedoria que tem me dado.

Aos meus pais, Paulo Cézar e Rosilande, que me deram todo o apoio necessário para chegar até aqui.

A minha namorada, Késia Fernandes, por todo o apoio, companheirismo e incentivo que me deu.

A minha família, que me incentivou.

A prof[a]. Dr. Paulyne Matthews Jucá, pela excelente orientação.

Aos professores participantes da banca examinadora, Arthur de Castro Callado e Bruno Góis Mateus pelo tempo, pelas valiosas colaborações e sugestões.

Aos demais professores que contribuíram para a minha formação.

Aos colegas da turma, em especial Amarildo Barros, Anderson Gonçalves, Bruno Barreto, Carlos Matheus, Diego Souza, Italos Estilon, Jacques Nier, Luan Lima, Lucas Sales, Lucas Teixeira e Mauro Roberto, pela amizade e pelos bons momentos compartilhados.

“Peçam, e lhes será dado; busquem, e encontrarão; batam, e a porta lhes será aberta.”

(Mateus 7:7)

RESUMO

A análise de sentimentos é uma das técnicas utilizadas para minerar opinião nas redes sociais. Dentre as redes sociais mais utilizadas, está o Twitter, sendo utilizada por cerca de 319 milhões de usuários. Ele é considerado como um *microblog* e permite que usuários, em tempo real, enviem e recebam de seus contatos textos de até 140 caracteres, conhecidos como *tweets*. Esses *tweets* podem conter diversos tipos de informações, como notícias, reclamações, opiniões, e muitos expressam sentimentos. Existem diversos métodos de análise de sentimentos. Dentre esses, o SentiStrength é um método que utiliza um dicionário léxico anotado por seres humanos e melhorado com o uso de Aprendizado de Máquina. O método demonstrou em algumas pesquisas um bom desempenho em análise de sentimento com *tweets* em inglês. O método disponibiliza alguns outros dicionários, incluindo um dicionário em português que foi pouco utilizado em pesquisas. Nesse contexto, este trabalho visa analisar a eficácia do método SentiStrength com o dicionário em português para *tweets* também em português. Para realização da análise da acurácia do método, foram realizadas duas coletas de 1000 *tweets* cada sobre temas diferentes e os resultados foram comparados com uma análise feita manualmente. Após uma sugestão de melhoria no dicionário, foi possível verificar uma melhora nos resultados, chegando à conclusão que a eficácia do método está diretamente relacionada com a generalização do dicionário utilizado.

Palavras-chave: SentiStrength. Twitter (Redes Sociais on-line).

ABSTRACT

The sentiment analysis is one of the techniques used to identify this customer satisfaction in social networks. Among the most used social networks is Twitter, being used by about 319 million users. It is considered a micro-blog and allows users, in real time, to send and receive texts of up to 140 characters, known as tweets, to from their contacts. These tweets can contain various types of information such as news, complaints, opinions, and many express feelings. There are several methods of sentiment analyzing. Among these, SentiStrength is a method that uses a lexical dictionary annotated by humans and improved with the use of Machine Learning. The method has shown in some surveys a good performance in sentiment analysis with tweets in English. The method provides some other dictionaries, including a dictionary in Portuguese that has been little used in searches. In this context, this work aims to analyze the accuracy of the SentiStrength method with the Portuguese dictionary for tweets also in Portuguese. To perform the analysis of the accuracy of the method, two 1000 tweets were collected each on different themes and the results were compared with a manual analysis. After a suggestion of improvement in the dictionary, it was possible to verify an improvement in the results, arriving to the conclusion that the effectiveness of the method is directly related to the generalization of the dictionary used.

Keywords: SentiStrength. Twitter (Social Networking on-line).

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

Como uma das estratégias para identificar a satisfação de seus clientes, as marcas/empresas têm usado as redes sociais virtuais e o seu grande volume de dados para analisar e extrair informações relevantes de opiniões de seus clientes sobre seus produtos e serviços (GOMES, 2013). A análise de sentimentos (AS) é uma das técnicas utilizadas para identificar essa satisfação de clientes em redes sociais, pois permite identificar a opinião geral de usuários das redes sociais sobre algum tema, marca ou produto.

O Twitter[1] é uma das redes sociais mais utilizadas, sendo utilizada por cerca de 319 milhões de usuários. Ele é considerado como um *microblog* (ALEXANDRINO, 2016) e permite que usuários, em tempo real, enviem e recebam de seus contatos textos de até 140 caracteres, conhecidos como *tweets.* Esses *tweets* podem conter diversos tipos de informações, como notícias, reclamações, opiniões, e muitos expressam sentimentos. Nesse contexto, existem muitos trabalhos que coletam esses dados e analisam o seu conteúdo em forma de polaridade (positivo/negativo) de sentimentos. Porém, poucos trabalhos exploram o idioma em português.

A análise de sentimento também pode ser explorada em outras áreas da tecnologia, como na inteligência artificial (IA), no uso de *bots* (abreviação de *robots* que significa robôs) que são softwares concebidos para simular ações humanas, permitindo que sejam capazes de interpretar os sentimentos, em *chatterbots* (junção das palavras *chatter* que significa conversador e *bots*) que são softwares que simulam a conversação humana, permitindo respostas de acordo com os sentimentos.

Existem diversos métodos de análise de sentimentos. Dentre esses, o SentiStrength[2] é um método que utiliza dicionários léxicos, especialista em textos curtos e de baixa qualidade, como os *tweets*, mostrando em algumas pesquisas uma boa precisão com *tweets* em inglês (GUZMAN; AZÓCAR; LI, 2014). O método disponibiliza dicionários em vários idiomas, incluindo um dicionário em português que até o presente momento não foi testado.

Pensando no problema de análise de sentimento em português de textos curtos e de baixa qualidade, como os *tweets*, e no dicionário em português ainda não testado do método SentiStrength que tem mostrado em algumas pesquisas uma boa precisão com *tweets* em inglês, este trabalho tem como objetivo analisar a eficácia do método de análise de

[1] https://twitter.com/
[2] http://sentistrength.wlv.ac.uk/

sentimento SentiStrength com o dicionário em português para textos de *tweets* (curtos e de baixa qualidade) também em português.

Para a realização deste estudo foram analisados um total de 2000 *tweets* em português, que faziam menção aos termos “motorola” e “liga da justiça”. Os resultados demonstraram que o dicionário em português para os dados coletados apresentou uma acurácia de 75,5% em comparação com a classificação realizada manualmente para o termo “motorola” e uma acurácia de 68,6% em comparação com a classificação realizada manualmente para o termo “liga da justiça”. Após uma breve proposta de melhoria para o dicionário, a acurácia para o termo “motorola” passou para 94,5% (uma melhoria de 19%) e a acurácia do termo “liga da justiça” passou para 83,3% (uma melhoria de 14,7%), confirmando o que diz Benevenuto et al. (2015) sobre a eficiência do método estar diretamente relacionada à generalização do dicionário léxico utilizado. Este trabalho também propõe melhorias para o dicionário em português do método SentiStrength.

O restante deste trabalho encontra-se dividido da seguinte maneira: a seção 2 apresenta os trabalhos relacionados. A seção 3 apresenta os conceitos técnicos e teóricos necessários para a realização do trabalho. A seção 4 apresenta o desenvolvimento do trabalho. A seção 5 apresenta os resultados obtidos. A seção 6 apresenta a conclusão. Por fim, a seção 7 apresenta os trabalhos futuros.

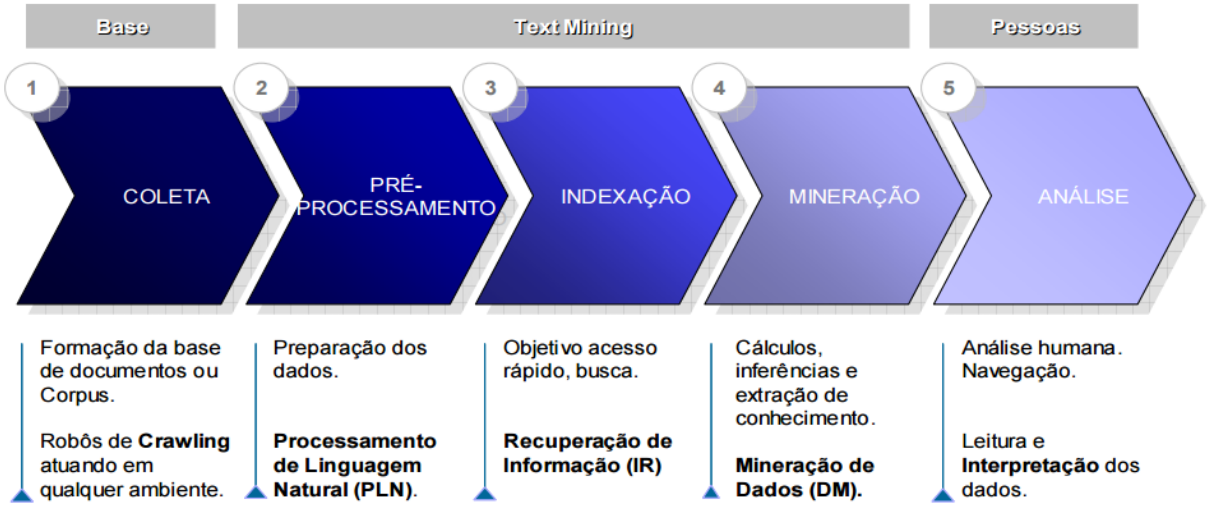
2 TRABALHOS RELACIONADOS

Para o embasamento deste trabalho, foi realizada uma pesquisa em artigos, trabalhos científicos e livros por métodos, conceitos e técnicas das áreas de análise de redes sociais, análise de sentimentos e mineração de textos. Os principais trabalhos relacionados identificados para este trabalho foram os trabalhos de Aranha (2007), Araújo, Gonçalves e Benevenuto (2013), Guzman, Azócar e Li (2014) e Reis, Gonçalves e Araújo (2015) que são apresentados a seguir.

2.1 Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional

Aranha (2007) propõe um modelo de processo para Mineração de Textos (Figura 1) com cinco grandes etapas: coleta, pré-processamento, indexação, mineração e análise da informação.

Figura 1 – Etapas do processo de Mineração de Textos



Fonte: Aranha (2007).

Na fase de coleta, o objetivo é compor a base textual que será analisada. Nessa fase, geralmente utiliza-se *web crawlers* que são softwares utilizados para a extração de dados externos.

Na fase de pré-processamento, o objetivo é estruturar os dados extraídos externamente visando a melhoria na qualidade e organização dos dados. Algumas técnicas são aplicadas, como o processamento de linguagem natural, dicionários léxicos ou simplesmente a exclusão de palavras consideradas desnecessárias.

Na fase de indexação, o objetivo é criar índices para o acesso mais rápido das informações analisadas.

Na fase de mineração, o objetivo é aplicar métodos ou algoritmos para a análise dos dados e extração do conhecimento.

Na fase de análise, o objetivo é avaliar e validar os resultados.

Os primeiros passos deste trabalho seguiram o modelo de processos proposto por Aranha (2007).

2.2 Métodos para Análise de Sentimentos no Twitter

Araújo, Gonçalves e Benevenuto (2013) comparam oito métodos de análise de sentimento propostos na literatura em diferentes contextos, utilizando duas bases de dados diferentes provenientes de redes sociais online. Os métodos analisados foram: LIWC, Happiness Index, SentiWordNet, SASA, PANAS-t, Emoticons, SenticNet e SentiStrength. A primeira base consistia de cerca de 1,8 bilhão de mensagens coletadas no Twitter, representando um histórico completo da rede no período coletado (em 2006 até Agosto de 2009). Dessa base de dados, foram filtrados *tweets* associados a seis eventos sociais relacionados a tragédias, lançamentos de produtos, política, saúde e esporte. A segunda base de dados consistia numa coleção de textos rotulados por humanos como positivo ou negativo. A partir de bases de dados reais, Araújo, Gonçalves e Benevenuto (2013) compararam os oito métodos para análise de sentimentos em termos de abrangência (a fração de mensagens capturadas por cada método) e concordância (a fração de sentimentos corretamente identificados por cada método).

Entre os resultados encontrados, Araújo, Gonçalves e Benevenuto (2013) observaram que os métodos possuíam diferentes graus de abrangência, variando entre 4% e 95% quando aplicados à base de dados associados a eventos reais, sugerindo que, dependendo do método utilizado, apenas uma pequena fração de mensagens será analisada, podendo levar a falsos resultados.

Nenhum dos métodos alcançou altos níveis de abrangência e concordância ao mesmo tempo. O método Emoticons atingiu a maior acurácia (acima de 85%), porém com uma das menores abrangências (4-13%).

Quando os métodos foram aplicados aos dados rotulados, a concordância variou entre 33% e 80%, sugerindo que a mesma amostra de dados poderia ser interpretada de forma diferente dependendo do método escolhido.

Os métodos demonstraram um desacordo na predição de sentimentos para diferentes eventos. Para o evento da queda de um avião, metade dos métodos detectaram mais positividade do que negatividade. O mesmo foi observado em outros eventos que se esperava uma maior quantidade de sentimentos negativos.

Araújo, Gonçalves e Benevenuto (2013) compararam oito métodos propostos pela literatura, em termos de concordância e abrangência com duas bases de dados no idioma em inglês. Também verificaram a polaridade a partir da segunda base de dados, onde mais próximo da polaridade da base de dados o método estava, melhor a predição da polaridade do método. Porém o método SentiStrength foi desconsiderado dessa verificação de predição esperada da polaridade por ter sido treinado utilizando essa mesma base de dados envolvida no processo.

Este trabalho difere por analisar somente a acurácia do método SentiStrength usando o dicionário em português para *tweets* também em português e por verificar a distância do erro da polaridade.

2.3 Sentiment Analysis of Commit Comments in GitHub: An Empirical Study

Guzman, Azócar e Li (2014) se propuseram a coletar comentários do GitHub[3] para fazer análise de sentimentos e responder a quatro questões: quais as emoções relacionadas às linguagens de programação em que um projeto é desenvolvido; quais emoções são relacionadas ao dia da semana ou hora em que o *commit* foi feito; quais emoções estão relacionadas às equipes de desenvolvimento distribuídas geograficamente e quais emoções estão relacionadas à aprovação do projeto.

O primeiro passo de sua execução foi a extração dos dados para serem analisados. Em um segundo passo, foi escolhida a estratégia de dicionário léxico, utilizando o método SentiStrength para fazer a análise de sentimentos. No último passo, foi realizada a análise de sentimentos.

Como resultado, foi possível verificar que as notas médias das emoções nos comentários de cada um dos projetos tendiam a neutralidade e as quatro questões levantadas puderam ser respondidas. As emoções relacionadas às linguagens de programação tiveram o Java como a linguagem com a pontuação mais negativa, seguida de C, C++, JavaScript, PHP, Python e Ruby. As emoções relacionadas ao dia da semana e hora que o *commit* foi feito tiveram a segunda-feira como o dia com emoções mais negativas e o sábado como o dia com

[3] www.github.com

emoções mais positivas, e o fim da tarde a hora com emoções mais negativas. As emoções relacionadas às equipes distribuídas geograficamente demonstrou que projetos com mais países envolvidos tiveram emoções mais positivas. Por fim, as emoções relacionadas à aprovação do projeto não tiveram correlações significativas com emoções positivas.

Guzman, Azócar e Li (2014) fizeram uma coleta de dados no GitHub no idioma em inglês, diferentemente do que é feito neste trabalho, onde os dados foram coletas no Twitter no idioma em português. Guzman, Azócar e Li (2014) não analisaram a acurácia do método SentiStrength (como é feito neste trabalho), porém, este trabalho utiliza a mesma estratégia de dicionário léxico, assim como o método SentiStrength, escolhido por Guzman, Azócar e Li (2014).

2.4 Uma abordagem Multilíngue para Análise de Sentimentos

Reis, Gonçalves e Araújo (2015) propuseram uma abordagem para detecção e análise de sentimentos para mensagens compartilhadas em aplicações da Web. Nessa abordagem, traduziram bases de dados rotuladas em nove idiomas para o idioma em inglês e executaram 13 métodos de análise de sentimentos em sua versão original (desenvolvida e validada para o inglês). Os métodos analisados foram: SentiWordNet, PANAS-t, SASA, SenticNet, Happiness Index, Emolex, NRC Hashtag Sentiment Lexicon, OpinionLexicon, Sentiment 140 Lexicon, VADER, LIWC, SentiStrength e Emoticons. Foram utilizados os idiomas: português, francês, espanhol, italiano, turco, russo, árabe, holandês e alemão.

Com os resultados foi possível verificar que a abrangência e a acurácia dos métodos analisados tiveram consistência mesmo quando submetidos em bases de dados de diferentes idiomas, desde que previamente fossem traduzidos para o idioma original do método, que na maioria dos casos é o inglês.

Reis, Gonçalves e Araújo (2015) propuseram uma abordagem para traduzir a base de dados a ser analisada para o idioma original do método e verificaram a abrangência e acurácia dos métodos analisados. Apesar de terem analisado o método SentiStrength e traduzido o idioma português para o inglês, Reis, Gonçalves e Araújo (2015) não analisaram esses métodos, em especial o método SentiStrength, no idioma português, como é o caso deste trabalho, que se propôs a analisar a acurácia do método SentiStrength no idioma português.

O Quadro 1 apresenta as diferenças e semelhanças dos trabalhos relacionados ao trabalho proposto, no qual PT se refere aos trabalhos que exploraram o idioma diretamente em português e IN se refere aos trabalhos que exploraram o idioma diretamente em inglês.

Quadro 1 - Diferenças e semelhanças dos trabalhos relacionados com o trabalho proposto

	PT	**IN**	**Processo de mineração**	**Coleta no Twitter**	**SentiStrength**	**Tema**
Aranha (2007)	Sim	Não	Sim	Não	Não	Modelo de processo para Mineração de Texto
Araújo, Gonçalves e Benevenuto (2013)	Não	Sim	Não	Sim	Sim	Métodos para Análise de Sentimentos no Twitter
Guzman, Azócar e Li (2014)	Não	Sim	Não	Não	Sim	Comentários do GitHub
Reis, Gonçalves e Araújo (2015)	Não	Sim	Não	Sim	Sim	Uma abordagem Multilíngue para Análise de Sentimentos
Presente Trabalho	Sim	Não	Sim	Sim	Sim	Análise de Sentimento de Tweets em Português

Fonte: elaborada pelo autor.

Sendo assim, o trabalho proposto executou os mesmos passos de mineração de textos proposto por Aranha (2007), além de explorar o método de análise de sentimento SentiStrength com a estratégia de dicionário léxico como Araújo, Gonçalves e Benevenuto (2013), Guzman, Azócar e Li (2014) e Reis, Gonçalves e Araújo (2015). Porém, este trabalho explorou o método SentiStrength com o dicionário diretamente em português para mensagens também em português, diferentemente dos trabalhos citados anteriormente.

3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, serão apresentados os conceitos que fundamentam o desenvolvimento deste trabalho. Na seção 3.1, é introduzido o conceito de Mineração de Textos (*text mining*). Na seção 3.2, os conceitos referentes à Análise de Sentimento (AS – *sentiment analysis*) são apresentados. Na seção 3.3, o Twitter é apresentado.

3.1 Mineração de Textos

"A Mineração de Textos, também conhecida como Descoberta de Conhecimento de Texto refere-se ao processo de extrair padrões interessantes e não triviais ou conhecimento a partir de textos desestruturados." (TAN, 1999, p.1, tradução livre).

Aranha (2007) descreve o processo de mineração de textos como contendo quatro principais etapas: coleta, pré-processamento, indexação e análise da informação.

A etapa inicial é a da coleta, onde o objetivo é compor a base de dados textual do trabalho a ser explorado. Essa base textual conterá as informações a serem extraídas pelas técnicas de mineração de texto.

A segunda etapa é a do pré-processamento e tem como objetivo estruturar a base de dados textual coletada para ser submetida em um formato propício aos algoritmos de extração automática de conhecimento.

A terceira etapa é a da indexação, que tem como objetivo aumentar o desempenho do processo. Geralmente, os dados pré-processados são indexados para o acesso mais rápido.

A quarta e última etapa é a de análise da informação, cujo objetivo é obter um o conhecimento padrão contido na base de dados textual.

Foram realizados, neste trabalho, os passos seguindo o modelo de processos proposto por Aranha (2007) que engloba todas as fases descritas acima. Na fase de análise, este trabalho utilizará de algumas métricas de análise de sentimento identificadas na literatura para avaliar os resultados.

Algumas métricas serão utilizadas para validar a análise dos resultados, tais como:

- Positivo – Percentagem de *tweets* (percentagem sobre o total de *tweets* analisados) classificados com sentimentos positivos;
- Negativo – Percentagem de *tweets* (percentagem sobre o total de *tweets* analisados) classificados com sentimentos negativos;
- Neutro – Percentagem de *tweets* (percentagem sobre o total de *tweets* analisados) classificados com sentimentos neutros;

- Verdadeiro Positivo – Percentagem de *tweets* com sentimentos positivos (percentagem sobre o total de *tweets* analisados), classificados com sentimentos positivos;
- Verdadeiro Negativo – Percentagem de *tweets* com sentimentos negativos (percentagem sobre o total de *tweets* analisados), classificados com sentimentos negativos;
- Verdadeiro Neutro – Percentagem de *tweets* com sentimentos neutros (percentagem sobre o total de *tweets* analisados), classificados com sentimentos neutros;
- Falso Positivo – Percentagem de *tweets* com sentimentos neutros ou negativos (percentagem sobre o total de *tweets* analisados), classificados com sentimentos positivos;
- Falso Negativo – Percentagem de *tweets* com sentimentos neutros ou positivos (percentagem sobre o total de *tweets* analisados), classificados com sentimentos negativos;
- Falso Neutro – Percentagem de *tweets* com sentimentos positivos ou negativos (percentagem sobre o total de *tweets* analisados), classificados com sentimentos neutros;
- Acurácia – Percentagem de *tweets* classificados como Verdadeiro Positivo, Verdadeiro Negativo e Verdadeiro Neutro (percentagem sobre o total de *tweets* analisados), cuja fórmula é: Acurácia = ((Verdadeiro Positivo + Verdadeiro Negativo + Verdadeiro Neutro) x 100) / (Verdadeiro Positivo + Verdadeiro Negativo + Verdadeiro Neutro + Falso Positivo + Falso Negativo + Falso Neutro);
- +Distância Erro – Distância de erro da polaridade positiva para mais ou para menos da polaridade positiva classificada manualmente;
- -Distância Erro – Distância de erro da polaridade negativa para mais ou para menos da polaridade negativa classificada manualmente.

3.2 Análise de Sentimento

Análise de Sentimento ou Mineração de Opinião faz parte de uma área da computação que estuda e visa identificar as opiniões, sentimentos e emoções em textos. As informações analisadas no texto podem ser classificadas como: fatos, que são expressões

claras sobre algo; Ou opiniões, que são expressões subjetivas que declaram os sentimentos (INDURKHYA; DAMERAU, 2010).

Muitas ferramentas e métodos têm sido propostos para fazer análise de sentimentos, acompanhado de diferentes tipos de estratégias, tais como aprendizagem de máquina, dicionários léxicos ou uma combinação delas (GONÇALVES *et al*, 2013).

As estratégias atuais podem ser divididas por duas classes principais: as classes baseadas em aprendizado de máquina, que utilizam um treinamento de um modelo com sentenças previamente rotuladas; e as baseadas em dicionários léxicos, que não utilizam sentenças previamente rotuladas nem treinos para a criação de um modelo (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

3.2.1 Estratégias

3.2.1.1 Aprendizado de Máquina

Segundo Benevenuto, Ribeiro e Araújo (2015), Aprendizagem de Máquina é composta por técnicas supervisionadas, empregando o termo supervisionado pelo fato de exigir uma etapa de treinamento de um modelo com amostras previamente classificadas. Os procedimentos dessa estratégia compreendem quatro etapas principais: obtenção de dados rotulados que serão utilizados para treino e para os testes; definição das características que permitam a distinção entre os dados; treinamento de um modelo computacional com algoritmo de aprendizagem; e aplicação do modelo.

Na etapa dos dados rotulados, o objetivo é obter uma entrada com o seu respectivo rótulo ou classificação. No contexto de análise de sentimentos, seria uma sentença acompanhada de sua polaridade.

Na etapa de definição das características, o objetivo é classificar os dados com características. Denominadas *features*, essas características devem ser atributos que identifiquem cada dado em um conjunto de dados a serem classificados, permitindo assim uma boa distinção entre o conjunto de dados.

Na etapa de treinamento do modelo com algoritmo de aprendizagem, o objetivo é gerar um modelo de classificação de palavras de acordo com a etapa de definições das características.

Na etapa de aplicação do modelo, o objetivo é receber uma sentença de entrada, classificar de acordo com o modelo treinado e ter como saída uma classificação. No contexto de análise de sentimentos, seria identificado no texto o sentimento de uma sentença.

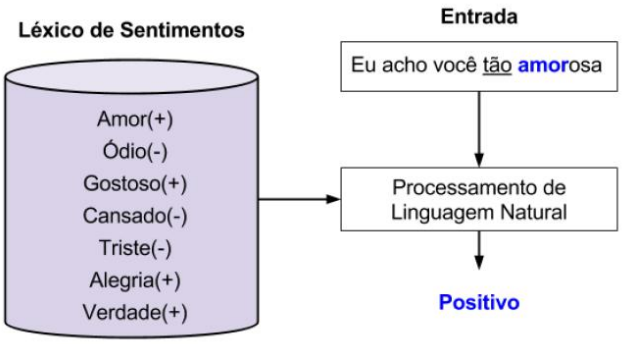
Para o conjunto de dados para qual foi treinado, aprendizado de máquina mostra-se eficaz, tendo uma alta aplicabilidade de seu modelo, porém é restrito ao conjunto de dados. Essa estratégia tem um alto custo computacional em termos de processamento da CPU e memória, característica que pode restringir a capacidade de avaliar um sentimento em dados de *streaming* (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

3.2.1.2 Dicionário Léxico

Dicionário Léxico é uma lista de palavras associadas a sentimentos específicos em que cada palavra possui um significado quantitativo ou qualitativo, que pode ser, por exemplo: um número entre -1 e 1, onde -1 é o valor do sentimento negativo e 1 é o valor do sentimento positivo; ou podem ser termos como positivo/negativo, feliz/triste. Esse tipo de abordagem possui o que é chamado de polaridade prévia, que é uma orientação semântica independente de contexto e que pode ser expressada com valores numéricos ou classes (TABOADA, 2011).

Segundo Benevenuto, Ribeiro e Araújo (2015), a análise de sentimento baseada na estratégia de dicionários léxicos é atualmente umas das estratégias mais eficientes na utilização de recursos computacionais e na capacidade de predição. Os autores descrevem o processo com três passos. O primeiro passo do processo de classificação é receber uma sentença de entrada. Um segundo passo é realizar um processamento de linguagem natural com uma pesquisa no léxico dos termos que formam a mensagem. Por fim, o método é capaz de resultar na polaridade ou sentimento da sentença de entrada. A Figura 2 mostra o processo generalizado.

Figura 2 – Léxico de sentimentos



Fonte: Benevenuto, Ribeiro e Araújo (2015).

Métodos baseados em aprendizagem de máquina funcionam melhor em contextos específicos, porém dependem de bases de dados rotuladas para treinar classificadores. Essa dependência é considerada como uma desvantagem, pois é alto o custo da obtenção desses dados e a aplicabilidade do modelo é restrita ao contexto específico em que foi criado. Por sua vez, métodos baseados em dicionários léxicos utilizam listas de palavras genéricas associadas a sentimentos específicos, não dependendo de dados rotulados para treinamento. Apesar de serem considerados menos eficientes em contextos específicos, a eficiência do método está diretamente relacionada à generalização do vocabulário do modelo utilizado e sua aplicação não está restrita ao contexto específico (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

Neste trabalho, foi utilizado o método SentiStrength que utiliza um dicionário léxico anotado por seres humanos e melhorado com o uso de Aprendizado de Máquina.

3.2.2 Métodos de análise de sentimentos

Existem na literatura muitos métodos de análise de sentimentos com diferentes estratégias como aprendizagem de máquina e dicionários léxicos. O Quadro 2 apresenta uma descrição dos principais métodos para análise de sentimentos disponíveis na literatura e o tipo de estratégia, Aprendizagem de Máquina (AM) ou Dicionário Léxico (DL) (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

Quadro 2 – Métodos para análise de sentimentos

Nome	**Descrição**	**DL**	**AM**
Emoticons	Possui uma lista de *emoticons* dividida em positivos (":)") e negativos(":("). O texto é classificado de acordo com a classe que tiver mais *emoticons*. Apesar de possuir uma alta taxa de acertos este método depende muito da presença do *emoticon* no texto.	✓	
Opinion Lexicon	Também conhecido como Sentiment Lexicon, consiste de uma lista com cerca de 6.800 palavras rotuladas como positivas e 6.800 palavras rotuladas como negativas, incluindo gírias e abreviações no idioma Inglês. Este é um método léxico criado a partir de textos coletados em *reviews* de produtos em sites de compra.	✓	
Opinion Finder (MPQA)	É uma ferramenta considerada híbrida pois utiliza um léxico de sentimentos mas utiliza Naive Bayes para distinguir se uma sentença é subjetiva ou objetiva.	✓	✓

Happiness Index	É uma escala de sentimentos que utiliza o popular ANEW (um conjunto de palavras ligadas a emoções do Inglês). Este método foi construído para avaliar textos entre 1 e 0, indicando a quantidade de felicidade existente. Em particular os autores utilizaram este método para mostrar que a "quantidade de felicidade" nas letras das músicas diminuiu entre 1961 e 2007.	✓	
SentiWordNet	É um léxico construído a partir de outro léxico já conhecido chamado WordNet. No WordNet os autores agruparam adjetivos, substantivos, verbos em conjuntos de palavras que fossem similares formando uma rede de palavras. Já os autores do SentiWordNet associaram uma polaridade entre algumas palavras-semestes do WordNet e propagaram essa polaridade nas palavras similares da WordNet criando um amplo léxico de sentimentos.	✓	✓
LIWC	O LIWC é uma ferramenta bem estabelecida e utilizada em diversas áreas, e contou com o aval de psicólogos, sociólogos e linguistas durante seu desenvolvimento. Ela possui um dicionário léxico de aproximadamente 4500 palavras e raízes de palavras, fazendo parte de oitenta categorias das mais variadas (ansiedade, saúde, lazer etc).	✓	
SenticNet	SenticNet é um dicionário semântico e afetivo para opinião em nível de conceito e análise de sentimento. Ele foi construído através do que é denominado pelos autores de sentic computing, um paradigma que explora Inteligência Artificial e técnicas de Web semântica para processar opiniões via mineração de grafos e redução de dimensionalidade. Ele é público e provê um bom material para mineração de opiniões em nível semântico e não apenas sintático.	✓	
AFINN	É um léxico construído a partir do ANEW mas com o foco em redes sociais, contendo gírias, acrônimos e palavras de baixo calão da língua Inglesa. Ele possui uma lista de 2.477 termos classificados entre -5(mais negativo) e +5(mais positivo).	✓	
SO-CAL	É um método léxico que leva em conta a orientação semântica das palavras (SO). Foi criado contendo unigramas (verbos, advérbios, substantivos e adjetivos) e multi-gramas (intensificadores e frases) numa escala entre -5 e +5. Os autores também incluíram analisador de partes do discurso e negação.	✓	
Emoticons DS (Distant Supervision)	É um léxico que possui termos gerados a partir de uma extensa base de dados do Twitter. Estes termos foram classificados automaticamente baseando-se na frequência de *emoticons* positivos ou negativos nas sentenças.	✓	

NRC Hashtag	É um léxico que utiliza a técnica de supervisionamento distante para classificar seus termos. De forma geral, ele classifica os termos provenientes do Twitter considerando as *hashtags* que o contém (i.e #joy, #sadness).	✓	
Pattern.en	É um pacote da linguagem python para lidar com processamento de linguagem natural. Um de seus módulos é responsável para inferir o sentimento no texto. Criado para ser rápido ele é baseado em polaridades associadas ao WordNet.	✓	
SASA	Foi criado para detectar sentimentos no Twitter durante as eleições presidenciais de 2012 nos Estados Unidos. Ele foi construído a partir de modelos estatísticos do classificador Naïve Bayes em cima de unigramas classificados. Ele também explora emoções em *emoticons* e exclamações.		✓
PANAS-t	Tem como objetivo inicial detectar as flutuações de humor dos usuários no Twitter. O método é um léxico adaptado a partir de uma versão adaptada do PANAS Positive Affect Negative Affect Scale. O PANAS é uma conhecida escala psicométrica que possui um grande conjunto de palavras associadas a 11 diferentes tipos de humor (surpresa, medo, serenidade etc).	✓	
EmoLex	É um léxico criado a partir do Amazon Mechanical Turk, no qual pessoas foram pagas para classificar os termos. Cada entrada está associada a 8 sentimentos básicos em inglês: *joy*, *sadness*, *anger* etc. A base do Emolex foi construída utilizando termos do Macquarie Thesaurus e palavras do General Inquirer e do Wordnet.	✓	
SANN	Foi construído para inferir a nota de avaliação de comentários dos usuários de produtos utilizando análise de sentimentos. Os comentários foram integrados em um classificador (kNN) ou K-Vizinhos mais próximos.		✓
Sentiment140 Lexicon	É um léxico criado de maneira similar ao NRC Hashtag. Foi utilizado um classificador SVM que utilizava *features* como: número e categoria de *emoticons*	✓	
SentiStrength	Constrói um dicionário léxico anotado por seres humanos e melhorado com o uso de Aprendizado de Máquina. SentiStrength atribui pontuações a *tokens* de um dicionário, onde *emoticons* também estão incluídos. A palavras com emoções positivas são atribuídos valores entre 1 e 5 e a palavras com emoções negativas são atribuídos valores entre -5 e -1. Os valores 1 e -1 são usados para indicar emoções neutras, enquanto que 5 e -5 são usados para indicar emoções muito positivas e muito negativas, respectivamente. SentiStrength divide o texto em trechos de uma ou mais sentenças e atribui valores positivos e negativos para cada sentença, informando a pontuação máxima ou mínima entre todas as palavras de uma	✓	✓

	sentença.		
Stanford Recursive Deep Model	Tem como proposta uma variação do modelo de redes neurais chamadas Redes Neurais Recursivas que processa todas as sentenças procurando identificar sua estrutura e computar suas interações. É uma abordagem interes, pois a técnica leva em consideração a ordem das palavras na sentença, por exemplo, que é ignorada por vários métodos.	✓	✓
Umigon	Pertence à família de léxicos e foi proposto para detectar sentimentos no Twitter, além de subjetividade. O método utiliza diversos recursos linguísticos como onomatopéias, exclamações, *emoticons* etc. Ele possui heurísticas responsáveis por desambiguar o texto baseadas em negações, palavras alongadas e *hashtags.*	✓	
Vader	Possui como base um dicionário léxico criado a partir de uma lista de palavras com base em dicionários já bem estabelecidos como LIWC, ANEW e GI. Em seguida, foram adicionadas construções léxicas presentes em microblogs tais como *emoticons*, acrônimos e gírias que expressam sentimentos.	✓	

Fonte: Quadro adaptado de Benevenuto, Ribeiro e Araújo (2015).

Por ser o método escolhido para realizar a análise de sentimentos neste trabalho, o SentiStrength será apresentado em detalhes a seguir.

3.2.2.1 SentiStrength

SentiStrength utiliza um dicionário léxico anotado por seres humanos e melhorado com o uso de Aprendizado de Máquina. SentiStrength atribui pontuações a *tokens* de um dicionário, onde *emoticons* também estão incluídos. Palavras com emoções positivas são atribuídos valores entre 1 e 5 e palavras com emoções negativas são atribuídos valores entre -5 e -1. Os valores 1 e -1 são usados para indicar emoções neutras, enquanto que 5 e -5 são usados para indicar emoções muito positivas e muito negativas, respectivamente. SentiStrength divide o texto em trechos de uma ou mais sentenças e atribui valores positivos e negativos para cada sentença, informando a pontuação máxima ou mínima entre todas as palavras de uma sentença (BENEVENUTO; RIBEIRO; ARAÚJO, 2015). O Quadro 3 apresenta um exemplo de como o SentiStrength analisa e classifica as frases.

Quadro 3 – Exemplo de análise e classificação do método SentiStrength

Frase	**Análise**	**Positivo/Negativo**	**Sentimento**
Eu amo minha mãe	Eu amo[4] minha mãe [sentence: 4,-1]	4, -1	Positivo
Fui ao cinema ontem	Fui ao cinema ontem [sentence: 1, -1]	1, -1	Neutro
Eu odeio quem fala com ignorância	Eu odeio[-4] quem fala com ignorância [sentence: 1, -4]	1, -4	Negativo

Fonte: elaborada pelo autor.

SentiStrength possui alguns dicionários em outros idiomas além do seu idioma nativo (inglês), tais como: finlandês, alemão, holandês, espanhol, italiano, russo. O dicionário em português, apesar de estar disponível no site do método, até o presente momento, ainda não tinha sido testado. Cada dicionário do SentiStrength é composto por arquivos que contém as palavras, *emoticons*, palavras de negação, gírias, palavras de intensidade, e seus respectivos pesos, sendo quatro os arquivos principais explorados neste trabalho no dicionário em português: EmoticonLookupTable.txt, EmotionLookupTable.txt, NegatingWordList.txt e BoosterWordList.txt.

EmoticonLookupTable.txt tem como objetivo anotar *emoticons* com as suas respectivas emoções, por exemplo, o *emoticon* "(-:" (carinha feliz) tem peso positivo 1.

EmotionLookupTable.txt tem como objetivo relacionar uma palavra com um peso positivo ou negativo. Esses pesos ficam entre -5 e -2, para palavras muito negativas e pouco negativas, respectivamente, e entre 2 e 5 para palavras poucos positivas e muito positivas, respectivamente. Por exemplo, a palavra "amor", foi atribuída peso 4, por sua vez a palavra "ódio", foi atribuída peso -4.

NegatingWordList.txt tem como objetivo anotar palavras de negação. Por exemplo, "não amo", o "não" presente na frase estaria negando o sentimento de amor.

BoosterWordList.txt tem como objetivo anotar palavras que intensificam as emoções, por exemplo, "muito amor", o "muito" presente na frase estaria intensificando a emoção amor, aumentando assim o seu peso positivo.

3.3 Twitter

O conceito de redes sociais é bastante associado às tecnologias da informação, sendo comum se pensar em comunidades virtuais, como o Twitter, Facebook, Linkedin e

Instagram, quando se ouve falar em redes sociais. Foi somente com o desenvolvimento da tecnologia e com o surgimento das redes sociais virtuais que esse conceito se tornou amplamente explorado (MACHADO; TIJIBOY, 2005). Porém, esse conceito vai além da tecnologia da informação e vem sendo historicamente estudado e discutido pelas Ciências Sociais antes da tecnologia da informação (ACIOLI, 2007). Alexandrino (2016, p.15) define uma rede social como "uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, podendo ser redes reais ou virtuais".

Neste trabalho, o termo "redes sociais" será usado referenciando as redes sociais virtuais. E a rede social Twitter será usada para a extração das informações na fase da coleta que será abordada na seção 4.2.

Sendo uma das redes sociais mais utilizadas (ALEXANDRINO, 2016), o *microblog* Twitter surgiu em 2006 com uma proposta simples: permitir que usuários divulguem o que estão fazendo em tempo real através de mensagens de textos de até 140 caracteres, conhecidos como *tweets*. Esse modelo foi determinante para o modo como as pessoas se expressariam, pois, apesar da proposta inicial, os usuários do Twitter utilizam essa rede social para compartilhar informações e opiniões sobre fatos, eventos em geral, produtos e serviços, marcas etc, permitindo assim que o Twitter se tornasse uma importante plataforma de troca de opiniões (NASCIMENTO; OSIEK; XEXÉO, 2015).

O Twitter fornece, para desenvolvedores de software, uma REST API[4], que possibilita acessar os dados dos usuários, como *tweets* e status em tempo real e em nível global (CARVALHO FILHO, 2014). REST (Representation State Transfer) é definido por Carvalho Filho (2014, p.18) como "uma arquitetura de redes de estilo híbrido derivada do estilo de várias arquiteturas baseadas em rede, que define uma interface conectora que permite aos clientes 'conversar' com servidores de maneira única".

O Twitter, por se tratar de uma das redes sociais mais utilizadas; e uma importante plataforma de troca de opiniões e de oferecer o uso de sua REST API, foi escolhida por este trabalho como fonte para a extração das informações para fazer análise de sentimentos em português.

[4] https://developer.twitter.com/

4 PROCEDIMENTOS METODOLÓGICOS

Nesta seção, serão descritas todas as fases da execução do trabalho.

O objetivo deste trabalho é analisar a eficácia do método de análise de sentimento SentiStrength utilizando o dicionário em português. Os cinco primeiros passos de execução realizados neste trabalho foram baseados no modelo de processo para mineração de textos proposto por Aranha (2007). Os passos realizados neste trabalho foram:

1. O primeiro passo foi a coleta do conjunto de dados a ser analisado;
2. O segundo passo foi o pré-processamento dos dados coletados;
3. O terceiro passo foi a indexação dos dados pré-processados;
4. O quarto passo foi a mineração, onde foi aplicado o método SentiStrength com o dicionário em português nos dados indexados;
5. O quinto passo foi a análise da informação, onde foi verificada manualmente a classificação do quarto passo com o objetivo de identificar erros na classificação dos sentimentos realizado pelo método SentiStrength;
6. O sexto passo foi sugerir uma melhoria no dicionário;
7. O sétimo passo foi a mineração, onde foi executado o método SentiStrength com a melhoria sugerida no dicionário nos dados indexados do quarto passo para comparar o resultado encontrado antes e depois da mudança no dicionário;
8. O oitavo passo foi a análise da informação, onde foi verificada manualmente a classificação do sétimo passo.

Para realizar os quatro primeiros passos da metodologia foi desenvolvido um software para automatizar o processo de coleta, pré-processamento, indexação e mineração. Apenas as análises manuais (passos 5 e 8) e a melhoria do dicionário (passo 6) foram feitas sem auxílio de software.

4.1 Desenvolvimento do Software

Inicialmente, para a execução do trabalho foi desenvolvido um software para automatizar os passos das seções 4.2, 4.3, 4.4 e 4.5. O software foi desenvolvido em NodeJS[5].

[5] https://nodejs.org

Sua escolha se deveu à alta escalabilidade, a facilidade de manutenção de código, o curto tempo para o desenvolvimento e o grande número de bibliotecas[6] de código aberto.

4.2 Coleta

O software desenvolvido utilizou a biblioteca em NodeJS disponível no site do Twitter para a comunicação com a API pública do Twitter, automatizando as buscas por *tweets* que fazem menção a um ou mais termos. O Twitter possui algumas limitações importantes: o número de chamadas feitas em um determinado período de tempo para a sua API; A quantidade de *tweets* obtidos em uma única chamada feita a sua API; e a data dos *tweets* buscados. Todas as limitações foram tratadas pelo software desenvolvido e estão sujeitas a mudanças.

Foram realizadas duas coletas de 1000 *tweets* em português através da API pública do Twitter. A primeira coleta fazia menção ao termo “motorola” (fabricante de celulares), por se esperar diferentes opiniões sobre a marca, seus produtos e serviços. A segunda coleta fazia menção ao termo “liga da justiça” (filme), também por se esperar diferentes opiniões sobre o tema e por ser em um contexto diferente ao da primeira coleta. Ao todo foram coletados 2000 *tweets* em Novembro de 2017.

4.3 Pré-processamento

Os dados foram pré-processados removendo quebras de linha e espaços múltiplos. Com o uso do método SentiStrength não é necessário outro tipo de pré-processamento, como por exemplo, remoção de links.

4.4 Indexação

Os dados coletados foram indexados e armazenados em um banco de dados não relacional denominado MongoDB[7] para o acesso mais rápido posterior. Sua escolha se deveu à fácil integração com o NodeJS e por seu modelo não estruturado de dados, já que os *tweets* tendem a mudar os seus atributos.

[6] Coleção de subprogramas.
[7] https://www.mongodb.com/

4.5 Mineração

O software desenvolvido executou a biblioteca em Java[8] com licença acadêmica e o dicionário em português disponíveis no site do método SentiStrength. A biblioteca recebe como parâmetros o local do arquivo onde está o dicionário a ser utilizado e o local do arquivo de texto ou simplesmente o texto a ser analisado. Como resultado, a biblioteca devolve um arquivo de texto com a classificação no mesmo local do arquivo de texto passado por parâmetro, ou, se for o caso, devolve como resultado apenas a classificação do texto passado por parâmetro.

4.6 Análise da Informação

Após a etapa de mineração cada *tweet* foi analisado manualmente pelo método com o dicionário em português com suas respectivas pontuações.

A análise manual levou em consideração o peso de cada palavra contida nos *tweets* de acordo com palavras classificadas pelo autor deste trabalho como positivo, negativo ou neutro e seus respectivos pesos, e deu pesos positivos e negativos a cada *tweet*, imitando o processo de análise do método SentiStrength (Quadro 3).

O Quadro 4 mostra as palavras classificadas pelo autor deste trabalho, usadas para comparar com a análise feita pelo método. Todas as palavras do Quadro 4 foram encontradas uma ou mais vezes na primeira base de dados de *tweets* que faziam menção ao termo "motorola". Palavras de baixo calão e abreviações foram consideradas.

Quadro 4 – Palavras encontradas na primeira coleta e classificadas pelo autor

Palavras positivas	adoro(3), agradou(3), amantes(4), amo(4), amor(4), amorzinho(4), apaixonada(4), bacana(2), bonita(2), bonitos(2), bons(3), Deus(4), elegante(2), elogiar(3), elogios(3), especial(3), excelente(4), feliz(4), felizes(4), felizmente(4), fofo(3), ganhar(2), ganhei(2), gostam(3), gostando(3), gostaria(3), gostava(3), graças(3), humilde(2), incrível(4), interessante(2), interesse(2), legal(2), lendários(4), lindo(3), lindos(3), maravilhas(3), maravilhosa(3), maravilhoso(3), massa(3), melhor(4), melhores(4), moral(3), obrigada(2), ótima(4), ótimos(4), parabéns(4), pfvr(2), poderosíssimo(4), positivo(2), potente(3), preferidos(3), prós(2),

[8] https://www.java.com/pt_BR/

	queridinho(3), querido(3), recomendo(3), reizinho(3), risos(2), santo(2), satisfeito(3), sensacional(3), sorte(3), status(2), talentosa(2), tesão(3), top(3), venero(4), vlw(2)
Palavras negativas	abomino(-4), arrebente(-3), arrepender(-2), arrependi(-2), bizarra(-3), bolado(-2), bosta(-3), bucetão(-3), bugada(-2), cacete(-3), cagar(-3), cagaram(-3), cagou(-3), cagueta(-2), caraio(-3), caralho(-3), chatas(-3), chorar(-4), choroso(-3), coitada(-3), complicado(-2), confusa(-2), confuso(-2), contras(-2), credo(-2), critiquei(-2), cu(-3), cuzão(-3), decepção(-3), decepciona(-3), denunciar(-3), desconfortável(-3), desgraça(-3), dor(-4), erros(-2), escandalosa(-2), estraga(-2), estressado(-3), estressou(-3), evitem(-2), expulsaram(-2), falência(-3), fdp(-3), feia(-2), feio(-2), feios(-2), ferrar(-3), fraca(-3), fracasso(-3), fraude(-3), frescura(-3), foda(-3), fodam(-3), fodas(-3), fodida(-3), fodidos(-3), fuder(-3), fudeu(-3), fudida(-3), furtaram(-3), horrível(-4), horrososo(-3), indesejados(-2), infelizmente(-4), inferno(-3), louca(-3), manchar(-2), matar(-4), maus(-2), merda(-3), morrendo(-3), morreu(-3), morto(-3), nojo(-3), obriga(-3), odiei(-4), ódio(-4), osso(-2), péssimo(-4), pika(-3), pior(-4), piranhas(-3), porcaria(-3), porra(-3), pqp(-3), prr(-3), puta(-3), puts(-3), quebra(-2), ranço(-4), reclamação(-2), reclamações(-2), reclamar(-2), revoltado(-3), ridícula(-3), roubado(-3), roubar(-3), roubava(-3), ruim(-4), ruinzinho(-4), socando(-2), socorro(-4), socos(-2), sofrer(-4), surtando(-3), tapada(-2), tnc(-3), treta(-3), trouxa(-3), viado(-3), vsf(-3), vtnc(-3), zoar(-2)
Palavras neutras	ai, aposentar, baixar, barato, beijo, breve, carga, caro, claro, como, deve, diminuir, ei, esperando, falta, grande, jogar, juros, lança, mano, muito, namorada, não, nix, ônix, padrão, parar, pobres, propaganda, salvar, saudade, tipo, usado, valor, valores, vão, velho, x
Palavras de intensidade	bastante(1), bem(1), demais(1), mais(1), mt(1), mto(1), muita(1), muito(1), super(2), tanto(1), tão(1), ultra(2)
Palavras de negação	n

Fonte: elaborada pelo autor.

Ao final dessa etapa, foi possível verificar a acurácia do método SentiStrength com o dicionário em português para as duas coletas de *tweets* que faziam menção aos termos “motorola” e “liga da justiça”. Também foi possível verificar a distância do erro das polaridades positivas e negativas na coleta de *tweets* que faziam menção ao termo “motorola”. Os resultados estão na seção 5.

4.7 Melhoria do Dicionário

Após verificar a acurácia do método SentiStrength para as duas coletas descritas na seção 4.2, foi feita uma sugestão de melhoria no dicionário em português.

O dicionário foi atualizado de acordo com as palavras do Quadro 4. Tais palavras foram observadas na primeira coleta de *tweets* que faziam menção ao termo “motorola”.

Todas as palavras contidas nos arquivos descritos na seção 3.2.2.1 do dicionário em português também foram analisadas. Algumas sofreram modificações de exclusão ou alteração de peso do sentimento, a fim de dar uma melhor classificação (positivo/negativo) de acordo com o seu significado. Palavras consideradas neutras (exemplos: “abertura”, “açougue” e “anexo”) e palavras que não estão presentes na língua portuguesa com exceções de abreviações e algumas palavras em inglês comumente usadas nos *tweets* (exemplos: “ok”, “okays” e “oks”) foram excluídas. Tais palavras podem ser observadas no Quadro 5.

Quadro 5 – Palavras retiradas e/ou modificadas do dicionário em português original do método SentiStrength

EmotionLookupTable.txt (Retiradas)	abertura(1), abrupto(-2), abusi(-4), acaso(-2), accus(-2), acostar(-2), açougue(-2), acrimon(-2), acrobacia(-2), admir(3), admoni(-2), ador(4), adulterat(-2), adventur(1), advers(-2), agarrar(-2), aggravat(-3), agitat(-2), agoniz(-4), agreeab(1), agu(-3), ai(-2), alegação(-2), alegar(-2), aleijado(-2), álibi(-2), alienat(-2), almejar(-2), alol(2), amaz(3), ambiguit(-2), ambíguo(-2), ambivalente(-2), amus(2), analfabeto(-2), anarquia(-2), anarquista(-2), anexo(1), angr(-4), antagoni(-2), antitrust(-2), anulação(-2), anular(-2), anxi(-3), aok(2), apagar(-2), aparência(-2), apath(-2), apesar(-4), aplicar(-2), aposentar(-2), appreciat(2), apprehens(-3), apreender(-2), apressado(-2), ardente(-2), argh(-2), argumentos(-2), arma(-1), arraste(-2), arrogan(-3), arruda(-2), arsehole(-3), artificial(-1), asham(-4), assinalada(-1), asswipe(-3), atrofia(-2), ausência(-2), avaliado(2), avaric(-3), aversi(-3), azedo(-2), baba(-2), baixa(-2), baixar(-2), bala(-2), banal(-3), banana(-3), barato(-2), barbari(-3), barf(-2), barreira(-2), batalha(-2), beaut(3), bebê(2), bebês(2), beicinho(-2), beijo(3), beijoca(-2), berk(-2), bff(4), bg(2), birdbrain(-2), blah(-2), blam(-2), bloco(-2),

	blur(-2), blurt(-2), bocejo(-2), bonehead(-2), brandir(-2), brillian(2), brincando(1), bulir(-2), bullshit(-3), bumhole(-2), burgl(-2), caçador(-2), calandra(-2), cancelar(-2), canhão(-2), capitular(-2), captura(-2), carenagem(-2), carga(-2), caro(-2), carrossel(-2), cavalo(2), caverna(-2), ceder(-2), cegos(-2), censor(-2), censura(-2), cerco(-2), cerda(-2), céu(2), chama(-3), chás(-3), cheio(-2), chicote(-2), chuckl(3), chupar(-2), cinzento(-2), claro(-2), coaxar(-2), cocksucker(-4), combatente(-2), comed(1), cometer(-2), como(2), compartilhada(1), compartilhar(1), competir(-2), comum(-2), concurso(3), condescen(-3), confissão(-2), confus(-2), contrariar(-2), contrário(-2), controvers(-2), cool(2), cortar(-2), courag(2), crap(-3), crappy(-3), craz(-1), crepitar(-2), critici(-2), crônica(-3), cruz(-2), crye(-4), cumplicidade(-2), damag(-1), defenc(-1), defensiva(-2), deitado(-4), delectabl(3), delgado(-2), delicious(3), deligh(3), dente(-2), dependentes(-2), depor(-2), derramar(-2), desafiador(-2), desafiar(-2), desafio(-2), desarmar(-2), desconhecido(-2), desculpe(-2), desenraizar(-2), desfazer(-2), desfeito(-2), desfiladeiro(-2), deslocar(-2), desperat(-3), despertaram(3), despis(-4), destruct(-1), desviar(-2), desvio(-2), determinado(1),deve(-2), devot(3), diâmetro(-2), dickhead(-2), digni(2), dilema(-2), diminuição(-2), diminuir(-2), din(-2), dinâmi(1), direito(-2), disagre(-2), disapprov(-2), disparar(-2), disputável(-2), dissatisf(-2), dissuadir(-2), distanciamento(-2), ditar(-2), divin(2), divisão(-2), dominação(-2), dominar(-2), dork(-3), douchebag(-2), downhearted(-2), easie(1), ecsta(4), eejit(-3), egotis(-3), ei(2), elegan(2), emocional(-2), empe(-2), empinar(-2), empt(1), encargos(-2), enemie(-2), energ(1), enervar(-2), enga(1), engolfar(-2), enrag(-4), enterrada(-3), enterrar(3), enthus(3), entorse(-2), envie(-3), esbanjar(-2), esboçado(-2), escaldadura(-2), escape(-2), escória(-3), escoriáceo(-2), escrutinar(-2), escurecer(-2), escuro(-2), esfregaço(-2), esotérico(-2), espera(2), esperando(2), espreitar(-1), estáticas(-2), estola(-2), estranged(-2), estrangeiro(-2), estrangeiros(-2), estridente(-2), estrita(-2), exasperat(-2), excel(2), excentricidade(-2), excêntrico(-2), excommunicat(-2), excruciat(-5), execuções(2), executar(-2), exigível(-3), exílio(-2), expediente(-2), expor(-2), exprobrar(-2), exterminat(-2), fab(3), fabricat(-2), faca(-2), fácil(1), faixas(4), fallout(-2), faroleiro(-2), fascista(-2), fathead(-3), fatigu(-3), feroc(-4), feroz(-2), fervilhar(-2), festiv(1), feudo(-3), fiesta(1), fiscal(-2), flexib(1), flexibilização(1), forgiv(2), forjado(-1), fricção(-2), fucker(-3), fuckface(-3), fucks(-3), fuckwit(-2), fuctard(-2), fud(-3), fugaz(-2), fugly(-4), fumaça(-3), fuming(-4), fundador(-2), funn(2), gaguejar(-2), galo(-2), garança(-3), garra(-2), gay(-2), geek(-1), genero(2), gentlest(3), germe(-2), gibberi(-2), giggl(3), glori(2), gmbo(3), graci(2), grande(3), gratef(2), grati(2), gratuito(-2), graves(-2), graxos(-2), griev(-4), grinn(2), grr(-3), gueto(-2), guincho(-2),

	h8(-4), ha(2), habitar(-1), handsom(2), happi(3), harmon(2), hater(-4), heartbroke(-4), heartwarm(3), hesita(-1), hijack(-2), hogwash(-2), homossexual(-3), hooey(-2), horr(-4), hugg(3),hungr(-2), ideal(1), ignor(-3), imóveis(-2), imóvel(-2), impatien(-2), impessoal(-1), importan(1), inclinação(-2), incompatib(-3), incompeten(-3), inconvenien(-2), indecis(-2), independentemente(-2), indetermin(-2), ineffect(-2), inevitável(-2), inexplicável(-2), infantil(-2), infiltração(-2), inflação(-2), injunção(-2), innocen(2), insecur(-2), insincer(-2), inspir(3), insufficien(-2), intell(2), intimidat(-4), intrud(-2), invigor(2), invisível(-1), isentar(-2), jerked(-1), jittery(-2), jogar(2), joll(3), jurar(-2), juros(2), jurou(-2), kindn(2), labuta(-2), laço(-2), laidback(2), lança(-2), lazie(-2), ligeira(-2), likeab(3), linguado(-2), liquidação(-2), liquidar(-2), litig(-2), livel(2), livrar(-2), lmao(3) lolol(2), lous(-3), loveless(-3), lucked(3), lucki(3), lucks(-3), lunkhead(-2), lutado(-3), lutar(-2), luv(3), maddest(-3), magnific(4), manipulat(-2), mano(2), mar(-2), masochis(-3), meanie(-2), melanchol(-4), merr(2), minar(-2), míope(-2), misses(-3), mistak(-2), misunderst(-2), mooch(-2), moodi(-3), moody(-3), mortal(-2), mortalha(-2), mortif(-3), mourejar(-2), muah(3), mudo(-2), muito(3), multicor(-2), mundano(-2), musaranho(-2), muthafuck(-4), namorada(3), não(-3), narcótico(-2), nast(-3), negligen(-3), nerd(-3), neutralização(-2), neutralizar(-2), ninhada(-2), nix(-2), nonsens(-2), notório(-2), novato(-2), ntre(-2), nucklehead(-2), numpty(-2), nurtur(1), nutter(-2), obes(-2), ociosidade(-2), oco(-2), ocultar(-2), offens(-3), omfg(2), openminded(1), opportun(1), optimi(1), ousadia(1), outrag(-3), padrão(-2), painf(-4), paining(-4), painl(2), palatabl(2), panelinha(-2), parafuso(-3), paranoi(-3), parar(-2), partes(1), partição(-2), partie(1), partilha(1), parto(-2), peculiar(-1), pendurar(-2), pequenino(-2), pequeno(-1), perambular(-2), persecut(-3), pessimis(-3), petrif(-4), pettie(-1), phobi(-3), pillock(-2), pinhead(-2), pitada(-2), pleasur(3), plebeu(-2), plonker(-2), pneu(-2), pobres(-2), popa(-2), porte(-2), positiv(3), prais(3), prannock(-2), prat(-2), prejudic(-3), pressentimento(-2), pressur(-2), prettie(3), privileg(2), prod(-2), promis(2), pronto(1), propaganda(-2), provação(-2), puk(-3), punho(-2), queimar(-2), quentes(1), questionável(-2), ração(-2), racis(-2), radiano(3), radical(-2), raid(-2), ralo(-2), rampante(-2), rasgar(-2), raso(-2), raspar(-2), rato(-2), reacionário(-2), reassur(2), reativa(-2), recalcitrante(-2), recessão(-2), recorrida(-2), recuo(-2), recusa(-2), recusar(-2), redundan(-2), refrão(-2), refugiados(-2), refutar(-2), regresso(-2), reluctan(-2), renúncia(-2), renunciar(-2), repartição(-2), resíduos(-2), resolv(2), reter(-2), revogação(-2), revogar(-2), ricos(1), ridicul(-2), riqueza(1), rock(2), rofl(3), romanc(4), rumor(-2), sábio(1), sacudiu(-2), sadde(-4), saída(1), saltitante(-2), salvar(2), sap(-2), sarcas(-3), satisf(2), saudade(-3), scariest(-4), se(1), secessão(-2), secur(2), seduzir(-2), segredo(-1),

	segregação(-2), senil(-2), separar(-2), separadamente(-2), sério(-1), servil(-2), shaki(-2), sigilo(-1), silli(2), simples(-2), simplista(-2), simulação(-2), simulado(-3), sincer(1), sintoma(-2), sitiar(-2), skank(-2), slur(-2), smart(2), smil(3), sociab(2), solteirona(-2), soluçando(-4), soluço(-4), soluços(-4), soluçou(-4), sonolência(-2), sonolento(-2), soulmate(3), sozinho(-1), spaz(-2), splend(3), startl(-2), straggler(-2), struggl(-2), stunk(-3), suavemente(2), suavemente(2), subjugat(-2), submisso(-2), subserviência(-2), substituição(-2), subterrâneo(-2), subtrair(-2), sucky(-2), suga(-2), sugado(-2), suk(-2), sunshin(2), sup(2), super(3), superstit(-2), supervisão(-2), suportada(2), suprem(3), surdez(-2), surdo(-2), surpreendente(1), suspicio(-2), sweetie(3), tabu(-2), tard(-2), tarifa(-2), tehe(3), temperamento(-3), temperamentos(-1), temporariamente(-2), terribl(-4), terrifyi(-4), terrori(-3), thankf(2), thanx(2), thnx(2), thrash(-2), tímido(-2), tipo(2), tiro(-2), toleran(2), torrent(-1), torto(-2), tortur(-4), tosser(-2), tosspot(-2), traged(-3), tratar(2), treacher(-3), treasur(2), trembl(-3), trivi(-2), trocadilho(-2), troubl(-2), truant(-2), trudg(-2), truque(-2), tubarão(-2), twat(-2), twunt(-2), ugh(-3), ugl(-3), unauthentic(-2), uncomfortabl(-3), undependability(-2) undependable(-2), undid(-1), uneas(-2), unhapp(-3), unimpress(-2), unlov(-3), unsavo(-3), unwelcom(-3), upchuck(-2), usado(-2), uva(2), vaga(-2), vaguear(-2), valuabl(2), velhaco(-2), velho(-2), veto(-2), vice(-2), virtuo(3), volatilidade(-2), vom(-3), vulgar(-2), vulnerab(-3), wack(-2), wanker(-2), wassock(-2), wazzock(-2), welcom(2), wickedn(-3), winn(1), wonderf(3), worr(-4), wow(3), wtf(-3), x(2), xox(2), xx(2), yay(2), yays(3)
EmotionLookupTable.txt (Modificadas)	altivo(2), ansioso(-2), apuro(2), astuto(2), bônus(2), brando(2), brilho(2), bruxa(-2), calmaria(2), calmarias(2), caprichoso(2), cauteloso(2), comemorar(3), conforto(3), contentamento(3), crédulo(2), cuidado(2), derrota(-2), desempenhado(2), determinação(3), dispor(2), enrijecer(3), esmagado(-2), espancado(-3), esperançosamente(2), facilidade(2), facilmente(2), festa(2), fofoca(-2), formidável(3), forte(2), ganhou(2), honesto(2), honra(2), incentivo(2), incerto(-2), incessante(2), incrível(4), indiferente(-2), indizível(4), inteligente(3), loucamente(-2), manso(2), melhor(4), merda(-3), multa(-3), negligência(-2), obscurecer(-2), obstinado(2), ok(2), okays(2), oks(2), original(2), paz(3), perdido(-2), perfeito(4), perverso(-3), pomposo(2), porra(-3), predicamento(2), prontidão(2), puro(3), querido(3), quitação(2), regozijar(2), relutante(2), resistente(2), responsável(2), retiro(2), revolução(2), rígida(2), risível(2), sabedoria(2), saciar(2), solene(3), tesouro(2), veemente(3), vigor(3), vital(3), vitória(2), vitórias(2)
BoosterWordList.txt (Retiradas)	assim(0), deve(-1), fez(-1), pode(-1), porra(2), realmente(1), soma(-1)

BoosterWordList.txt (Modificadas)	maldito(-2)
NegatingWordList.txt	acostumado, pelo

Fonte: elaborada pelo autor.

4.8 Execução da Melhoria

Aplicado a melhoria no dicionário em português do método SentiStrength, adicionando as palavras classificadas pelo autor deste trabalho que foram observadas na primeira coleta de *tweets* que faziam menção ao termo “motorola” e atualizando as palavras contidas no dicionário original em português com exclusões e alterações de pesos, foi executada novamente a fase de mineração descrita na seção 4.5, afim de aplicar o dicionário melhorado para as duas coletas realizadas na seção 4.2.

4.9 Análise da Melhoria

Após a execução da mineração com o dicionário melhorado, foi feita uma nova análise, onde foi comparada a análise manual descrita na seção 4.6 com a nova análise feita pelo método SentiStrength com o dicionário melhorado, conforme descrito na seção 4.8.

Ao final dessa etapa, foi possível verificar a nova acurácia do método SentiStrength com o dicionário melhorado, assim como a nova distância do erro das polaridades positivas e negativas na coleta de *tweets* que faziam menção ao termo “motorola”. Na coleta de *tweets* que faziam menção ao termo “liga da justiça” foi possível verificar também a nova acurácia. Os resultados serão apresentados a seguir na seção 5.

5 RESULTADOS

Nesta seção, serão apresentados os resultados obtidos no trabalho.

5.1 Coleta e análise das palavras dos *tweets*

As coletas foram realizadas em Novembro de 2017, iniciando com a coleta dos *tweets* que faziam menção ao termo “motorola”, em seguida, a coleta dos *tweets* que faziam menção ao termo “liga da justiça”.

O autor deste trabalho observou que a forma de escrita dos *tweets* analisados eram semelhantes, assim como as palavras utilizadas. O Quadro 4 mostra as palavras que foram observadas uma ou mais vezes, essas palavras foram comuns entre os *tweets*, e a forma como eram expressadas ajudou ao autor deste trabalho a classificar os seus pesos.

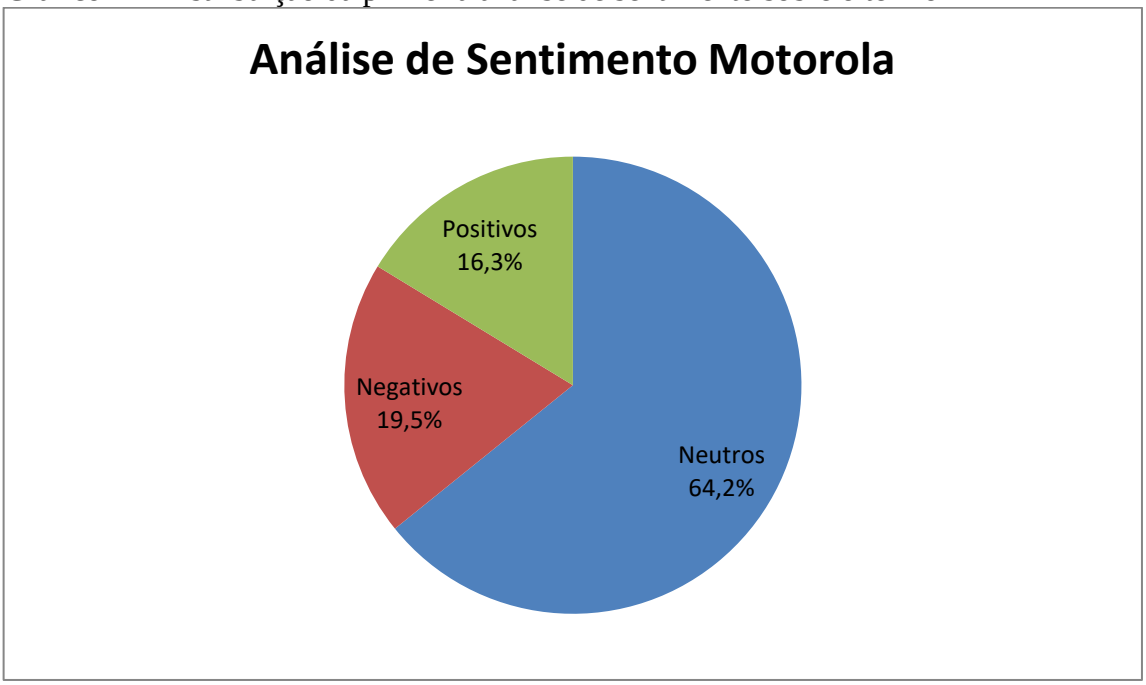
5.2 Acurácia do método SentiStrength com o dicionário em português

O método SentiStrength analisou em dois momentos um total de 2000 *tweets*. No primeiro momento, ele analisou *tweets* que faziam menção aos termos “motorola” (1000 *tweets*) e “liga da justiça” (1000 *tweets*) utilizando o seu dicionário em português sem modificações. No segundo momento, ele analisou *tweets* que faziam menção aos termos “motorola” (1000 *tweets*) e “liga da justiça” (1000 *tweets*) utilizando o seu dicionário em português com melhorias sugeridas pelo autor deste trabalho.

5.2.1 Análise no primeiro momento Motorola

O método SentiStrength com o dicionário em português analisou um total de 1000 *tweets* que fazia menção ao termo “motorola”. O Gráfico 1 mostra os resultados obtidos com o uso do método, onde 163 (16,3% do total de 1000) *tweets* foram analisados como positivos, 195 (19,5% do total de 1000) *tweets* foram analisados como negativos e 642 (64,2% do total de 1000) *tweets* foram analisados como neutros.

Gráfico 1 – Distribuição da primeira análise de sentimento sobre o termo “Motorola”



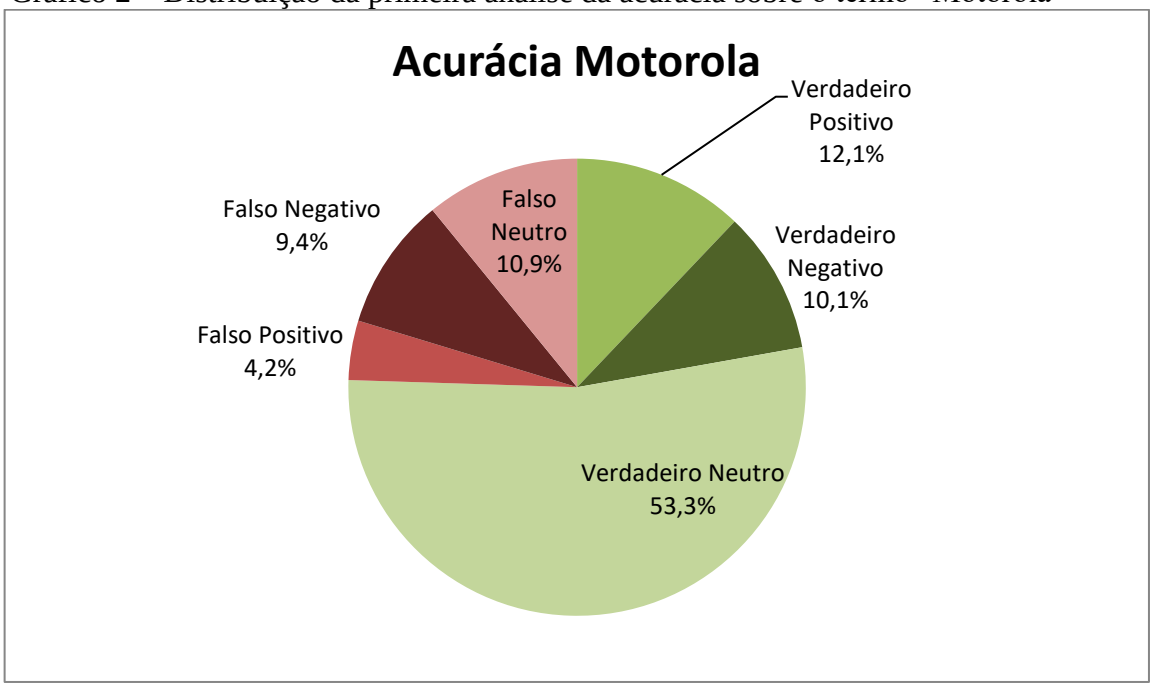
Fonte: elaborada pelo autor.

5.2.1.1 Análise dos resultados no primeiro momento Motorola

Foi verificada uma acurácia de 75,5% do método SentiStrength com o seu próprio dicionário em português sem modificações. Esse dado significa que em comparação com a análise de sentimentos feita manualmente, o método com o dicionário em português acertou o sentimento de 755 *tweets* no total de 1000.

O Gráfico 2 mostra a distribuição dos acertos e erros da análise do método, onde apresentou 12,1% de *tweets* verdadeiramente classificados como positivos, 10,1% de *tweets* classificados verdadeiramente como negativos, 53,3% classificados verdadeiramente como neutros, 4,2% de *tweets* classificados erroneamente como positivos, 9,4% de *tweets* classificados erroneamente como negativos e 10,9% de *tweets* classificados erroneamente como neutros.

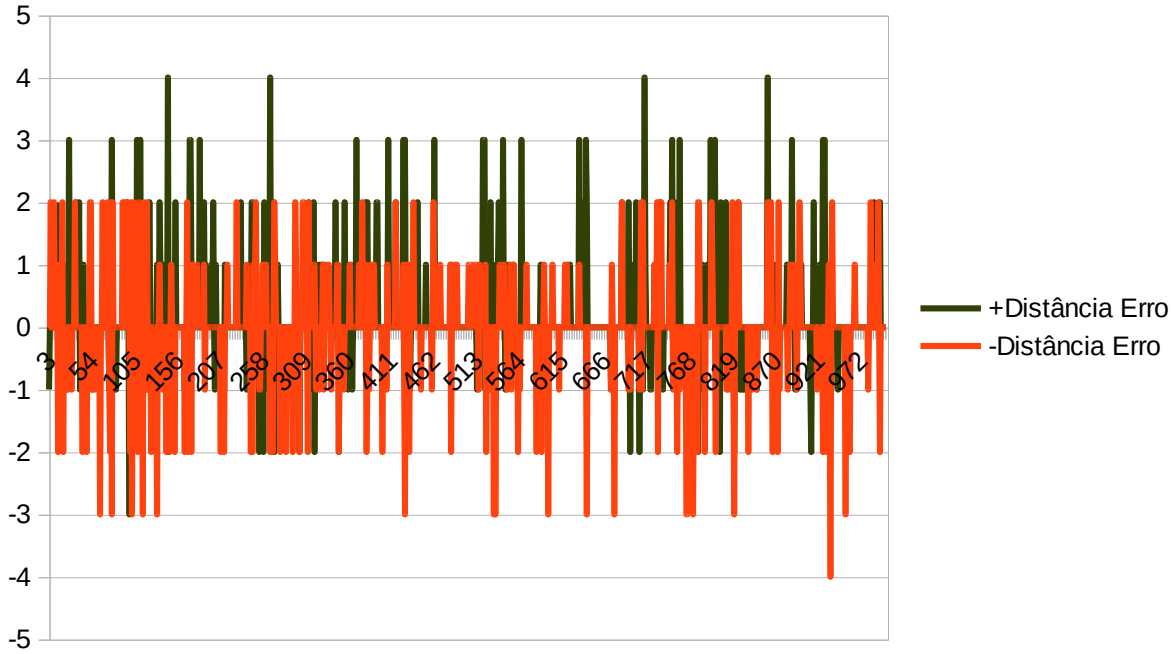
Gráfico 2 – Distribuição da primeira análise da acurácia sobre o termo “Motorola”



Fonte: elaborada pelo autor.

A análise do método obteve uma média de distância de erro de 0,155 para positivos e -0,028 para negativos (quanto mais próximo de zero, melhor). O Gráfico 3 mostra a distância da polaridade dos erros para positivo e para negativo em relação ao que foi analisado manualmente, onde +Distância Erro significa a distância da polaridade positiva analisada pelo método da polaridade positiva analisada manualmente, e –Distância Erro significa a distância da polaridade negativa analisada pelo método da polaridade negativa analisada manualmente.

Gráfico 3 – Distribuição da primeira distância de erros da análise de sentimento sobre o termo “Motorola”



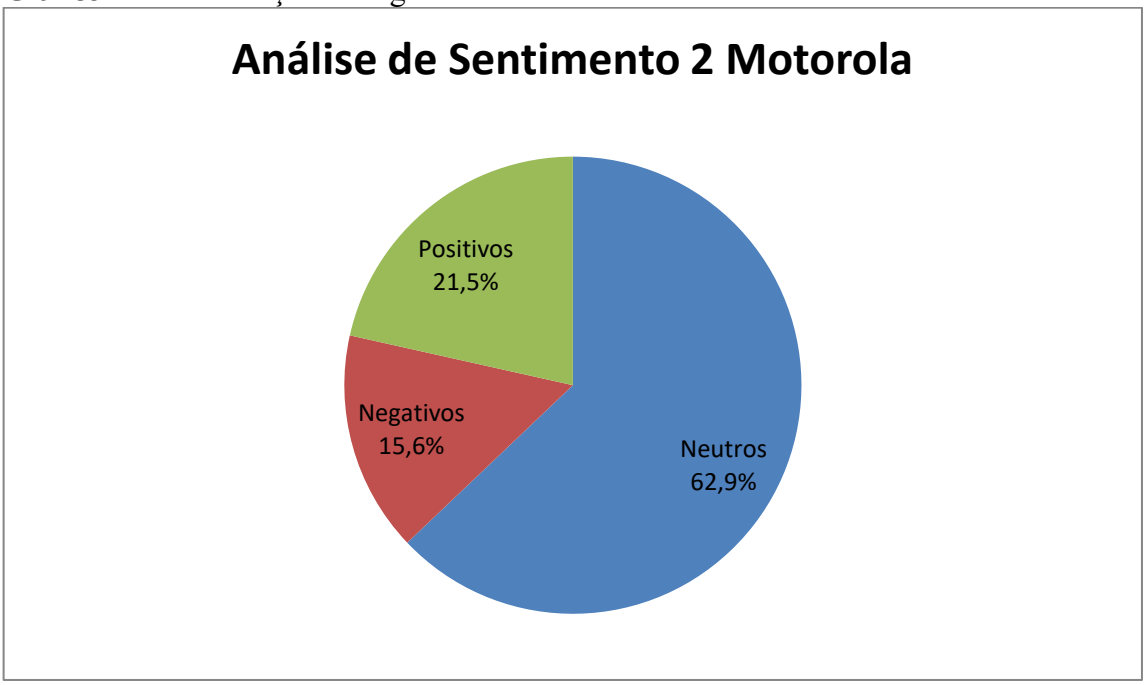
Fonte: elaborada pelo autor.

É possível perceber as distâncias (para mais ou para menos) de erro das polaridades positivas e negativas classificadas pelo método SentiStrength comparadas com as polaridades positivas e negativas classificadas manualmente. As distâncias positivas e negativas que permaneceram em zero foram as que obtiveram classificações idênticas à classificação manual.

5.2.2 Análise no segundo momento Motorola

O método SentiStrength com o dicionário em português melhorado pelo autor deste trabalho, analisou um total de 1000 *tweets* que faziam menção ao termo “motorola”. O Gráfico 6 mostra os resultados obtidos com o uso do método, onde 215 (21,5% do total de 1000) *tweets* foram analisados como positivos, 156 (15,6% do total de 1000) *tweets* foram analisados como negativos e 629 (62,9% do total de 1000) *tweets* foram analisados como neutros.

Gráfico 4 – Distribuição da segunda análise de sentimento sobre o termo “Motorola”



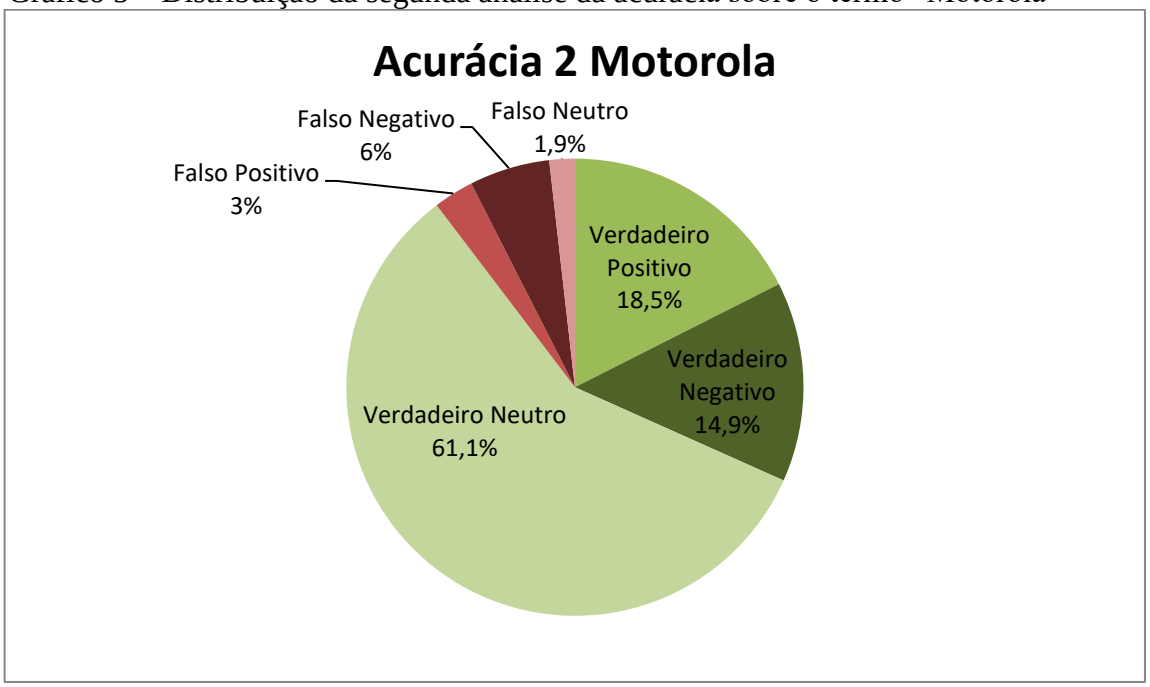
Fonte: elaborada pelo autor.

5.2.2.1 Análise dos resultados no segundo momento Motorola

Foi verificada uma acurácia de 94,5% do método SentiStrength com o dicionário em português melhorado pelo autor deste trabalho. Esse dado significa que em comparação com a análise de sentimentos feita manualmente, o método com o dicionário em português acertou o sentimento de 945 *tweets* no total de 1000, com 190 mais acertos do que comparado com a análise feita com o dicionário original, uma melhora de 19% do total de 1000.

O Gráfico 7 mostra a distribuição dos acertos e erros da análise do método com o dicionário melhorado pelo autor deste trabalho, onde apresentou 18,5% de *tweets* verdadeiramente classificados como positivos, 14,9% de *tweets* classificados verdadeiramente como negativos, 61,1% classificados verdadeiramente como neutros, 3,0% de *tweets* classificados erroneamente como positivos, 0,6% de *tweets* classificados erroneamente como negativos e 1,9% de *tweets* classificados erroneamente como neutros.

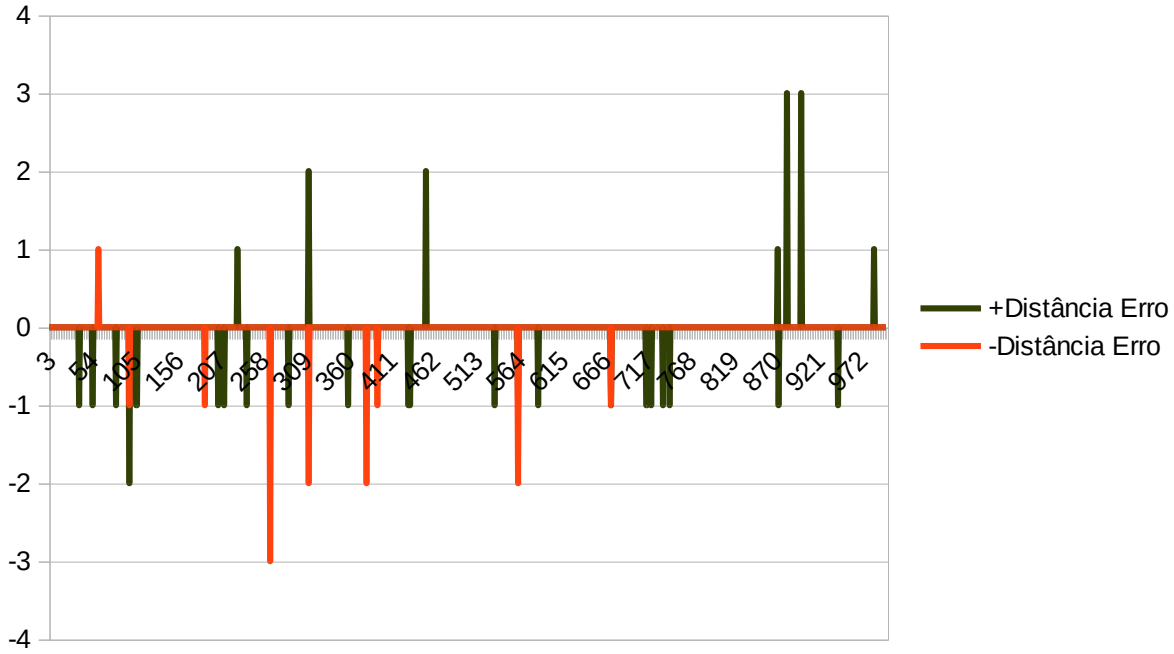
Gráfico 5 – Distribuição da segunda análise da acurácia sobre o termo “Motorola”



Fonte: elaborada pelo autor.

A análise do método obteve uma média de distância de erro de -0,012 para positivos e -0,012 para negativos (quanto mais próximo de zero, melhor). Uma melhora na média de distância do erro de 0,143 para positivos e 0,016 para negativos comparados com a análise feita com o dicionário original. O Gráfico 8 mostra a curva da polaridade dos erros para positivo e para negativo em relação ao que foi analisado manualmente, onde +Distância Erro significa a distância da polaridade positiva analisada pelo método da polaridade positiva analisada manualmente, e –Distância Erro significa a distância da polaridade negativa analisada pelo método da polaridade negativa analisada manualmente.

Gráfico 6 – Distribuição da segunda distância de erros da análise de sentimento sobre o termo “Motorola”



Fonte: elaborada pelo autor.

É possível perceber as distâncias (para mais ou para menos) de erro das polaridades positivas e negativas classificadas pelo método SentiStrength com o novo dicionário comparadas com as polaridades positivas e negativas classificadas manualmente. As distâncias positivas e negativas que permaneceram em zero foram as que obtiveram classificações idênticas à classificação manual.

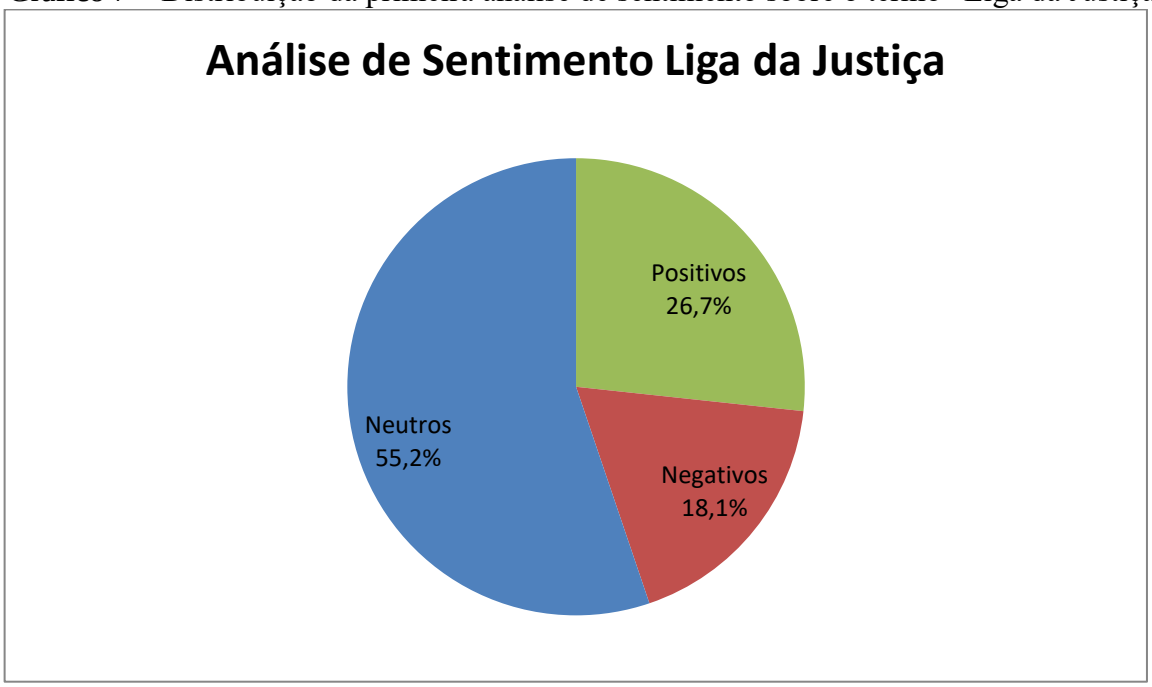
Houve uma melhora nas distâncias (para mais ou para menos) de erro das polaridades positivas e negativas classificadas pelo método SentiStrength com o novo dicionário comparadas com as distâncias (para mais ou para menos) de erro das polaridades positivas e negativas classificadas pelo método SentiStrength com o dicionário original, pois é possível perceber o aumento de distâncias zero com o novo dicionário.

Para garantir que as melhorias encontradas pela modificação do dicionário não estavam acontecendo pelo fato de que as mudanças no dicionário foram motivadas em parte pelas palavras encontradas nos conjunto de *tweets* coletados com a palavra “motorola”, uma nova coleta com outro assunto foi realizada e os dados foram avaliados. O resultado é apresentado a seguir.

5.2.3 Análise no primeiro momento Liga da Justiça

O método SentiStrength com o dicionário em português analisou um total de 1000 *tweets* que fazia menção ao termo "liga da justiça". O Gráfico 4 mostra os resultados obtidos com o uso do método, onde 267 (26,7% do total de 1000) *tweets* foram analisados como positivos, 181 (18,1% do total de 1000) *tweets* foram analisados como negativos e 552 (55,2% do total de 1000) *tweets* foram analisados como neutros.

Gráfico 7 – Distribuição da primeira análise de sentimento sobre o termo "Liga da Justiça"



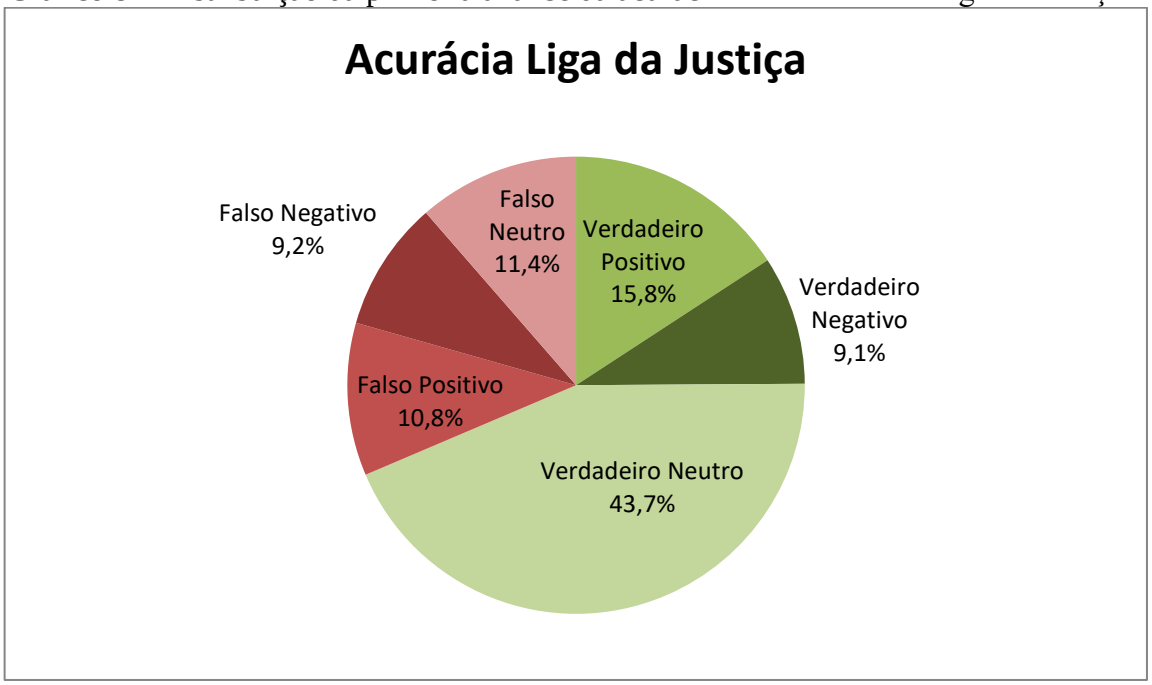
Fonte: elaborada pelo autor.

5.2.3.1 Análise dos resultados no primeiro momento Liga da Justiça

Foi verificada uma acurácia de 68,6% do método SentiStrength com o seu próprio dicionário em português sem modificações. Esse dado significa que em comparação com a análise de sentimentos feita manualmente, o método com o dicionário em português acertou o sentimento de 686 *tweets* no total de 1000.

O Gráfico 5 mostra a distribuição dos acertos e erros da análise do método, onde apresentou 15,8% de *tweets* verdadeiramente classificados como positivos, 9,1% de *tweets* classificados verdadeiramente como negativos, 43,7% classificados verdadeiramente como neutros, 10,8% de *tweets* classificados erroneamente como positivos, 9,2% de *tweets* classificados erroneamente como negativos e 11,4% de *tweets* classificados erroneamente como neutros.

Gráfico 8 – Distribuição da primeira análise da acurácia sobre o termo “Liga da Justiça”

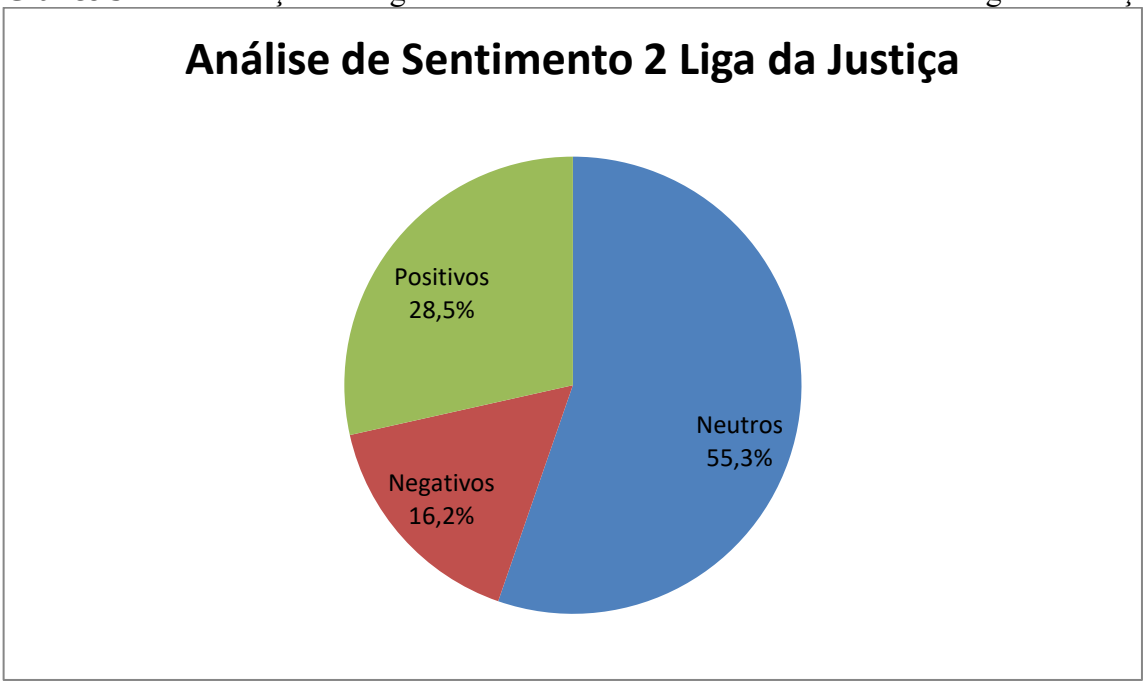


Fonte: elaborada pelo autor.

5.2.4 Análise no segundo momento Liga da Justiça

O método SentiStrength com o dicionário em português melhorado pelo autor deste trabalho, analisou um total de 1000 *tweets* que faziam menção ao termo “liga da justiça”. O Gráfico 9 mostra os resultados obtidos com o uso do método, onde 285 (28,5% do total de 1000) *tweets* foram analisados como positivos, 162 (16,2% do total de 1000) *tweets* foram analisados como negativos e 553 (55,3% do total de 1000) *tweets* foram analisados como neutros.

Gráfico 9 – Distribuição da segunda análise de sentimento sobre o termo “Liga da Justiça”



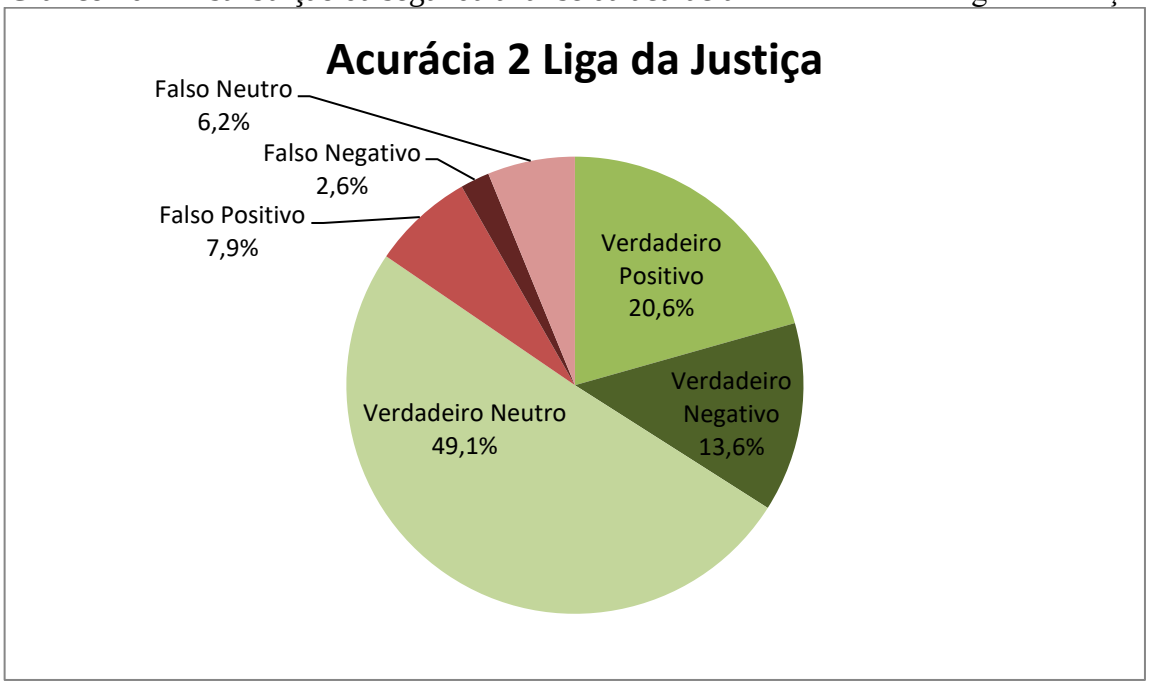
Fonte: elaborada pelo autor.

5.2.4.1 Análise dos resultados no segundo momento Liga da Justiça

Foi verificada uma acurácia de 83,3% do método SentiStrength com o dicionário em português melhorado pelo autor deste trabalho. Esse dado significa que em comparação com a análise de sentimentos feita manualmente, o método com o dicionário em português acertou o sentimento de 833 *tweets* no total de 1000, com 147 mais acertos do que comparado com a análise feita com o dicionário original, uma melhora de 14,7% do total de 1000.

O Gráfico 10 mostra a distribuição dos acertos e erros da análise do método com o dicionário melhorado pelo autor deste trabalho, onde apresentou 20,6% de *tweets* verdadeiramente classificados como positivos, 13,6% de *tweets* classificados verdadeiramente como negativos, 49,1% classificados verdadeiramente como neutros, 7,9% de *tweets* classificados erroneamente como positivos, 2,6% de *tweets* classificados erroneamente como negativos e 6,2% de *tweets* classificados erroneamente como neutros.

Gráfico 10 – Distribuição da segunda análise da acurácia sobre o termo “Liga da Justiça”



Fonte: elaborada pelo autor.

6 CONCLUSÃO

Este trabalho teve como objetivo analisar a eficácia do método SentiStrength com o dicionário em português em *tweets* também em português. Foram realizadas duas coletas de 1000 *tweets* cada, que faziam menção aos termos “motorola” e “liga da justiça”. Ao todo foram coletados 2000 *tweets* em português em Novembro de 2017.

Para análise foram utilizadas métricas de análise de sentimento consideradas mais importantes pela literatura. Os *tweets* classificados pelo método SentiStrength que faziam menção ao termo “motorola” obtiveram uma acurácia de 75,5% com o dicionário original quando comparados a classificação feita manualmente. Os *tweets* classificados pelo método SentiStrength que faziam menção ao termo “liga da justiça” obtiveram uma acurácia de 68,6% com o dicionário original quando comparados a classificação feita manualmente. Foram identificadas e classificadas pelo autor deste trabalho as palavras com maior ocorrência e que davam sentido a classificação de sentimentos na coleta que fazia menção ao termo “motorola”. Foi observado pelo autor que a forma como os *tweets* eram escritos, assim como as palavras utilizadas, eram semelhantes. Foi sugerida uma melhoria no dicionário em português do método SentiStrength com as palavras identificadas na coleta de *tweets* que faziam menção ao termo “motorola” e com a exclusão e alteração dos pesos das palavras contidas no dicionário original. Foi verificada uma melhora de 19% na acurácia da análise dos *tweets* que faziam menção ao termo “motorola”, ficando com 94,5% de acurácia, e uma melhora de 14,7% na acurácia da análise dos *tweets* que faziam menção ao termo “liga da justiça”, ficando com 83,3% de acurácia, confirmando o que diz Benevenuto, Ribeiro e Araújo (2015), que a eficácia do método está diretamente relacionada com a generalização do dicionário.

Devido ao grande tempo gasto com a classificação manual dos *tweets* dando pesos positivos e negativos na mesma frase, imitando o método SentiStrength, apenas duas coletas de 1000 *tweets* foram analisadas.

Assim, foi verificada por este trabalho a eficácia do método SentiStrength com o dicionário em português para *tweets* também em português. O método demonstrou uma melhora significativa com uma pequena sugestão de melhoria com inclusão/exclusão/modificação das palavras e seus respectivos pesos, avaliando assim o dicionário em português do método de análise de sentimentos SentiStrength.

7 TRABALHOS FUTUROS

Foram aplicadas algumas métricas de análise de sentimento neste trabalho que são consideradas como as principais pela literatura, porém, é possível desfrutar de mais métricas para realizar novas análises de sentimento, assim como algumas técnicas de validação estatística.

Foram realizadas duas coletas em contextos diferentes de 1000 *tweets* cada, que faziam menção aos termos “motorola” (fabricante de celulares) e “liga da justiça” (filme). Para uma análise mais ampla é possível coletar outros termos, em outras quantidades, explorando a análise de sentimento em português com o método SentiStrength em outros contextos.

Foi sugerida uma melhoria no dicionário em português do método SentiStrength com palavras observadas na coleta de *tweets* que faziam menção ao termo “motorola” e com exclusões e atualizações nas palavras contidas no dicionário original. É possível aperfeiçoar o dicionário, assim como os pesos das emoções contidas nas palavras com a ajuda de um ou mais linguistas. Os métodos de aprendizagem de máquina também podem ser empregados para a identificação automática de termos que possam ser incluídos no dicionário de forma a melhorar o resultado sem a necessidade de intervenção humana.

O autor deste trabalho observou que a forma de escrita dos *tweets* analisados eram semelhantes, assim como as palavras utilizadas. É possível explorar uma melhoria na abrangência e acurácia do método SentiStrength com o dicionário em português para a classificação de *tweets* também em português com uma coleta maior de *tweets* sobre vários temas, minerar as palavras mais utilizadas, classificá-las com sentimentos (positivos/negativos) e atualizar o dicionário.

Foi realizada uma análise de sentimento com o método SentiStrength com o dicionário em português. Para obtenção de melhores resultados na classificação, é possível aperfeiçoar o método SentiStrength adicionando novas regras no método, tais como: negação após o termo e não apenas antes (exemplo: “gosto não” também ser uma negação equivalente a “não gosto”); Palavras de intensidade após o termo (exemplo: “quero bem muito” também ser uma intensidade equivalente a “quero muito bem”); e redução de seguidos caracteres repetidos para uma possível identificação da palavra.

REFERÊNCIAS

ACIOLI, Sonia. **Redes sociais e teoria social:** revendo os fundamentos do conceito. **Informação & Informação**, v. 12, n. 1esp, p. 8-19, 2007.

ALEXANDRINO, Alex de Oliveira. **Análise de redes sociais aplicada a tweets sobre séries de tv**. 2016. 48 f. TCC (graduação em Sistemas de Informação) - Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2016. Disponível em: <http://www.repositoriobib.ufc.br/000027/00002792.pdf>. Acesso em: 08 nov. 2017.

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português:** Sob o Enfoque da Inteligência Computacional. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2007.

ARAÚJO, Matheus et al. Métodos para análise de sentimentos no twitter. In:**Proceedings of the 19th Brazilian symposiom on Multimedia and the Web (WebMedia'13)**. 2013.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para Análise de Sentimentos em mídias sociais. In: **Brazilian Symposium on Multimedia and the Web (Webmedia)**. Manaus, Brazil. 2015.

CARVALHO FILHO, José Adail. **Mineração de textos: análise de sentimentos utilizando tweets referentes à Copa do Mundo 2014**. 2014. 44 f. TCC (graduação em Engenharia de Software) – Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2014. Disponível em: <http://www.repositoriobib.ufc.br/000017/0000179f.pdf>. Acesso em: 08 nov. 2017.

GOMES, Helder Joaquim Carvalheira. **Text Mining:** análise de sentimentos na classificação de notícias. Information Systems and Technologies (CTISTI), 2013 8th Iberian Conference on. Lisboa. 2013.

GONÇALVES, Pollyanna et al. **Comparing and cobining sentiment analysis mthods**. In: Proceedings of the first ACM conference on Online social networks. ACM, 2013. p. 27-38.

GUZMAN, Emitza; AZÓCAR, David; LI, Yang. Sentiment analysis of commit comments in GitHub: an empirical study. In: **Proceedings of the 11th Working Conference on Mining Software Repositories**. ACM, 2014. p. 352-355.

INDURKHYA, Nitin; DAMERAU, Fred J. **Handbook of natural language processing**. 2ed. Florida: CRC Press, 2010.666 p.

MACHADO, Joicemengue Ribeiro; TIJIBOY, Ana Vilma. Redes Sociais Virtuais: um espaço para efetivação da aprendizagem cooperativa. **RENOTE**, v. 3, n. 1, 2005.

NASCIMENTO, Paula; OSIEK, Bruno Adam; XEXÉO, Geraldo. ANÁLISE DE SENTIMENTO DE TWEETS COM FOCO EM NOTÍCIAS/SENTIMENT ANALYSIS OF NEWS TWEET MESSAGES. **Revista Electronica de Sistemas de Informaçao**, v. 14, n. 2, p. 1, 2015.

REIS, Julio CS et al. Uma abordagem multilıngue para análise de sentimentos**.** In: **IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)**, 2015.

TABOADA, Maite et al. Lexicon-based methods for sentiment analysis. **Computational linguistics**, v. 37, n.2, p. 276-307. 2011.

TAN, Ah-Hwee et al. Text mining: The state of the art and the challenges. In: **Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases**. sn, 1999. p. 65-70.