Bruna Emídia de Assis Almeida Fraga

Análise de Sentimentos nas Redes Sociais

Timóteo

2019

Bruna Emídia de Assis Almeida Fraga

Análise de Sentimentos nas Redes Sociais

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação.

Centro Federal de Educação Tecnológica de Minas Gerais

Campus Timóteo

Graduação em Engenharia de Computação

Orientador: Prof. Dr. Maurílio Alves Martins da Costa

Timóteo

2019

Bruna Emídia de Assis Almeida Fraga

Análise de Sentimentos nas Redes Sociais

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação.

Trabalho aprovado. Timóteo, 01 de Março de 2019:

Prof. Dr. Maurílio Alves Martins da Costa
Orientador
CEFETMG

Prof. Me. Odilon Corrêa da Silva
Professor Convidado
CEFETMG

Prof. Me. Deisymar Botega Tavares
Professor Convidado
CEFETMG

Timóteo
2019

Ao meu companheiro de vida
Breno Rodrigues Silva (In Memorian).

Agradecimentos

Primeiramente agradeço à Deus por nunca me abandonar e me dar a força necessária para continuar a lutar por todos os meus sonhos e projetos.

Ao meu companheiro e exemplo, Breno Rodrigues Silva, responsável pela minha transformação e por quem hoje sou. Por todos os valores e princípios, mas principalmente pela coragem. Esteve sempre ao meu lado, me incentivando a cada etapa e comemorando cada vitória.

À minha família, pelo incentivo dado nos momentos mais difíceis e por me fazer enchergar sempre além do que meus olhos pudessem ver. Obrigada pela crença em meu potencial e por sempre me fazer lembrá-lo, pois nos momentos de dor, angustia e ansiedade desacreditamos de nós mesmos e necessitamos de palavras amigas ao nosso lado. Em especial à minha vó Izabel Emidia de Assis pelo exemplo de dedicação e perseverança que sempre me serviram de lição, aos meus pais Paulo e Irene Fraga, ao meu irmão e companhia diária Herick Fraga, à minha prima pela paciência e presença em minha vida, Maysa Emídia e à todos os meus tios e tias que me proferiram palavras de carinho e força que me possibilitaram prosseguir.

Ao meu professor Dr. Maurílio Alves, sempre serei grata pela dedicação, carinho e amizade ao longo de toda a minha trajetória desde à sala de aula até a realização deste trabalho.

Aos meus amigos e amigas que foram grandes companheiros ao longo da graduação e que são um dos pilares que me sustentam como pessoa, em especial a Elisa Daniel Araújo, Lucas Bourguinon, Jéssica Cardilo e Renan Batista.

Aos meus amigos de turma e de curso pelos momentos que passamos juntos ao longo destes anos e aos professores pelos ensinamentos e por todas as experiências transmitidas à mim que me fizeram evoluir como pessoa e profissional.

E por fim, gostaria de agradecer a todos que direta ou indiretamente me auxiliaram na elaboração deste trabalho.

Resumo

A análise de sentimentos é uma área de estudo computacional que permite o processamento de grandes quantidades de textos para a extração e classificação destes textos em sentimentos.Tendo em vista que as redes sociais são utilizadas diariamente por milhões de usuários para compartilharem ideias, pensamentos, acontecimentos cotidianos e também opiniões de apoio e indignação sobre objetos ou fatos, este trabalho se objetivou a utilizar a análise de sentimentos para extrair o sentimento contido nas mensagens textuais publicadas pelos usuários de uma rede social. Para isto, foi necessário formar uma base de dados a partir das mensagens extraídas de uma rede social, classificar as mensagens nos sentimentos positivo, negativo e neutro através de um método de análise de sentimentos e comparar o resultado obtido com uma classificação feita manualmente. Realizou-se, então, uma pesquisa exploratória e descritiva que utilizou um método de mineração de textos somado à uma tecnica de análise de sentimentos, permitindo a classificação das mensagens publicadas por usuários e evidenciando a viabilidade de se extrair sentimentos de textos publicados em redes sociais.

Palavras-chave: análise de sentimentos; mineração de texto; redes sociais; mineração de opinião.

Abstract

Sentiment analysis is a computational study area that allows large amounts of texts to be processed for extraction and classification of these texts into feelings. Given that social networks are used daily by millions of users to share ideas, thoughts, everyday events and also opinions of support and indignation about objects or facts, this study aimed to use sentiment analysis to extract the feeling contained in the textual messages published by users of a social network. For this, was necessary to form a database with the messages extracted from a social network, classify this messages into positive, negative and neutral feelings through a method of sentiment analysis and compare the results with a manually classification. This was an exploratory and descriptive research that used a text mining method allied with a sentiment analysis technique allowing the classification of messages published by users and showing the feasibility of extracting feelings from texts published in social networks.

Keywords: sentiment analysis; text mining; social network sites; opinion mining.

Lista de ilustrações

Lista de Quadros

Lista de abreviaturas e siglas

API	*Application Programming Interface*
OAuth	*Open Authorization*
AS	Análise de Sentimentos
MO	Mineração de Opinião
MT	Mineração de Textos
RSO	Redes Sociais Online

Sumário

1 Introdução

A humanidade, segundo Russell (2013), precisa ser ouvida, satisfazer a sua curiosidade, se sentir importante e valorizada. Para isso, é necessário que ela estabeleça conexões entre si, compartilhe pensamentos, ideias e experiências. Em meados do século XX, surge a ideia de rede social. A rede é formada por indivíduos que se conectam através das relações sociais já existentes entre eles. Essa conexão faz com que as ações dos indivíduos que pertencem à uma mesma rede, se adaptem de acordo com ela (FERREIRA, 2011). Assim, as pessoas que estão inseridas em grupos familiares, religiosos, etc, crescem e se moldam tendo como espelho aqueles indivíduos com os quais se convive, por isso, o indivíduo se adapta de acordo com os valores e a moral, presentes dentro das relações. Souza e Quandt (2008), afirma na mesma linha que as redes sociais são formadas por indivíduos que partilham valores e objetivos em comum, permitindo a existência de uma estrutura complexa e dinâmica.

As redes sociais estudam as relações humanas em geral e existem independentemente da tecnologia. Porém, a tecnologia evidenciou a estrutura das redes sociais e fez com que as pessoas tomassem consciência de que ela existe. A tecnologia, segundo Russell (2013), têm como propósito melhorar a experiência dos seres humanos, já Ferreira (2011), diz que a tecnologia aproximou os indivíduos e intensificou o contato entre eles, fazendo com que a barreira do espaço físico fosse rompida. A tecnologia trouxe ferramentas de comunicação que permitem a interação das pessoas independente da distância entre elas. Um exemplo dessas ferramentas são as Redes Sociais Online (RSO), capazes de suprir as necessidades dos seres ditas e ainda promover uma comunicação fácil e rápida.

As RSO estão dentro do universo das mídias sociais. Estes dois conceitos aparecem com frequência na literatura, porém possuem significados sutilmente diferentes. As mídias sociais são de acordo com Kaplan e Haenlein (2010) aplicações online que têm como foco o compartilhamento de conteúdos por seus usuários. E as RSO, apesar de possuirem a funcionalidade de compartilhar conteúdos, têm seu foco voltado para o relacionamento e o engajamento entre os usuários de forma que seja estabelecida uma interação virtual entre eles. Portanto, as RSO são um tipo específico de mídias sociais.

As redes sociais online são gratuitas e interativas. Alguns exemplos são: Facebook, com mais de 2 bilhões de usuários, Youtube, Instagram, Twitter, LinkedIn, Pinterest, Google+ entre outros. Elas foram desenvolvidas para que as pessoas criem perfis, estabeleçam uma rede de amigos e publiquem conteúdos. Estes conteúdos podem ser de diversos formatos como links, fotos e vídeos mas principalmente mensagens textuais. E é através das mensagens textuais que os usuários fazem comentários expressando as suas opiniões e sentimentos sobre acontecimentos diários, assuntos do momento ou outras publicações. O conteúdo produzido por um usuário é conhecido como *User-Generated Content* (UGC) e dada a riqueza de informações contidas nas publicações geradas sobre variados temas e assuntos e a significância do volume de dados produzido, os autores Balazs e Velásquez (2016) dizem que é

humanamente impossível compreender toda essa informação em um período de tempo satisfatório, sendo necessários mecanismos para extrair e analisar esses conteúdos.

Surge a partir desta demanda o campo de pesquisa da Análise de Sentimento (AS) ou Mineração de Opinião (MO). Este estudo computacional é segundo Liu (2010), um desafio do processamento de linguagem natural ou mineração de texto. Portanto, apesar dos diversos formatos publicados por usuários de uma rede social, o estudo têm seu foco voltado para análise textual. Ainda segundo o autor, os textos são manifestações linguísticas que podem ser categorizados em fatos ou opiniões e a análise de sentimento estuda através da computação opiniões, emoções e sentimentos expressos em textos.

A análise de sentimento quando aplicada a diversas mensagens publicadas em uma rede social permite que se tenha uma percepção geral da opinião dos usuários da rede em um dado momento sobre um assunto, ou ainda a captura da reação do público, seja ela positiva ou negativa, sobre algum tema.

Portanto, este trabalho busca responder a seguinte questão de pesquisa: Como determinar sentimentos presentes nas mensagens textuais publicadas por usuários de uma rede social?

1.1 Objetivos

1.1.1 Geral

O objetivo geral deste estudo é extrair o sentimento exposto nas mensagens textuais publicadas pelos usuários de uma rede social, utilizando um método de análise de sentimentos.

1.1.2 Específico

Para atender à este objetivo geral, foram elencados os seguintes objetivos específicos:

1. Extrair dados de uma rede social possibilitando a criação de uma base de dados robusta.
2. Classificar as mensagens extraídas utilizando um método de análise de sentimentos.
3. Comparar os resultados obtidos através da análise de sentimentos com os resultados de uma classificação manual.

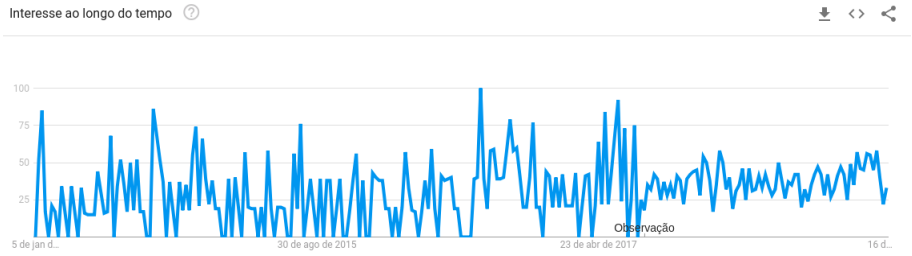
1.2 Motivação

As redes sociais online possuem milhões de usuários, no Brasil por exemplo, 62% da população já é usuária das redes sociais (SOCIAL; HOOTSUIT, 2018). Todas essas pessoas reunidas em um ambiente online, interagindo entre comentários e curtidas produzem um grande volume de informações e dados. Tendo em vista que estas pessoas se comunicam e opinam sobre os mais variados temas e assuntos, as redes sociais passam a ser uma grande fonte de opiniões e sentimentos sobre diversos temas.

Para Pang, Lee et al. (2008) o interesse ao redor da opinião faz parte do processo de tomada de decisão. Segundo os autores, este é um processo natural que pertence à cada uma das pessoas que quando se depara com uma situação difícil ou incomum, tende a buscar opiniões daqueles considerados mais experientes. Este processo de tomada de decisão pertence também às empresas que precisam ser cada vez mais assertivas. Sob a visão de Liu (2010), a opinião pública se torna importante por ser capaz de influenciar tanto pessoas como empresas, em suas crenças e ações.

Um exemplo comum deste interesse ao redor da opinião é o aumento nos últimos anos da busca por resenhas de produtos na rede social Youtube, como mostra a Figura 1. Nessas resenhas, as pessoas falam sobre as vantagens, desvantagens e das suas experiências com o produto. Então antes de fazer uma compra, o usuário pode pesquisar sobre o produto na rede e ver o que está sendo falado sobre ele. Assim, é possível descobrir se ele realmente vale a pena através das opiniões positivas e negativas ditas sobre ele. Os vídeos e propagandas da marca deixaram de ser suficientes, as pessoas estão em busca de relatos que mostrem a realidade. Assim, as RSO expõe para as pessoas as informações necessárias para que elas decidam realizar ou não a compra.

Figura 1 – Evolução de busca do termo ’resenha’ no Youtube ao longo de 5 anos



Fonte: Google (2019)

Porém, devido a quantidade de opiniões expostas sobre os temas, acaba se tornando inviável para o ser humano conseguir compreender todas as informações das redes e tirar alguma conclusão em tempo hábil. Então, a computação através da análise de sentimentos, consegue processar grandes quantidades de textos e classificar estes textos em sentimentos positivos, negativos e também neutros. Assim, a partir do resultado obtido as pessoas ou empresas podem reagir de alguma forma, tomando uma decisão ou ação.

Com a possibilidade de se obter dados opinativos sobre diferentes temas, a análise de

sentimentos passa a ser uma alternativa barata, não invasiva, rápida e autêntica às pesquisas de opinião convencionais (MALHEIROS; LIMA, 2013). De acordo com os autores, a análise de sentimento processa rapidamente milhares de mensagens textuais que já foram publicadas de forma espontânea no ambiente online e também é um método impessoal que não requer materiais impressos.

Os dados contidos dentro das redes sociais podem ser explorados de forma gratuita, através das *Application Programming Interface* (API). Geralmente, as RSO liberam o acesso a dados públicos por meio desta interface tornando assim possível capturar e trabalhar com estas informações.

Diante dos fatos apresentados, este estudo se motivou na necessidade de se determinar o sentimento contido em uma mensagem textual de forma computacional, através de um algoritmo de classificação automática de análise de sentimentos. Devido à existência de um grande volume de dados, que pode ser acessado gratuitamente e trazer informações novas a fim de auxiliar pessoas e empresas. Para isso, é necessário descobrir o processo de analisar sentimento como um todo. Por meio dos trabalhos já publicados, percebe-se a necessidade de um conhecimento prévio de redes sociais e extração de dados, porém, com as informações dispersas o estudo torna-se moroso.

1.3 Estrutura

A presente monografia encontra-se estruturada em sete capítulos. Os tópicos a seguir explicitam cada um deles resumidamente:

- Este primeiro capítulo propôs situar o leitor sobre o contexto no qual este trabalho se encontra, apresentando o problema e as motivações que serviram de incentivo para a busca da solução do mesmo.
- O Estado da Arte apresenta trabalhos relacionados com o problema proposto.
- No capítulo 3, encontram-se os fundamentos e conceitos necessários para o entendimento e desenvolvimento desta monografia.
- O quarto capítulo, apresenta os métodos e passos utilizados para se alcançar a resolução do problema proposto.
- Os resultados obtidos estão detalhados no capítulo 5.
- No capítulo 6 é apresentada a conclusão
- E o último capítulo, Trabalhos Futuros indica melhorias para a continuidade deste.

2 Estado da Arte

A ascensão dos problemas de análise de sentimentos levou à publicação de um grande volume de pesquisas (PANG; LEE et al., 2008). As áreas de destaque destas publicações foram: a de saúde (CHEE; KARAHALIOS; SCHATZ, 2009), (CHEW; EYSENBACH, 2010) (VYDISWARAN; ZHAI; ROTH, 2011); área empresarial, com o monitoramento do sentimento do público em relação às marcas e produtos (EVANGELISTA; PADILHA, 2013) (CHAMLERTWAT et al., 2012) (GHIASSI; SKINNER; ZIMBRA, 2013), social (NASCIMENTO et al., 2012) (MALHEIROS; LIMA, 2013) (GRASSI et al., 2011)(HASSAN; QAZVINIAN; RADEV, 2010) e política (MALINI; CIARELLI; MEDEIROS, 2017) (STIEGLITZ; DANG-XUAN, 2012).

Observa-se nas publicações uma grande variedade de técnicas aplicadas para um mesmo fim: a extração e análise de sentimentos a partir de textos. Em Medhat, Hassan e Korashy (2014), 54 artigos do campo de pesquisa da análise de sentimentos foram coletados e as técnicas utilizadas por eles estão apresentadas na figura 2 em forma de fluxograma. Como pode ser visto na imagem, a análise de sentimentos pode ser realizada a partir de duas abordagens principais: detecção baseada em aprendizado de máquina ou baseada em análise léxica.

O aprendizado de máquina pode ser definido como um conjunto de métodos capazes de reconhecer padrões a partir de um conjunto de dados e fazer uma predição futura (MURPHY, 2012). Os métodos de aprendizado de máquina em análise de sentimentos, utilizam algoritmos de aprendizado em uma base textual. Esses algoritmos ainda podem seguir por duas vertentes, a do aprendizado supervisionado ou não supervisionado. Na aprendizagem supervisionada, a resposta desejada já é conhecida e por este motivo é possível calibrar os pesos para melhorar a assertividade do algoritmo, fazendo com que ele retorne o que se espera. Já na aprendizagem não supervisionada, não se têm ideia prévia da resposta do problema, porém ele é utilizado para explorar os dados e se fazer análises. No aprendizado supervisionado, quatro tipos de classificadores foram mapeados por Medhat, Hassan e Korashy (2014): Árvore de decisão, linear, baseado em regras e probabilísticos .

Para a execução do método de aprendizado de máquina na análise de sentimentos, são necessárias bases de treinamento e teste previamente classificadas em sentimentos para que o algoritmo consiga aprender e classificar textos desconhecidos (BENEVENUTO; RIBEIRO; ARAÚJO, 2015). Alguns destes algoritmos utilizados nesta área são:

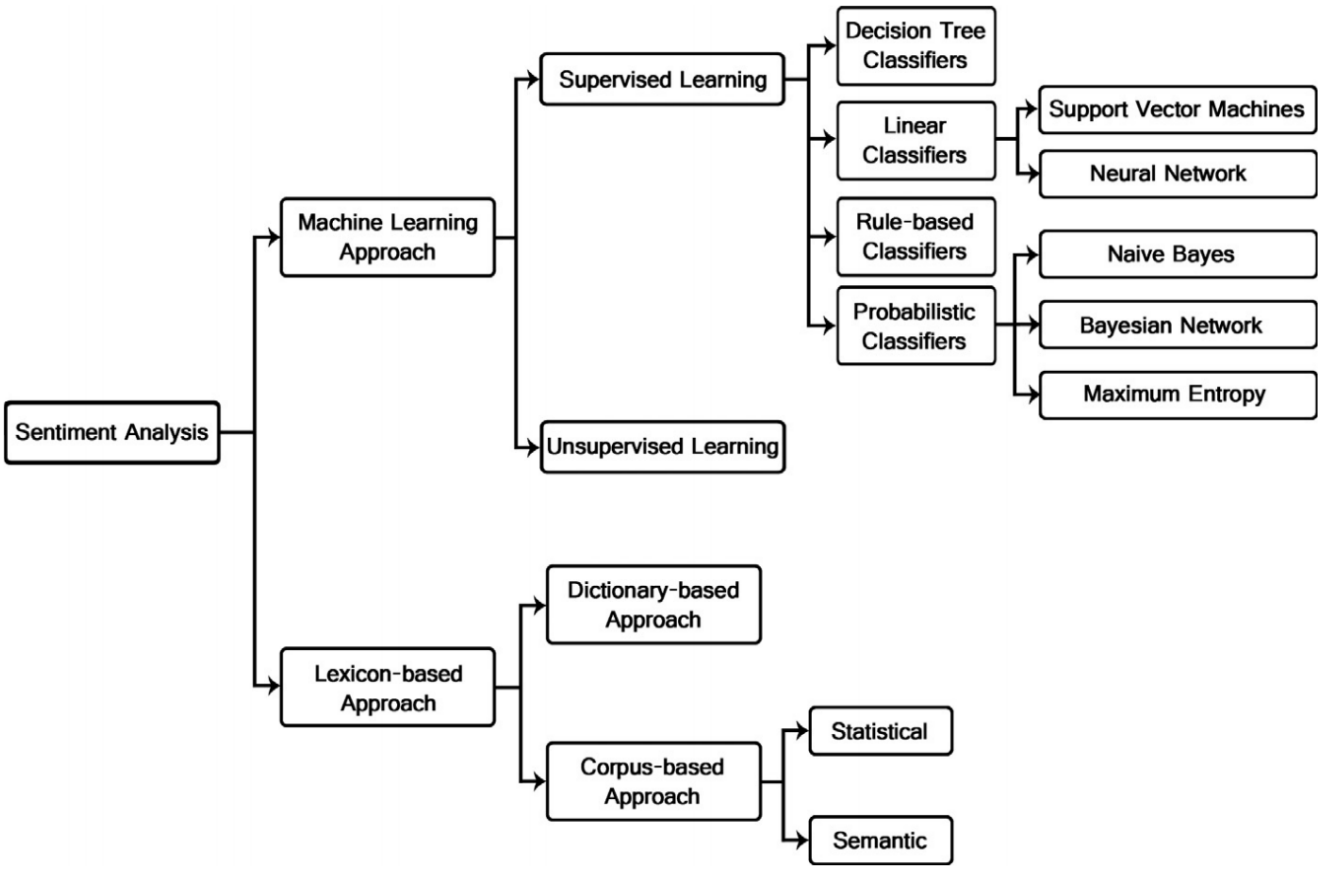
- Support Vector Machines (SVM): O classificador SVM é um modelo de aprendizado de máquina para reconhecimento de padrões introduzido por (CORTES; VAPNIK, 1995).
- Neural Network : Modelo proposto baseado nas redes neurais biológicas presentes no cérebro humano capazes de reconhecer padrões e realizar aprendizados.
- Bayesian Network: Modelos gráficos de dependencia probabilística baseados na incerteza pelo teorema de Bayes, onde os nós representam as variáveis e os arcos a relação

entre eles (KORB; NICHOLSON, 2010).

- Naive Bayes (NB): Classificador probabilístico baseado no teorema de Bayes. É a forma mais simples da rede bayesiana onde todos os nós de atributo são independentes e se relacionam apenas com o nó de classe (ZHANG, 2004).
- Maximum Entropy (MaxEnt): Classificador probabilístico que uniformiza a distribuição da probabilidade dos dados quando o cenário não é conhecido (NIGAM; LAFFERTY; MCCALLUM, 1999).

A outra abordagem para a análise de sentimentos é a análise léxica, que segundo Benevenuto, Ribeiro e Araújo (2015), relaciona valores quantitativos ou qualitativos às palavras para atribuir-lhes sentimento. Assim, é formado um conjunto de palavras opinativas que compõe um dicionário léxico. Nesta linha de abordagem, podem ser utilizados também padrões sintáticos para reconhecer novas palavras além daquelas que já existem no conjunto (MEDHAT; HASSAN; KORASHY, 2014).

Figura 2 – Técnicas de Classificação de Sentimentos



Fonte: Medhat, Hassan e Korashy (2014)

Após a execução destes métodos a análise de sentimentos obtêm como resultado uma classificação do sentimento que pode envolver três aspectos: polaridade, emoção e força do

sentimento (BENEVENUTO; RIBEIRO; ARAÚJO, 2015). Segundo os autores, a saída comum dos métodos é a polarizada. A polaridade indica uma orientação que pode ser positiva, negativa ou neutra, assim, o texto analisado terá uma saída que pode ser binária (positiva ou negativa) ou terciária (positiva, negativa e neutra). Porém, alguns métodos são ainda mais específicos e classificam os textos em emoções. O aspecto da emoção, segundo Liu (2010), já indica sentimentos subjetivos específicos que podem ser de seis tipos de acordo Parrott (2001): amor, alegria, surpresa, raiva, tristeza e medo. As emoções podem ser divididas em muitas outras e apresentar diferentes intensidades (LIU, 2010). Outros métodos utilizam o aspecto da intensidade ou força do sentimento , segundo Benevenuto, Ribeiro e Araújo (2015), que pode ser tanto de um sentimento quanto de uma emoção. Sua saída é um valor numérico de um intervalo (-1 a 1, -inf a +inf, etc) que representa a força do sentimento de um texto.

2.1 Análise de sentimentos na academia

O trabalho de Souza (2015) criou um método para identificar usuários influentes da rede social Twitter e classificar automaticamente a categoria na qual o usuário é influente, baseando-se nos conteúdos postados. Os conteúdos poderiam pertencer a três categorias: política, esporte ou outros assuntos. Para a classificação das postagens, utilizou-se técnicas de mineração de texto e o método de aprendizado supervisionado Naive Bayes. Apesar de não se tratar da classificação de sentimentos, este trabalho utilizou métodos muito semelhantes para atingir o objetivo e encontrar as categorias desejadas. Como resultado, a autora conclui que o processo de classificação das mensagens foi eficaz apenas para a categoria política.

Em Cavalin et al. (2014), foi desenvolvido um sistema de tempo real capaz de analizar o sentimento das mensagens dos usuários do Twitter. Ao longo da pesquisa foram analizadas as mensagens durante os jogos da Copa das Confederações de 2013. Quanto à análise de sentimentos a abordagem de aprendizado de máquina foi utilizada e o classificador probabilístico de Naive Bayes polarizou as mensagens em sentimentos positivo, negativo e neutro.

Um sistema em tempo-real foi desenvolvido para acompanhamento da eleição federal Irlandesa do ano de 2011, por Bermingham e Smeaton (2011). A técnica utilizada foi a abordagem aprendizado de máquina. Para isso, a base de treino foi preparada manualmente por anotadores que classificaram as mensagens em sete grupos. Porém, foram considerados para o treinamento somente os grupos positivo, negativo e neutro. Dentre os algoritmos de classificação presentes estão *Support Vector Machines* (SVM) e *Multinominal Naive Bayes* (uma variação de Naive Bayes que estima a probabilidade das palavras de acordo com a sua frequência). Um de seus resultados foi o mapeamento de sentimentos durante as eleições. Notou-se que alguns dias antes da votação, houve uma variação significativa dos sentimentos públicos em relação aos partidos. Essa variação pode ser observada até algumas semanas após a votação (BERMINGHAM; SMEATON, 2011).

Wang et al. (2012) também desenvolveu um sistema em tempo real que monitorou as eleições presidenciais americanas de 2012 no Twitter. Em seu estudo foram considerados quatro tipos de sentimentos: negativo, positivo, neutro e incerto. Para formar as bases de treino e teste do modelo Naive Bayes (NB) unigrama, uma série de mensagens foram classificadas

por voluntários.

Go, Bhayani e Huang (2009), formaram a base de dados com tweets diversos sem uma definição de tema. Para a formação da base de treinamento e aplicação dos algoritmos de aprendizado de máquina, os autores capturaram tweets que contivessem emoticons [1]. Assim, rapidamente conseguiriam uma base robusta já categorizada em dois sentimentos positivo e negativo. Porém, no momento do treinamento, os emoticons foram retirados dos textos para não influenciarem os algoritmos de classificação e atrapalhar a acurácia deles. Os autores testaram classificadores baseados em palavras-chave, *Naive Bayes*, *Maximum Entropy*, *Support Vector Machines*. E o melhor acerto obtido foi de 83% para o *Maximum Entropy* quando combinado a unigramas e bigramas, em comparação ao resultado alcançado pelos outros algoritmos classificadores usados.

Em Gonçalves et al. (2012), uma nova escala psicométrica para o Twitter foi proposta tendo como base uma vesão anterior da escala conhecida como PANAS-x (*Positive Affect Negative Affect Scale*). A escala PANAS-x é composta por um acervo de palavras que se relacionam com um estado afetivo, ou emoção. O propósito da nova escala denominada pelos autores, PANAS-t, seria categorizar textos publicados por usuários do Twitter sobre eventos diversos, em 11 sentimentos: (i) Medo, (ii) Hostilidade, (iii) Culpa, (iv) Tristeza, (v) Jovialidade, (vi) Auto-Confiança, (vii) Atenção, (viii) Timidez, (ix) Fadiga, (x) Serenidade e (xi) Surpresa. Para isso, foram utilizados cerca de 1,8 bilhões de tweets. Esta escala psicométrica juntamente com mais sete outros métodos (*SentiWordNet, SASA, Emoticons, SentiStrength, LIWC, SenticNet, e Happiness Index*) foram comparados em Araújo et al. (2013). As métricas avaliadas foram: abrangência, concordância, positivo verdadeiro, negativo verdadeiro, acurácia (taxa em que o método acerta o sentimento), precisão e F-measure. Os autores chegaram a conclusão de que os métodos apresentam diferentes graus de abrangência e acurácia, portanto não existe um método que é sempre melhor. Porém, de acordo com o autor, um conjunto de métodos pode suprir a necessidade do usuário e apresentar melhores resultados.

Existem ferramentas na literatura capazes de polarizar textos automaticamente como: *SentiWordNet, Sentiment140, SailAil Sentiment Analyzer(SASA), SenticNet, iFeel* (ARAÚJO et al., 2013). Em Teixeira e Azevedo (2011), os autores utilizaram o SentiWordNet [2], um recurso léxico para a mineração de opinião para a análise de sentimento nas redes Twitter e Facebook. O objetivo era utilizar as informações destas redes para verificar a percepção dos consumidores quanto à alguns produtos e relacionar esta opinião com os seus valores de mercado. Os produtos analisados foram filmes ainda em cartazes, devido à movimentação das redes proporcionada por eles com uma grande probabilidade de conter opiniões explícitas em suas mensagens. Como resultado, conseguiram relacionar o percentual de mensagens positivas e negativas com os valores de bilheteria alcançados na semana de estreia. Attux (2017) também utilizou uma ferramenta de classificação automática, Sentiment140, para realizar uma predição eleitoral para a presidência do Brasil no ano de 2014 tendo como base informações do Twitter. Segundo ele, os resultados obtidos podem ser considerados aceitáveis e bem parecidos com

[1] *"O termo emoticon é a aglutinação de duas palavras inglesas: emotion (emoção) e icon (ícone), e consiste na utilização de caracteres gráficos para representar emoções humanas"* (SIGNIFICADOS, 2018, p.1).

[2] http://sentiwordnet.isti.cnr.it/

o real resultado das eleições.

3 Fundamentos históricos e teóricos

O capítulo que se inicia objetiva-se a introduzir e apresentar conceitos cruciais para o entendimento deste trabalho utilizados na resolução do problema em questão.

3.1 Redes Sociais

O foco deste trabalho será as ferramentas de relacionamento que nas palavras de Ferreira (2011) podem ser consideradas manifestações especiais e particulares de algumas redes sociais.

3.1.1 As redes sociais Online

As RSO oferecem um serviço que permite a comunicação das pessoas através da internet. A definição dessas plataformas pode ser expressa como serviços baseados na web no qual as pessoas podem criar perfis em um sistema, podendo ser eles públicos ou semi-públicos (BOYD; ELLISON, 2007). Ainda de acordo com o autor, através desse serviço os usuários se apresentam, conectam-se uns aos outros e podem visualizar e explorar a tanto à sua lista de conexões como as de outras pessoas. O principal objetivo das plataformas sociais é formar e tornar visíveis as conexões entre pessoas.

Pode-se dizer que o principal objetivo das redes sociais é conectar pessoas virtualmente. Por mais que as redes ofereçam um ambiente propício e toda uma estrutura para se conhecer novas pessoas, as conexões estabelecidas na rede social são muitas vezes um reflexo de uma relação pré-existente no mundo real. A pesquisa de Ellison, Steinfield e Lampe (2007) mostrou que os usuários utilizavam o Facebook[1] com mais frequência para manter ou intensificar as relações reais do que procurar por novos amigos. Para transformar um relacionamento real em virtual não é necessário que haja um vinculo forte entre as pessoas (BOYD; ELLISON, 2007). O simples fato de se conhecerem, terem amigos em comum ou apenas terem se visto uma única vez, é suficiente para que elas se reconheçam e se aceitem virtualmente.

Apesar de possuírem um objetivo central em comum, as redes se diferem em alguns aspectos. Cada rede social disponibiliza recursos e funcionalidades de acordo com o seu propósito. Por exemplo, o Facebook é uma RSO que têm como propósito a Amizade. Já o Pinterest[2] e o Instagram[3] incentivam seus usuários à compartilharem imagens. Com foco nos negócios a rede LinkedIn[4] permite que seus participantes se apresentem de forma profissional revelando aptdões e habilidades, compartilhem conteúdos relacionados à temas profissionais relacionados ao trabalho e favorece o *networking*: estabelecimento de uma rede de contatos

[1] https://www.facebook.com
[2] https://br.pinterest.com
[3] https://www.instagram.com
[4] https://br.linkedin.com

profissional entre os indivíduos. Em uma outra frente, têm-se o Twitter, uma rede que foca em conteúdos e notícias rápidas.

Essas características únicas provocam no público graus distintos de afinidade, portanto, cada rede social atrai um grupo específico que mais se identifica com o que a rede propõe (KAPLAN; HAENLEIN, 2010).

3.1.2 Elementos das redes sociais

As plataformas sociais possuem alguns elementos chaves que as caracterizam como uma rede social. Segundo Benevenuto, Almeida e Silva (2011) foram identificados nas principais plataformas elementos relevantes que serão explorados à seguir:

3.1.2.1 Perfis

Assim que uma conta é criada em uma ferramenta social, o passo seguinte é o preenchimento do perfil. Todo membro, possui este espaço individual. Nele o usuário insere informações básicas como: nome, idade, sexo, localização etc. Descreve a si mesmo, apresentando as suas características e interesses para a visualização daqueles que farão parte da sua rede. Há também no perfil, listas de amigos e/ou seguidores. Assim que o cadastro do usuário é concluído, ele pode convidar pessoas conhecidas para fazer parte da sua rede. Alguns dos termos utilizados pelas ferramentas são amigos, contatos ou fãs. Para estabelecer uma amizade, a maioria das RSO requerem uma confirmação bidirecional (BOYD; ELLISON, 2007). Essa confirmação ocorre quando uma solicitação de amizade enviada por um usuário, precisa ser aceita por outro. Já as conexões unidirecionais, conhecidas por fãs, ou seguidores não precisam de aceite.

Os amigos ou seguidores, são os atores com os quais pode-se interagir dentro deste ambiente. As interações são diversas e variam conforme as funcionalidades disponibilizadas pela rede. Alguns exemplos são: ações curtir, cutucar, comentar e compartilhar. As informações criadas pelos usuários são chamadas publicações geralemente compostas por textos, imagens, vídeos, gifs, enquetes etc.

De acordo com Boyd e Ellison (2007), a visibilidade que o perfil de um usuário possui depende de alguns fatores como: regras do próprio site ou discrição do usuário. Alguns sites já possuem perfis abertos de forma que sejam acessíveis para qualquer pessoa, seja ela usuária da ferramenta ou não. Outros, disponibilizam tipos de contas que possuem níveis. Esses níveis variam de acordo com a assinatura escolhida pelo usuário. Dessa forma, a visualização fica atrelada às diferentes contas que determinam o que o usuário poderá ver ou não. Sites como o Facebook, permitem que os usuários vejam os perfis daqueles que estão em sua rede de amigos. Porém, caso seja de interesse deles, há a possibilidade de negar a permissão de visualização para determinadas pessoas. O Twitter possui por *default* o perfil público, o que significa que as informações nele contidas são visíveis para todos os membros da rede. Porém, há a opção de torná-lo privado caso o usuário deseje. O foco do Twitter são as mensagens transmitidas entre utilizadores, sendo o perfil algo secundário (TEIXEIRA; AZEVEDO, 2011).

3.1.2.2 Atualizações

As redes sociais normalmente têm uma página inicial, responsável por reunir as atualizações da rede de amigos ou fãs do usuário e por serem imediatas, as publicações se tornam visíveis à rede logo após o compartilhamento. As atualizações segundo, Benevenuto, Almeida e Silva (2011), permitem que os usuários descubram novos conteúdos, os encorajam a interagir com as publicacões já compartilhadas e também incentivam os usuários a produzirem os seus próprios conteúdos.

3.1.2.3 Comentários e Avaliações

Os comentários e avaliações são funcionalidades das redes sociais que promovem interações entre os usuários no ambiente online. Os comentários são espaços em que os usuários expressam sua opinião sobre uma publicação. Geralmente o comentário é textual, porém algumas RSO permitem a adição de imagens, gifs entre outros recursos. Os comentários podem ser adicionados na maioria das publicações compartilhadas (BENEVENUTO; ALMEIDA; SILVA, 2011). Enquanto as avaliações são uma forma de os usuários demonstrarem a sua opinião e sentimento relacionado à publicacão através de recursos já existentes, como por exemplo o botão 'gostei'. No Instagram quando o usuário gosta da publicação é possível curtir com um coração, já o Youtube, disponibiliza duas opções: o gostei (*like*) e o não gostei (*dislike*). No Facebook essas avaliações foram aprimoradas com o tempo e o botão que antes possibilitava somente curtir agora disponibiliza mais cinco emoções: o amor , alegria, surpresa, tristeza e raiva.

3.1.2.4 Listas de Favoritos e Listas de Mais Populares (Top Lists)

Tanto a lista de favoritos como a lista de mais populares permitem a seleção do conteúdo publicado. No caso da lista de favoritos, o próprio usuário adiciona à ela os conteúdos mais relevantes para si. Já nos *Top Lists* as redes sociais identificam as publicações de maior engajamento (promovem maior interação), que são aquelas que possuem grande número de visualização, curtidas ou avaliações e comentários (BENEVENUTO; ALMEIDA; SILVA, 2011).

3.1.2.5 Metadados

Segundo Dijck (2013), os metadados são informações estruturadas que viabilizam ou facilitam a recuperação da informação. São exemplos de metadados de acordo com Benevenuto, Almeida e Silva (2011), título, descrição e tags que podem ser adicionados à publicação pelo próprio usuário ou ser obtidos pelo sistema de forma automática. No Instagram, Facebook e Twitter é comum o uso de *tags* denominadas *hashtags*. As *hashtags* têm como objetivo agrupar todo o conteúdo produzido sobre um assunto ou tema através de palavras-chave. Assim, a busca pela *hashtags* devolve tudo o que foi produzido utilizando a palavra-chave buscada.

3.1.3 As redes sociais no Brasil

Um relatório divulgado pela agência *We are Social* em conjunto com a plataforma *Hootsuit* revelou como é a atividade dos brasileiros quando se trata de internet (SOCIAL;

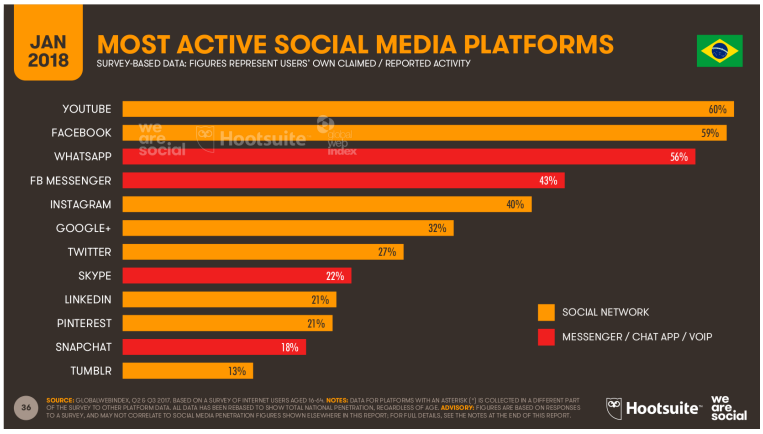
HOOTSUIT, 2018). As informações utilizadas neste relatório são provenientes de diversas empresas como Google, Statcounter, Ookla, Alexa e Ericsson que capturaram dados durante o ano de 2017 e forneceram a agência.

O Brasil já possuia em 2018, 139,1 milhões de usuários da internet e destes, 130 milhões eram membros de redes sociais. O tempo médio gasto por dia pelos brasileiros na internet colocou o Brasil como terceiro colocado no ranking mundial: Foram 9 horas e 14 minutos gastos com acesso à internet a partir de qualquer dispositivo e 3 horas e 39 minutos gastos somente com as redes sociais.

Foi relatado que 85% dos usuários usaram a internet todos os dias e que 60% dos usuários acessaram a internet somente através dos smartphones.

Na figura 3 abaixo, foram listadas as doze plataformas de mídias sociais mais utilizadas no Brasil. Nesta imagem as redes sociais estão apresentadas de amarelo enquanto as plataformas de mensagens instantâneas estão de vermelho. Assim, as cinco maiores redes sociais entre os brasileiros foram: Youtube, Facebook, Instagram, Google+ e Twitter.

Figura 3 – Plataformas mais usadas no Brasil



Fonte: Social e Hootsuit (2018)

3.1.3.1 Youtube

O Youtube foi fundado em fevereiro de 2005, nos Estados Unidos por três ex-funcionários do pay-pal: Chad Hurley, Steve Chen e Jawed Karim. A plataforma permite o compartilhamento de vídeos por seus usuários e a interação deles através de curtidas e comentários nas publicações. Em 2006, foi comprado pela Google por US$ 1,65 bilhão.

A missão do Youtube é "dar a todos uma voz e revelar o mundo"(YOUTUBE, 2019), seus valores baseiam-se,de acordo com a própria companhia, na liberdade de expressão, direito à informação, direito à oportunidade e liberdade de pertencer. O Youtube está presente em 88 países e disponível em 76 idiomas. A versão em português chegou o Brasil em junho de 2007 e em 2018 a plataforma atingiu a marca de 1,8 bilhão de usuários ativos, se tornando a segunda maior rede social do mundo.

3.1.3.2 Facebook

O Facebook foi criado pelos alunos da Universidade de Harvard Mark Zuckerberg, Dustin Moskovitz, Eduardo Saverin e Chris Hughes em 4 de fevereiro de 2004. A princípio a rede era limitada aos estudantes desta universidade, mas acabou se expandindo com o tempo e se tornando aberto à população com mais de 13 anos.

O propósito da rede é permitir a conexão e o compartilhamento de conteúdos com todos aqueles considerados importantes. Assim, a rede foi adiquirindo cada vez mais adeptos e despertando também o interesse de empresas que viram a ferramenta como uma poderosa estratégia de marketing. Em 27 de junho de 2017, o Facebook atingiu a marca de 2 bilhões de usuários ativos se consolidando como a maior rede social do mundo.

3.1.3.3 Instagram

Kevin Systrom e Mike Krieger criaram o Instagram em 2010. Esta rede social é direcionada a imagens, permitindo o compartilhamento de fotos e vídeos além de comentários e curtidas de seguidores nas publicações. Com cerca de um ano, o número de usuários com dispositivos Apple, chegava a 10 milhões. Em 2012, o Instagram foi adquirido pelo Facebook por 1 bilhão de dólares e disponibilizado também para os dispositivos Android. A companhia têm como valores a simplicidade, criatividade e privacidade.

3.1.3.4 Google+

A rede social do Google, o Google+ surgiu em 2011 para competir com o Facebook. A rede permite o compartilhamento de mensagens, links, imagens entre outros através de recursos como: os círculos, que permitia a organização dos amigos da rede em grupos, Hangouts, fotos, eventos, comunidades etc. A plataforma foi desenvolvida para web e smartphones, com aplicações para Android e IOS.

Apesar dos seus mais de 100 milhões de usuários, o engajamento da plataforma era fraco assim, no início do ano de 2019 o Google anunciou que o Google+ seria desativado e que a partir de abril deste mesmo ano, todos os perfis seriam desativados e seus conteúdos excluídos.

3.1.3.5 Twitter

A rede social Twitter foi fundada em março de 2006 por Jack Dorsey, Evan Williams e Biz Stone. Suas atualizações curtas em formato de imagem ou texto a caracterizam como um serviço de *microblogging*. A princípio, o Twitter foi desenvolvido para ser um serviço de troca de status, ser o SMS da internet. Por isso, a característica principal desta rede é a publicação instântanea de tamanho limitado. Por muito tempo o limite ficou estabelcido em 140 caracteres, mas acabou sendo extendido até 280. Esse status, conhecido como Tweet que significa gorjeio ou pio de passarinhos, gera movimentação na rede quando os usuários respondem ao questionamento *What's happening?* (O que está acontecendo?). Neste espaço, os usuários se expressam livremente sobre quaisquer temas. Como valor principal, o Twitter acredita na livre expressão e no poder que as vozes têm de mudar o mundo (TWITTER, 2018b).

3.1.3.6 Facebook x Twitter

Os tópicos anteriores exploraram as cinco redes sociais mais populares entre os brasileiros. Observou-se que as redes Youtube e Instagram são direcionadas à vídeos e imagens respectivamente e o Google+, apesar de se assemelhar ao Facebook quanto aos tipos de mídia que podem ser compartilhadas, perde preferência por ter menor engajamento entre usuários. Dessa forma, as redes Facebook e Twitter receberão maior destaque ao longo deste estudo por se tratar de redes que possuem características relacionadas à mensagens textuais e público mais coerentes com os objetivos deste estudo. As duas redes sociais Facebook e Twitter possuem uma série de semelhanças e diferenças, e nesta secção serão exploradas algumas delas.

O Facebook estabelece a conexão entre os usuários de forma bidirecional, assim um usuário solicita a amizade e outro deve confirmar. As amizades geralmente já existem no meio offline e posteriormente se tornam virtuais. No Twitter, ocorre a unitelaridade da conexão, em que só existe a opção seguir. Dessa forma, os usuários buscam seguir e receber atualizações somente daqueles que produzem conteúdos considerados para eles relevantes. Por esta razão Webtegrity (2014) diz que o Facebook conecta pessoas, enquanto o Twitter conecta tópicos e idéias.

As duas redes possuem em comum recursos como *hashtags*, menções, tópicos e é possível pesquisar pessoas, negócios, tópicos e empresas (WEBTEGRITY, 2014). Ainda segundo o autor o Facebook ao disponibilizar mais recursos que o Twitter, se torna mais difícil de ser usado pelos usuários.

O quadro 1, compara outras características das redes porém há duas que interferem diretamente nesta pesquisa: a configuração da privacidade e o tamanho da publicação. No que tange à publicação, o Facebook permite que o usuário se expresse livremente sem impor limites, enquanto o Twitter o dispõe apenas 280 caracteres. Quanto à privacidade o Facebook permite selecionar a privacidade a cada publicação, assim o usuário tem a liberdade de escolher quais conteúdos serão exibidos para o público, amigos, amigos de amigos, bem como criar uma visualização especial para pessoas selecionadas. Já no Twitter os *tweets* são públicos por padrão. Para proteger suas publicações, o usuário deve proteger o seu perfil por completo, devendo autorizar cada nova solicitação de seguir.

Quadro 1 – Semelhanças e diferenças das redes sociais Facebook e Twitter

	Facebook	**Twitter**
Registro	Requerido	Requerido
Tipo do Site	Serviço Rede Social	Serviço Rede Social
Website	www.facebook.com	https://www.twitter.com
Mensagens Privadas	Sim	Sim
Características	Amigos, Fãs, Mural, Feed de Notícias, *Fanpage*, Grupos, *Apps*,*Likes*, Fotos, vídeos, textos, pesquisas, links, jogos, *status*, cutucadas, presentes mensagem, seção classificada.	*Tweet*, *Retweet*, Mensagem Direta Seguir, Assuntos do momento, *Links*, fotos, vídeos
Número Usuários	2 bilhões	Acima 500 milhões
Linguagens	Disponível em 140 línguas	Disponível em 29 línguas
Aprovação pelos usuários	Curtir, compartilhar	*retweet*, favoritar
Seguir usuários	Somente se o usuário habilitar a opção seguir	Sim
Adicionar amigos	Sim	Não
Tamanho da postagem	Ilimitado	280 caracteres
Jogos	Sim	Não
Editar publicação	Sim	Não
Configuração privacidade	Sim	Pública ou privada
Expressar opinião sobre o assunto	Comentar	Responder
Menção aos usuários	@(user)	@user

Fonte: Adaptado de (DIFFEN, 2016)

3.2 Extração de dados das Redes Sociais

As redes sociais em geral, disponibilizam interfaces para a interação ou captura de seus dados. Para que os desenvolvedores consigam usufruir deste recurso, as redes dispõe do site *developers* que contêm documentação e guias para auxiliar no processo de comunicação com as novas aplicações. Este site, também expõe as regras de utilização e de acesso aos dados, bem como o processo de segurança de autenticação exigido. Essas interfaces são conhecidas como *Application Programming Inteface* (API) e para realizar a extração de dados é preciso conhecer as especificidades da API com a qual se irá trabalhar. Como por exemplo, descobrir qual a linguagem ela consome e retorna, bem como a sua estrutura e regras. A seguir, esses itens ganharão mais profundidade tanto no panorama geral, como no específico.

3.2.1 Application Programming Interface

Uma API é uma interface que concede acesso aos desenvolvedores tanto aos dados como aos serviços de uma ferramenta para a criação de aplicações (JACOBSON; BRAIL; WOODS, 2011)(BUCCAFURRI et al., 2015). Jacobson, Brail e Woods (2011), ainda ressalta que ela funciona como um contrato, devido à sua interface devidamente documentada, consistente e previsível. Tais características transmitem confiabilidade e eficiência para ambos os lados, tanto para o provedor como para o consumidor da API. As APIs são para Makice (2009) uma forma de tornar os dados antes pertencentes à uma ferramenta acessíveis ao resto do mundo.

As APIs possuem três atores: os provedores, desenvolvedores e usuários finais. Os provedores são as companhias que as fornecem, como Google, Youtube, Facebook, Twitter etc. Os desenvolvedores são considerados audiência primária das regras, pois eles irão utilizá-

las para desenvolver aplicações para a audiência secundária que são os usuários finais. Estes farão uso direto das aplicações desenvolvidas. A presença dos usuários finais, interfere na forma com que o provedor deve se preocupar com os dados disponibilizados por sua API, seja na forma de endereço, *copyright*, usos legais, entre outros.

O desenvolvimento de uma API requer, segundo Jacobson, Brail e Woods (2011), definições quanto a sua estrutura, formato de dados utilizado para consumo e retorno, bem como detalhes de segurança. Essas definições devem atender aos desenvolvedores que são os consumidores diretos da API. O autor recomenda para o desenvolvimento de APIs: o formato de dados JSON (*JavaScript Object Notation*), e o *framework* de autenticação OAuth. O JSON segundo ele é mais fácil de se desenvolver e consumir do que o formato XML (Extensible Markup Language). Já o OAuth, serve para proteger os dados sigilosos dos usuários na internet e pode ser utilizado de maneiras variadas.

Quanto aos tipos de APIs, existem dois: as públicas e as privadas. As primeiras são aquelas disponibilizadas por um provedor diante do aceite dos termos de uso, de forma que a relação existente referente ao contrato entre provedor e desenvolvedor seja simplória. Nas APIs privadas, o contrato é complexo com acordos formais estabelecidos segundo uma base legal. Jacobson, Brail e Woods (2011) ainda diz que as APIs públicas representam hoje a ponta do Iceberg e que em contrapartida, as privadas, que são a grande maioria, são imperceptíveis. Acredita que são as privadas que ajudam a transformar as companhias e que elas trazem benefícios reais para os negócios.

Este estudo coletará dados das redes sociais através de uma API pública. Segundo França et al. (2014), existem duas formas de fazê-lo: através de extração retroativa ou em tempo real, baseada em *streaming*. Na extração retroativa, os alvos ou termos de busca são procurados na base de dados do provedor e retornados pela API. Anteriormente, as APIs permitiam que se buscasse conteúdo publicados em qualquer período ou data, porém com o grande aumento da quantidade de dados, foi instaurado em algumas delas como o Twitter, uma restrição quanto ao período de busca, sendo permitido coletar dados até uma data limite. A extração baseada em *streaming* mantêm uma conexão aberta com o servidor e a medida que os dados são publicados com o alvo ou termo de busca, são devolvidos pela API permitindo uma coleta em tempo real.

3.2.2 Framework

Um *framework* é definido por Schmidt e Buschmann (2003), como um conjunto de padrões criados apartir de uma abstração, tanto das estruturas como do mecanismo das aplicações. O *framework* ainda fornece controle específico de comportamentos e estruturas da aplicação. Esse controle é dado aos desenvolvedores e permite assim, o reuso de código.

O *framework* mais importante deste estudo e necessário para a extração de dados é o OAuth que será explorado com mais detalhes na sua secção específica a seguir.

3.2.2.1 OAuth

A *Open Authorization* mais conhecida pela sigla OAuth, é um *framework* desenvolvido para permitir que as aplicações de terceiros, sejam elas das plataformas Web, mobile ou Desktop, possam interagir de forma segura com servidores HTTP, por meio de um método padronizado simples (OAUTH, 2017).

Antes de ser criado, as aplicações pediam as credenciais da conta do usuário como uma forma de autorização para executar determinados serviços. Porém, identificou-se que essa forma de autenticação dava à terceiros acesso irrestrito e permanente às contas das pessoas e também à todos os seus dados, com permissão para executar as ações: adicionar dados, excluir e fazer publicações (LEIBA, 2012). Com o OAuth, os dados sensíveis das pessoas como por exemplo, nome de usuário e senha, são preservados, limitando o acesso à conta pelas aplicações. Os usuários passam a deter controle sobre as aplicações, gerenciando o nível de acesso fornecido à elas e permitindo a sua revogação à qualquer momento (RUSSELL, 2013).

Desde a criação deste *framework*, algumas ferramentas sociais o estão utilizando para preservar os seus usuários, em especial o Facebook e o Twitter. O OAuth evoluiu, e a sua primeira versão OAuth 1.0 (Documento técnico do IETF (Internet Enginnering Task Force) de número RFC 5849) foi substituída pela versão OAuth 2.0 (RFC 6749).

Para a compreensão do OAuth, segundo (LEIBA, 2012) algumas terminologias são importantes:

- Cliente: O Cliente da transação OAuth é o serviço que solicita a autorização.
- Proprietário de Recursos: Aquele que possui os dados requeridos pelo cliente.
- Servidor de Recurso: Serviço que concederá acesso às informações dos proprietários.
- Servidor de Autorização: Serviço verificador de credenciais que validará a autorização.
- Chave de Acesso (Token): Chave concedida pelo servidor de autorização ao cliente para que ele consiga fazer requisições ao servidor de recurso. Essa chave é utilizada pela aplicação em substituição das credenciais do Proprietário de Recursos.
- Garantia de Autorização: Informação que serve para validar uma transação pelo servidor de autorização.

Essas terminologias são necessárias para a compreensão das etapas de autenticação. A Figura 4, mostra o fluxo e a troca de informações entre os atores nas 6 etapas do OAuth 2.0. O cliente é quem dá início ao processo, requisitando autorização de acesso ao proprietário do recurso. O proprietário de recursos concede a autorização retornando uma credencial que é uma garantia de autorização. O cliente com posse da garantia de autorização requer uma chave de acesso ao servidor de autorização. O servidor de autorização por sua vez, autentica o cliente e valida a garantia de autorização, e somente após o sucesso de todas as validações o servidor concede a chave de acesso. Esta chave é utilizada pelo cliente para requerer ao

Figura 4 – Fluxo de autenticação OAuth

Fonte: Adaptado de Hardt (2012)

servidor de recurso a informação desejada. O servidor faz a validação desta chave e retorna a informação, caso a chave seja válida.

3.2.3 Formato de arquivo - JSON

O *JavaScript Object Notation* (JSON) segundo JSON (2018), é um formato para troca de dados, desenvolvido com o intuito de ser familiar aos humanos proporcionando tanto leitura quanto escrita fácil. Este formato também é vantajoso para as máquinas, porque facilita a análise sintática e seu uso. A Figura 5, apresenta um objeto com as características: nome, sobrenome, idade e sexo. O objeto têm seu início e fim determinados por chave esquerda e chave direita respectivamente. Os valores do objeto são pares formados pelo conjunto ("nome :valor"). E quando o objeto possui mais de um valor, eles são separados entre si por vírgulas como mostra o exemplo.

Figura 5 – Exemplo de objeto do formato JSON

```
{
        nome        :    Maria,
        sobrenome   :    Silva,
        idade       :    25,
        sexo        :    feminino
}
```

Fonte: Autora

3.2.4 APIs específicas

No Quadro abaixo 2, estão apresentadas as cinco maiores redes sociais do Brasil de acordo com Social e Hootsuit (2018), suas APIs de acesso aos dados e recursos, bem como as formas de autenticação por elas exigidas. A documentação das APIs e as regras podem ser encontradas no site *developers* de cada rede. Com o intuito de exemplificar as características específicas de uma API, as secções abaixo exploram as principais APIs do Facebook e Twitter.

Quadro 2 – Informações das APIs de algumas redes sociais utilizadas no Brasil

Rede Social	**Site**	**API - Acesso a dados**	**Autenticação**
Youtube	https://developers.google.com/youtube/	Youtube Data API YouTube Analytics API YouTube Live Streaming API	OAuth 2.0
Facebook	https://developers.facebook.com/	Graph API API de Feed Público	OAuth 2.0
Instagram	https://developers.facebook.com/	Instagram Graph API	OAuth 2.0
Google+	https://developers.google.com/+/	Google+ API	OAuth 2.0
Twitter	https://developer.twitter.com/	REST API, Streaming API	OAuth 2.0, HTTP Basic Authentication

Fonte: Autora

3.2.4.1 API do Facebook

A principal API da rede social Facebook é a Graph API. Segundo Facebook (2018), todas as outras APIs existentes são expansões desta, e todos os produtos do Facebook interagem com esta API. Uma outra API que permite acesso à dados é a API de Feed Público, porém, para que seja usada o desenvolvedor necessita de aprovação prévia do Facebook, já que a inscrição dela ainda não se encontra aberta.

É através da Graph API que ocorre a inserção e a extração de dados do Facebook. Esta é uma API que carrega diversas funcionalidades e possibilidades, com ela é possível fazer publicações, consultas, gerenciar anúncios entre outras tarefas. Mas para realizar solicitações à Graph API são necessárias tokens de acesso que fornecem autorizacão para tal. A necessidade dessas tokens se deve ao uso do OAuth 2.0. Após realizadas essas solicitações o retorno para as buscas feitas na Graph API são do formato JSON.

A Graph API estabelece uma limitação de volume relacionada à quantidade de requisições que podem ser feitas ao servidor da ferramenta dentro de um intervalo de tempo. Durante uma hora, uma aplicação pode realizar uma quantidade de chamadas igual a duzentas vezes o número de usuários que ela possui. Ou uma página, pode fazer no máximo 4.800 chamadas por usuário em um período de 24 horas.

3.2.4.2 API do Twitter

O Twitter fornece aos desenvolvedores uma grande variedade de APIs para a construção de aplicações ou soluções para a sua ferramenta. Ao final do ano de 2018, existiam cinco categorias de APIs: Standard, Premium, Enterprise, Ads (para anúncios) e Twitter para websites. Porém, quando se trata de extração de dados, deve-se considerar as três primeiras

categorias. A API Standard, que será utilizada neste trabalho, é uma categoria de API gratuita, já as duas outras categorias, Premium e Enterprise, são pagas e possuem funcionalidades que permitem análises e monitoramento profundo voltado para os negócios (TWITTER, 2018a).

A categoria Standard conta com dois módulos chamados API REST e API Streaming. A API REST permite a realização de consultas no banco de dados disponibilizado, a execução de ações de inserção, atualização e exclusão (BUCCAFURRI et al., 2015). As consultas porém, tem um endpoint de apenas sete dias, o que significa que ao usar essa categoria, só se consegue obter publicações feitas há no máximo sete dias. Já as categorias pagas, conseguem ter um limite retroativo maior que varia de 30 dias à todo o histórico do Twitter, desde 2006.

As pesquisas feitas utilizando a API REST da categoria Standard são pautadas na relevância, dessa forma, os dados capturados são aqueles considerados os mais importantes pela ferramenta, de forma que possa haver publicações que não serão exibidas. Além disso, neste módulo é necessária autenticação para ser utilizado, as respostas são retornadas para o desenvolvedor em JSON e existe um limite quando as chamadas direcionadas ao servidor, o *15 Minute Windows*. O twitter seccionou o tempo em intervalos de quinze minutos e definiu que dentro desse período seria possível operar em dois modos, realizando 180 requisições ou 450 requisições dentro desta janela de tempo. Quando a aplicação atinge a quantidade limite de requisições, o seu acesso ao servidor é bloqueado impedindo novos acessos. Caso haja insistências em ultrapassar o limite de requisições permitidas pelo servidor o IP requisitante pode ser banido por tempo indeterminado.

No módulo de API Streaming os desenvolvedores tem acesso às publicações em tempo real. Por este motivo, aplicações requerentes de informações reais e completas devem fazer o uso deste modo de acesso. Ela consegue estes feitos por fornecer as informações a uma latência muito baixa e possuir conexão de longa duração. Portanto, enquanto o link estiver ativo o retorno da requisição será incrementado.

3.3 Mineração de Texto

A Mineração de textos (MT) é uma extensão da área de mineração de dados (TAN et al., 1999; HEARST, 1999). Esta subárea foi desenvolvida para que a partir de uma grande quantidade de informações textuais, contidas em textos não estruturados ou semi-estruturados, fosse possível encontrar padrões relevantes e não triviais e ainda extrair novas informações (HEARST, 1999; KANTARDZIC, 2011). Para Hearst (2003) a mineração de textos é a descoberta de informações através da computação. Nesse processo, conteúdos textuais de diferentes fontes são extraídos automaticamente para satisfazer o objetivo de revelar uma informação completamente desconhecida.

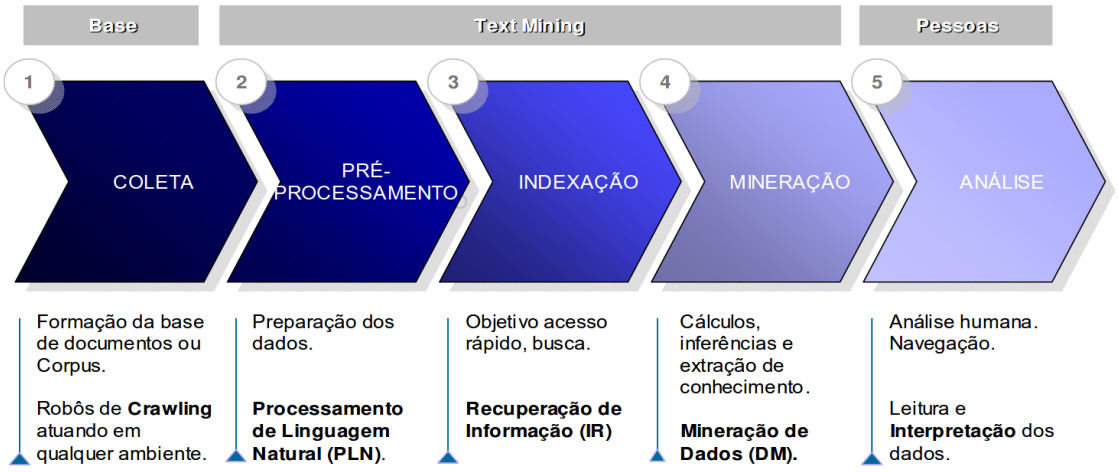
Para encontrar padrões, as técnicas de mineração de dados e textos necessitam de dados altamente estruturados. Portanto, os dados adquiridos precisam ter um estrutura compatível com as técnicas. Caso não estejam no formato requisitado, eles precisarão passar por uma extensa transformação para que se adequem e adiquiram a estrutura necessária (WEISS

et al., 2010). Ainda segundo o autor, tanto a mineração de dados como a mineração de textos são conceitos que se baseam no passado, ou seja, através de amostras do passado elas conseguem abstrair algo completamente novo. Os dois conceitos também fazem uso das mesmas técnicas de aprendizagem, a diferença entre os dois está no objeto de análise: o primeiro analiza números e o segundo textos.

A mineração de dados e a mineração de textos, se diferenciam na forma de reconhecer os padrões de seus objetos de análise. Algoritmos computacionais são incapazes de ler ou compreender textos escritos, dessa forma, os padrões encontrados pela mineração de textos estão relacionados à linguagem natural, ou seja à linguística (HEARST, 2003). Para minerar uma forma textual, segundo Weiss et al. (2010), se faz necessária a transformação dos textos em uma representação numérica . Já na mineração de dados, as informações estão estruturadas dentro do banco de dados e são formadas por fatos, já que foram armazenadas automaticamente por sistemas (HEARST, 2003).

Este trabalho utilizará a mineração de textos a fim de extrair os sentimentos dos documentos online produzidos por usuários das redes sociais. Para isto, foi escolhido o modelo de mineração textual proposto por (ARANHA; VELLASCO; PASSOS, 2007). Este modelo, engloba todo o processo de mineração e por este motivo é composto por uma série de etapas que serão exploradas juntamente com as suas técnicas. A Figura 6, expõe todas as etapas do processo de mineração deste modelo de forma sucinta.

Figura 6 – Etapas de mineração a serem executadas



Fonte: Aranha, Vellasco e Passos (2007)

O modelo é formado por cinco etapas sequenciais. A primeira delas é a formação da base. O objetivo desta etapa é a coleta de informações para a construção de um universo de dados a serem explorados. A segunda é o pré-processamento. Nesta etapa os dados são transformados para se adequarem a formatos passíveis de leitura para os algoritmos de extração automática de conhecimento. Posteriormente têm-se a indexação que cria índices para facilitar a recuperação da informação. A mineração é a quarta etapa e nela se aplica algoritmos para a extrair o conhecimento, em seguida, realiza-se a quinta fase: a análise dos resultados

obtidos.

3.3.1 Coleta

Em mineração de textos a coleta dos dados é uma fase que têm como principal propósito a construção de um arcabouço de dados. O próprio conceito da mineração de textos diz que este processo é feito a partir de um volume de dados robusto. Então, para que estes documentos textuais consigam ser reunidos é necessário primeiramente descobrir onde eles podem ser encontrados.

Segundo Aranha, Vellasco e Passos (2007) os dados podem estar presentes em três ambientes: No diretório de pastas dentro de um dísco rígido; Em tabelas de Bancos de Dados; E na Internet. Após identificar o alvo da extração é preciso que se defina métodos para acessá-los nesses ambientes e posteriormente conseguir direcionar a coleta para textos relevantes a fim de produzir novas informações dentro do contexto desejado.

Como o presente trabalho é voltado para as redes sociais, a ênfase da coleta estará voltada para dados contidos na Internet. Dentro do ambiente virtual os textos estão espalhados em sites, blogs, fóruns, páginas de notícias, em documentos oficiais e acadêmicos como artigos, livros e revistas.

Uma forma de obter os dados de páginas web é através de *Web Crawlers*. Também conhecidos como *Robot* ou *Spider*, são programas desenvolvidos para navegar pelas páginas Web através dos *hyperlinks*. De acordo com Pant, Srinivasan e Menczer (2004), esses programas recebem uma lista de sites sementes e a partir deles outros sites são consumidos em cadeia através dessas conexões existentes entre eles. Ainda segundo o autor, os crawlers são muito velozes capazes de visitar milhares de páginas web em poucos minutos.

Os Crawlers são grandes consumidores de recursos em função da sua atividade de acumular informações. Nas palavras de Menczer, Pant e Srinivasan (2004), esses algoritmos necessitam de largura de banda para baixar as páginas visitadas; memória do sistema para sustentar a sua execução; Unidade de Processamento (CPU) para avaliar as URLs de acordo com os critérios do algoritmo e Armazenamento em Disco para guardar todos os textos encontrados bem como os *hyperlinks* para incrementar a lista para a sua próxima visita.

A coleta de dados direcionada à uma rede social deve ser feita através de uma API. A maioria das redes sociais (Youtube, Facebook, Instagram, Twitter, Pinterest, Tumblr e Linkedin), disponibiliza essa aplicação para que os desenvolvedores consigam acesso aos dados da rede.

3.3.2 Pré-Processamento

O sucesso dos algoritmos de mineração de textos está atrelado à qualidade dos dados formadores do *Corpus* - conjunto de documentos (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006). Por isso, após a etapa de coleta faz-se necessário o pré-processamento dos dados. Essa etapa é responsável por transformar os dados adquiridos, como os textos por exemplo, a ganharem estrutura melhorando assim a qualidade e a organização de todo aquele

amontoado de dados facilitando a submissão destes aos algoritmos de indexação e mineração de dados (ARANHA; VELLASCO; PASSOS, 2007).

Segundo Kotsiantis, Kanellopoulos e Pintelas (2006), quando os dados se apresentam de forma ruidosa, estranha ou quando trazem informações irrelevantes a base prejudica a acurácia dos algoritmos de mineração de dados fazendo com que eles não encontrem informações úteis. O pré-processamento além de aprimorar o resultado, reduz o custo da etapa de mineração agindo na solução de problemas como: Dados ruidosos, faltantes e ou redundantes. Deve-se levar em consideração que as bases de dados trabalhadas são geralmente muito extensas, então uma simples redundância sem sentido pode elevar o tempo de execução do algoritmo sem agregar nenhuma informação nova, favorecendo o consumo de recursos e ainda tornando o algoritmo menos assertivo.

Porém, um desafio dessa etapa é manter a integridade dos dados, de forma que após passar por todas as técnicas os dados estejam tranformados e consigam ainda representar com fidelidade o conjunto de dados original sem que tenha havido perda de informações. Duas técnicas de pré-processamento utilizadas neste estudo serão exploradas a seguir: Tokenização e remoção de stopwords.

3.3.2.1 Tokenização

A tokenização é um processo que acontece no início da fase de pré-processamento. Nele, os documentos adquiridos através da extração são divididos em pequenas unidades chamadas *Tokens*. Os *Tokens* normalmente representam palavras, números ou símbolos de pontuação (MANNING; SCHÜTZE, 1999).

Existem diferentes abordagens dentro da tokenização, até porque este processo esta ligado à uma linguagem natural. Como as línguas se diferem em diversos aspectos as formas de manipulá-las devem ser também específicas. De acordo com Indurkhya e Damerau (2010) existe uma diferença crucial na tokenização de línguas que são delimitadas por espaço e das que são não segmentadas.

As línguas não segmentadas são formadas por palavras sucessivas que não possuem um símbolo ou caractere de limitação. São exemplos dessas línguas, o Chinês, Japonês e Thailandês. Por não serem limitadas essas linguas requerem outras informações para serem tokenizadas. Informações do tipo léxico ou morfológico. Já as línguas delimitadas por espaço, como o próprio nome já diz, correspondem aquelas que possuem o espaço como limitador de palavras. Este é o caso do português brasileiro, inglês e a maioria das línguas européias (INDURKHYA; DAMERAU, 2010).

Este trabalho irá construir o seu *Corpus* a partir de documentos na língua Portugês do Brasil, por isso serão utilizadas somente abordagens delimitadas por espaçamento. Abaixo, um exemplo de tokenização da frase: "A vida começa onde termina o medo !".

[A] [vida] [começa] [onde] [termina] [o] [medo] [!]

Como pode ser visto, o processo utiliza o espaço como uma forma de subdividir o texto em partes menores. Assim, uma simples frase se transforma em oito tokens.

3.3.2.2 Remoção de stopwords

Um texto é formado por uma sequência de termos que garantem que o texto seja flúido e consiga transmitir uma mensagem. Porém, alguns destes termos possuem um papel mais sintático do que semântico e não carregam consigo um conteúdo útil (WILBUR; SIROTKIN, 1992). Este é o caso de algumas classes da nossa língua, como os pronomes, artigos ,preposições e conjunções.

Os termos pertencentes às classes citadas acima são chamados *stopwords*. São exemplos de *stopwords* as palavras "de", "a", "um", "o", "até". Estes termos se encontram espalhados por todo o texto e aparecem com uma grande frequência porém como não agregam informações para a análise que será realizada, são removidos do *Corpus* textual.

3.3.3 Indexação

A indexação é a terceira etapa da metodologia de Aranha, Vellasco e Passos (2007), ela pertence à uma subàrea da mineração de dados conhecida como recuperação de informação. Ainda de acordo com o autor do método, a etapa permite que o *Corpus* seja manipulado de maneira mais eficiente.

Diante de um volume de dados muito grande, uma simples consulta acaba se tornando uma tarefa árdua e difícil porque todo o conjunto de documentos precisa ser percorrido de forma a ser possível encontrar aquela ocorrência em meio a milhares ou às vezes milhões de caracteres. É uma tarefa que demanda tempo e muito esforço computacional (SOARES, 2013).

A indexação então surge com o objetivo de se criar índices. Estes são coleções de termos únicos extraídos do conjunto de dados, juntamente com ponteiros referentes aos lugares onde aqueles termos podem ser encontrados no documento (PAL; TALWAR; MITRA, 2002). Assim, essas estruturas permitem a consulta de termos sem a necessidade de percorrer toda a base de dados (SCHÜTZE; MANNING; RAGHAVAN, 2008). Proporcionando otimização da velocidade e desempenho na busca realizada (SOARES, 2013).

3.3.4 Mineração

A fase de mineração irá trabalhar em cima dos dados provenientes dos processos anteriores para a produção efetiva de novos conhecimentos úteis. Para isto, podem ser aplicados sobre os dados, algoritmos pertencentes à diversas áreas do conhecimento: Estatística, Aprendizado de Máquina, Redes Neurais e Banco de Dados (ARANHA; VELLASCO; PASSOS, 2007). Portanto, faz-se necessário escolher dentre as opções de abordagens disponíveis, aquela que mais se adequa ao objetivo que se quer alcançar.

Nas palavras de Junior (2007), para se verificar a similaridade entre documentos a técnica utilizada deve ser a de clusterização, já que esta técnica reúne os dados formando grupos, ou *clusters*. Outra técnica existente é a de classificação que categoriza documentos para que eles pertençam a um conjunto. O autor ainda diz que ambas as técnicas pertencem

tanto à mineração de dados quanto à mineração de textos, porém a mineração de textos possui ainda outras técnicas exclusivas como a sumarização e a extração de características.

3.3.5 Análise

Por fim, a última etapa têm início. A etapa de análise e interpretação dos resultados. Esta fase do processo irá a partir dos dados obtidos fazer um verificação para avaliar o classificador de acordo com algumas métricas que podem ser taxa de erro, tempo de CPU e complexidade do modelo (ARANHA; VELLASCO; PASSOS, 2007).

Ou ainda as métricas podem ser qualitativas, levando em consideração o conhecimento de especialistas do domínio (JUNIOR, 2007). Esses especialistas irão fazer verificações dos resultados conquistados com o conhecimento disponível do domínio e deixando a cargo do usuário a decisão sobre a aplicação dos resultados (ARANHA; VELLASCO; PASSOS, 2007).

3.4 Análise de Sentimentos

A análise de sentimentos nasce da necessidade de se extrair opiniões de textos opinativos de forma automática devido ao volume de textos produzidos na Web (LIU, 2010). Ainda segundo o autor, este é um problema que pertence ao processamento de linguagem natural ou mineração de texto.

Existem nas palavras de Liu (2010), duas categorias às quais o texto pode pertencer: fato ou opinião. Um fato é uma verdade indiscutível, já uma opinião carrega consigo uma subjetividade. E é em cima dessa subjetividade que a análise de sentimentos trabalha. "A análise de sentimentos ou mineração de opinião é o estudo computacional de opiniões, sentimentos e emoções expressas no texto"[Tradução Nossa] (LIU, 2010, p.629).

A análise de sentimentos se divide entre as abordagens de aprendizado de máquina, análise léxica ou híbrida (MEDHAT; HASSAN; KORASHY, 2014). Na híbrida, a análise léxica e o aprendizado de máquina são usados em conjunto para se chegar ao resultado. O autor ainda diz que a análise de sentimentos se dividem quanto aos objetivos, elas podem buscar a classificação de sentimentos ou classificação de emoções. As secções abaixo se aprofundam nestes itens apresentados.

3.4.1 Opinião, Sentimento e Emoção

Seguindo a linha de pensamento de Liu (2010), têm-se que a opinião, o sentimento e a emoção são conceitos diferentes mas que se relacionam dentro da área de estudo da análise de sentimentos. Segundo ele a opinião é uma visão, atitude ou emoção sobre um objeto. Essa opinião se traduz em um sentimento ou orientação, também conhecida como orientação de sentimento, orientação semântica ou polaridade da opinião, que pode ser positiva, negativa ou neutra. Dessa forma, a opinião expressa de forma textual pode ter o seu sentimento encontrado e classificado.

As emoções por sua vez, são de acordo com Liu (2010), sentimentos subjetivos específicos. Alguns exemplos de emoções são: amor, alegria, surpresa, medo, raiva, tristeza.

As emoções, que podem ainda ser subdivididas em diversas categorias e intensidades, são os estados mentais das pessoas e a análise de sentimentos se interessa pela linguagem utilizada pelas pessoas para traduzir esse estado em palavras. Assim, expressões textuais de amor ou alegria trasmitem sentimentos positivos, enquanto as de raiva e medo, transmitem sentimentos negativos. A análise de sentimento tenta então, inferir o sentimento das pessoas através da linguagem (LIU, 2010).

3.4.2 Análise de Sentimentos baseada em aprendizado de máquina

A abordagem baseada em aprendizado de máquina é uma técnica que através dos algoritmos de aprendizado e de uma base já classificada passa por etapas de treinamento e consegue aprender a classificar sentenças de acordo com aquelas que lhe foram apresentadas. Assim, ele se torna capaz de classificar sentimentos de sentenças nunca antes vistas. De acordo com Benevenuto, Ribeiro e Araújo (2015) a aprendizagem de máquina têm quatro etapas principais:

1. Dados rotulados para treino e teste (sentenças já classificadas)
2. Definição das características que distinguem os dados (Bag of Words - Array que mapeia cada palavra da sentença para um dicionário de palavras)
3. Treinamento de um modelo computacional com um algoritmo de aprendizado
4. Aplicação do modelo

3.4.3 Análise de Sentimentos baseada em análise léxica

A análise de sentimentos baseadas em abordagens léxicas, segundo Benevenuto, Ribeiro e Araújo (2015), são técnicas não supervisionadas que utilizam dicionários léxicos de sentimento. Neste dicionário, as palavras estão relacionadas à valores quantitativos ou qualitativos que se referem à sentimentos. Por exemplo, uma palavra pode ter um valor entre +1 à -1, onde cada extremo se refere ao sentimento mais positivo e mais negativo respectivamente, ou ter um valor qualitativo como , por exemplo: positivo, negativo, feliz ou triste.

A abordagem léxica é vantajosa porque não necessitar de classificação prévia de sentenças e nem de base de treino como ocorre nos métodos supervisionados, assim, ela fica livre de contexto por justamente não sofrer especialização durante o treinamento (BENEVENUTO; RIBEIRO; ARAÚJO, 2015). Porém, ela precisa de um conjunto de palavras denominadas sementes que já são previamente conhecidas e ter os seus sentimentos já estabelecidos (MEDHAT; HASSAN; KORASHY, 2014).

Esta abordagem é uma das mais eficientes tanto na predição como na utilização dos recursos computacionais, porém para sê-lo são necessárias aplicações de heurísticas para a classificação das sentenças, já que o somatório simples dos valores das palavras presentes na sentença podem ser insatisfatórios (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

4 Materiais e Métodos

Neste capítulo serão apresentados o método e as configurações definidas durante a pesquisa para a resolução do problema em questão.

4.1 Introdução

O presente estudo é uma pesquisa caráter exploratório e descritivo. A pesquisa exploratória nas palavras de Piovesan e Temporini (1995), não existe por si só, mas se mantêm como parte integrante de outra principal. Seu objetivo é conhecer o objeto de estudo de forma a compreender a sua contextualização e o seu significado. Este tipo de pesquisa ainda pode ser considerada uma "investigação em área onde há pouco conhecimento sistematizado, acumulado."(VERGARA, 1990, p.4). Já a pesquisa descritiva, é definida por Vergara (1990) como sendo uma exposição de características de um fenômeno em que fatores e variáveis podem ser corelacionados.

Dentre as fontes que compõe este estudo, constam artigos recentes sobre o tema de extração de dados das mídias sociais, mineração de texto e análise de sentimentos, buscando sempre autores de referência e de autoridade no assunto, como: Bing Liu, Bo Pang, Lillian Lee, Danah Boyd, Nicole Ellison, Alec Go e o brasileiro Fabrício Benevenuto. Estes autores foram identificados devido à grande quantidade de citações feitas à eles, evidenciando o valor de suas contribuições para a área. E também à quantidade de publicações feitas por eles na área estudada. Os resultados deste estudo por sua vez serão apresentados de forma qualitativa e quantitativa utilizando de recursos visuais como gráficos, tabelas e quadros, buscando uma melhor exibição dos resultados obtidos a partir da análise realizada.

A presente pesquisa realizará uma análise de sentimentos das mensagens da rede social Twitter: os *tweets*. Nesta análise de sentimentos os *tweets* serão polarizados em três sentimentos: positivo, negativo e neutro. Para a sua execução, a pesquisa utilizará como base o método proposto por Aranha, Vellasco e Passos (2007) ilustrado na figura 6. Para isto, devem estar previamente definidos a rede social e o evento a ser analisado. As seções abaixo detalham os pontos considerados para cada escolha.

4.1.1 Escolha da Rede Social

Dentre as diversas redes sociais disponíveis online, as cinco maiores redes do Brasil no ano de 2018 foram respectivamente: Youtube, Facebook, Instagram, Google+ e Twitter segundo Social e Hootsuit (2018). Todas essas redes são gratuitas e de fácil acesso, porém , existe uma rede que reúne características que viabilizam a mineração de texto: O Twitter.

A primeira característica que o destaca é o propósito da rede. Apesar de ser permitido o compartilhamento de links, gifs, vídeos e fotos a rede têm como principal objetivo a propagação de mensagens textuais. Estes outros formatos de mídia foram acrescentados porteriormente

na rede para incrementar a comunicação. Levando em consideração o conteúdo, enquanto o Twitter esta voltado para as mensagens, as redes Youtube e Instagram possuem como foco o compartilhamento de vídeos e fotos respectivamente, e por este motivo foram descartadas já que o objetivo do trabalho é a mineração textual.

Outra característica vantajosa do Twitter é a configuração padrão de visibilidade do perfil. Um novo perfil criado no Twitter possui a visibilidade definida como pública por padrão, o que quer dizer que todas as informações compartilhadas são visíveis à toda a rede. Caso o usuário deseje, há a possibilidade de restringir a visualização de suas publicações alterando estas configurações para privado, assim, somente os usuários que fazem parte de sua rede poderão enxergá-las. As aplicações criadas para acessar os dados de uma plataforma online têm acesso direto às postagens públicas. Porém, em um perfil privado é possível conseguir acesso somente se o usuário autorizar explicitamente este acesso da aplicação às suas publicações. Assim, os perfis privados acabam sendo uma barreira que dificulta o acesso aos dados. Estes perfis são mais comuns por exemplo, no Facebook do que no Twitter, o que favorece a segunda rede no quesito acesso.

A terceira e última característica que favorece o Twitter sobre os demais neste estudo é a mensagem curta. O Twitter é uma rede voltada para a comunicação rápida e por este motivo libera um espaço limitado para que os usuários se expressem. Segundo Liu (2010) quando a mensagem subjetiva é longa demais e possui um nível muito alto de detalhes se torna ainda mais difícil conseguir obter os sentimentos.

4.1.2 Eventos

O ano de 2018 é um ano eleitoral no Brasil e deverão ser escolhidos candidatos para as vagas de deputado federal e estadual, senadores, governador e presidente. Com o tema político em alta, este trabalho estará direcionado aos presidenciáveis e nos sentimentos que os circundam. Para isto, foram escolhidos os seguintes eventos ,como mostra o quadro abaixo,3, para o monitoramento das mensagens publicadas nas redes sociais. Dentre os motivos para esta seleção foram considerados a abrangência em todo o território nacional e o grande volume de mensagens a ser produzido já que o tema político gera muita repercussão.

Quadro 3 – Acontecimentos monitorados durante o período de eleições em 2018

Evento	**Período da Coleta (2018)**
Oficialização do Haddad como candidato	11 de Setembro
Debate no SBT (TV)	26 de Setembro
Debate na Record (TV)	30 de Setembro
Debate na Globo (TV)	04 de Outubro

Fonte: Autora

4.2 Formação da base de dados

Para se ter acesso aos dados da rede social Twitter é preciso se registrar no site dos desenvolvedores (*developers* [1]) e posteriormente fazer o registro da sua aplicação (App). Isso é necessário, porque atualmente o método de autenticação utilizado é o OAuth que requer chaves únicas para autorizar a aplicação a ter acesso às informações. Com as chaves *(keys)* em mãos e os eventos determinados (Quadro 3), o próximo passo foi desenvolver um algoritmo de extração de dados na linguagem Python para a comunicação com o servidor do Twitter através de interações com a API *Streaming*.

O Python foi escolhido por ser uma linguagem que possui bibliotecas que agem como facilitadoras na programação, contendo funções que tornam o processo de comunicação com a API, mais transparente ao nível do programador. Algumas dessas bibliotecas são: Python-twitter e tweepy.

A coleta de dados foi realizada no dia de ocorrência dos eventos mostrados no quadro 3. Para realizar uma busca nas mensagens é necessário determinar um alvo. Assim, o servidor retornará o resultado daquilo que se busca. Os alvos podem ser palavras presentes nos tweets, hashtags, menções e usuários. Esta pesquisa utilizou menções e hashtags como alvo dos algoritmos de busca por concentrarem as mensagens do tema em questão que foram: a oficialização da candidatura de Haddad e os debates nas emissoras SBT, Record e Globo. Os alvos de busca utilizados estão listados abaixo no quadro 4.

Quadro 4 – Alvos dos algoritmos de busca

Evento	**Alvo**
Oficialização do Haddad como candidato	#OBrasilFelizDeNovo, @Haddad_Fernando, #HaddadPresidente
Debate no SBT (TV)	#DebateSBT
Debate na Record (TV)	#DebateNaRecord
Debate na Globo (TV)	#DebateNaGlobo

Fonte: Autora

O retorno da API é um arquivo *JavaScript Object Notation* (JSON) que contém os *tweets* e seus atributos. No algoritmo de extração esse formato de saída foi mantido pela facilidade de manuseio do arquivo e obtenção das informações mais relevantes. Era também possível obter o retorno no formato texto (.txt) porém o JSON se sobresai quando se considera o fator de manipulação.

Ao final do período de coleta, a extração obteve como resultado quatro grandes arquivos com um total de 651.837 *tweets*, como mostra o quadro 5.

O *tweet*, componente essencial do microblog Twitter, é exibido na Figura 7. Nela, percebe-se as informações que a ferramenta disponibiliza visualmente para os seus usuários, que são basicamente: foto de perfil, nomes, data , texto e botões para interação. Quando uma ação é executada (curtir,retweetar e comentar) números indicadores aparecem ao lado para evidenciar o número de interações de cada ação. Foi inserido na figura, uma tarja preta para preservar a identidade do usuário que retweetou a notícia.

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

Quadro 5 – Arquivos dos eventos monitorados durante as eleições de 2018

Data Coleta	Nome Arquivo	Tamanho Inicial	Qtd Tweets
11/09	Haddad.json	464 MB	83.842
26/09	SBT.json	416 MB	48.402
30/09	Record.json	572 MB	60.736
04/10	Globo.json	2,5 GB	458.857

Fonte: Autora

Figura 7 – Informações exibidas de um Tweet

Fonte: (TWITTER, 2018c)

Um *tweet* carrega porém, muito mais informações do que aparenta. Na figura 8, têm-se o mesmo *tweet* representado acima, extraído no primeiro evento e armazenado em arquivo de formato .json. Este pequeno arquivo, quando armazenado em disco, possui geralmente um tamanho aproximado à 4 KB de dados. Além de informações referentes à própria mensagem como texto, menções e hashtags, é possível obter dados de usuário como: nome, quantidade de seguidores e seguidos, status de perfil, quantidade de curtidas bem como outras informações de perfil e configurações de conta.

Para a execução da pesquisa em questão não são necessárias todas estas informações, por isso, somente algumas foram eleitas para compor a base de dados. Assim, o tamanho dos arquivos foram consequentemente reduzidos para uma melhor execução dos algoritmos das próximas etapas. As informações selecionadas foram: o id do *tweet* - número que identifica a mensagem como única. o texto - a mensagem compartilhada na rede social. O nome dos usuários e das hashtags mencionados no tweet como também a data de publica-

Figura 8 – Tweet extraído no formato .json


```
{"quote_count": 0, "contributors": null, "truncated": false, "text": "RT @exame: Turbinado pelo Datafolha, Haddad deve ser oficializado hoje pelo PT
https://t.co/XJ6V0yvNpp", "is_quote_status": false, "in_reply_to_status_id": null, "reply_count": 0, "id": 1039454644513239041, "favorite_count": 0, "entities":
{"user_mentions": [{"id": 30070288, "indices": [3, 9], "id_str": "30070288", "screen_name": "exame", "name": "Exame"}], "symbols": [], "hashtags": [], "urls": [{"url":
"https://t.co/XJ6V0yvNpp", "indices": [79, 102], "expanded_url": "https://abr.ai/2QiA7ea", "display_url": "abr.ai/2QiA7ea"}]}, "retweeted": false, "coordinates": null,
"timestamp_ms": "1536660274042", "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>", "in_reply_to_screen_name": null,
"id_str": "1039454644513239041", "retweet_count": 0, "in_reply_to_user_id": null, "favorited": false, "retweeted_status": {"quote_count": 4, "contributors": null,
"truncated": false, "text": "Turbinado pelo Datafolha, Haddad deve ser oficializado hoje pelo PT https://t.co/XJ6V0yvNpp", "is_quote_status": false, "in_reply_to_status_id":
null, "reply_count": 10, "id": 1039445338300600320, "favorite_count": 46, "source": "<a href=\"http://publicize.wp.com/\" rel=\"nofollow\">WordPress.com</a>", "retweeted":
false, "coordinates": null, "entities": {"user_mentions": [], "symbols": [], "hashtags": [], "urls": [{"url": "https://t.co/XJ6V0yvNpp", "indices": [68, 91], "expanded_url":
"https://abr.ai/2QiA7ea", "display_url": "abr.ai/2QiA7ea"}]}, "in_reply_to_screen_name": null, "id_str": "1039445338300600320", "retweet_count": 12, "in_reply_to_user_id":
null, "favorited": false, "user": {"follow_request_sent": null, "profile_use_background_image": true, "default_profile_image": false, "id": [redacted], "default_profile":
false, "verified": true, "profile_image_url_https": "https://pbs.twimg.com/profile_images/890525951557283840/AkjJ2Ufa_normal.jpg", "profile_sidebar_fill_color": "EFEFEF",
"profile_text_color": "000000", "followers_count": 2301651, "profile_sidebar_border_color": "FFFFFF", "id_str": [redacted], "profile_background_color": "180101",
"listed_count": 5625, "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png", "utc_offset": null, "statuses_count": 148048, "description":
"As not\u00edcias, an\u00e1lises, posts e dicas mais interessantes do site EXAME\r\n\r\nhttp://www.exame.abril.com.br", "friends_count": 256, "location": "S\u00e3o Paulo -
Brasil", "profile_link_color": "CA0300", "profile_image_url": "http://pbs.twimg.com/profile_images/890525951557283840/AkjJ2Ufa_normal.jpg", "following": null, "geo_enabled":
true, "profile_banner_url": "https://pbs.twimg.com/profile_banners/30070288/1347997696", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
"name": "Exame", "lang": "pt", "profile_background_tile": false, "favourites_count": 198, "screen_name": "exame", "notifications": null, "url": "http://www.exame.com.br",
"created_at": "Thu Apr 09 21:20:19 +0000 2009", "contributors_enabled": false, "time_zone": null, "protected": false, "translator_type": "none", "is_translator": false},
"geo": null, "in_reply_to_user_id_str": null, "possibly_sensitive": false, "lang": "pt", "created_at": "Tue Sep 11 09:27:35 +0000 2018", "filter_level": "low",
"in_reply_to_status_id_str": null, "place": null}, "user": {"follow_request_sent": null, "profile_use_background_image": true, "default_profile_image": false, "id":
[redacted], "default_profile": false, "verified": false, "profile_image_url_https": "https://pbs.twimg.com/profile_images/986657297584873472/Dm9UTpsy_normal.jpg",
"profile_sidebar_fill_color": "FFF7CC", "profile_text_color": "0C3E53", "followers_count": 26, "profile_sidebar_border_color": "F2E195", "id_str": [redacted],
"profile_background_color": "BADFCD", "listed_count": 0, "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme12/bg.gif", "utc_offset": null,
"statuses_count": 1980, "description": null, "friends_count": 128, "location": null, "profile_link_color": "FF0000", "profile_image_url":
"http://pbs.twimg.com/profile_images/986657297584873472/Dm9UTpsy_normal.jpg", "following": null, "geo_enabled": false, "profile_banner_url":
"https://pbs.twimg.com/profile_banners/52696338/1513514224", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme12/bg.gif", "name": "Johnee 2018",
"lang": "pt", "profile_background_tile": false, "favourites_count": 3284, "screen_name": [redacted], "notifications": null, "url": null, "created_at": "Wed Jul 01 11:19:49
+0000 2009", "contributors_enabled": false, "time_zone": null, "protected": false, "translator_type": "none", "is_translator": false}, "geo": null,
"in_reply_to_user_id_str": null, "possibly_sensitive": false, "lang": "pt", "created_at": "Tue Sep 11 10:04:34 +0000 2018", "filter_level": "low",
"in_reply_to_status_id_str": null, "place": null}
```


Fonte: Autora

ção da mensagem. O quadro 6 abaixo resume as campos extraídos, que compuseram o novo documento. Este novo documento foi criado por um algoritmo JAVA que a partir do arquivo original de cada evento, selecionou os atributos escolhidos e os preservou na nova base. Dessa forma, os *tweets* tiveram um redução significativa de tamanho (figura 9).

Quadro 6 – Informações dos tweets relevantes para a pesquisa

Informação	**Campo**
Identificador	id
Texto	text
Menção	entities/user_mentions/screen_name
Hashtag	entities/hashtag/text
Data	created_at

Fonte: Autora

Figura 9 – Tweet reduzido


```
{"hashtags":["DebateNaGlobo"],"created_at":["Fri","Oct","05","03:48:58","+0000","2018"],"text":"De Ciro
Gomes @cirogomes no #DebateNaGlobo: \"Essa divisão não vai permitir que o Brasil supere sua crise\"Diz
o candidato que dá soco e tapa em entrevistadores que fazem perguntas difíceis e chamou o Sul do Brasil
de nazista.","id":1048057430591066113,"mencoes":["Desesquerdizada","cirogomes"]}
```


Fonte: Autora

Nesta pesquisa optou-se por utilizar a ferramenta Sentiment140 para a classificação dos *tweets* conseguindo por consequência a polarização dos sentimentos. Porém, como a ferramenta opera na lingua inglesa toda a base de dados necessitava tradução.

Em um primeiro momento se utilizaria na tradução a API Goslate[2] para Python que segundo os próprios criadores é um serviço gratuito, rápido e simples que ao receber uma chamada de serviço abre uma requisição ao Google Tradutor retornando assim o texto traduzido. Porém, o Google desenvolveu uma forma de barrar pequenas API's como esta de terem

[2] https://pypi.org/project/goslate/

acesso ao Tradutor o que acarretou na descontinuidade desta e outras API's que utilizavam mecanismos similares (GOSLATE, 2016).

Além disso, a tradução de arquivos começou a ser explorada comercialmente se tornando um serviço pago oferecido pelas gigantes da internet. Foi criado pelo Google o serviço Cloud Translation[3] que oferece API para a tradução de textos em diversos idiomas sob um custo de 20 dólares por milhão de caracteres, já a Amazon possui o serviço Amazon Translate[4] que também realiza traduções porém com um preço levemente inferior de 15 dólares por milhão de caracteres.

Diante deste cenário com uma média aritmética de aproximadamente 162.959 *tweets* por evento e considerando o pior cenário que seria um *tweet* com todos os seus caracteres utilizados (280), esta pesquisa teria em média 45.628.590 de caracteres por evento, acarretando em um grande custo para a execução da tradução: de US$ 2.738 a US$ 3.650.

Por este motivo a pesquisa reduziu a base de dados à um total de 20.000 *tweets*, tornando assim possível a utilização da plataforma gratuita do Google Tradutor para a tradução. Sendo que a base de cada evento deverá ser dividida em fragmentos ainda menores para serem submetidos ao website. Os detalhes da tradução serão tratados mais a frente, em sua respectiva seção.

A base foi dividida igualmente entre todos os eventos e cada um deles recebeu um total de 5.000 *tweets*. Estes, foram selecionados de maneira aleatória sem seguir critérios e regravados em um novo arquivo por ordem cronológica. Neste momento, o *Corpus* está formado e pronto para a primeira fase da mineração de texto: o pré-processamento.

4.3 Pré-Processamento

4.3.1 Duplicados e Retweet

Antes de apresentar como esta pesquisa irá tratar os duplicados e retweets é importante salientar a diferença entre estes dois conceitos. São considerados duplicados os tweets dentro de uma base de dados que são redundantes ou idênticos, carregando consigo a mesma mensagem e o mesmo código de identificação (id). Já os retweets são compartilhamentos de mensagens escritas por outros autores podendo ou não ter acréscimo de novas idéias. Esta pesquisa removeu os duplicados e manteve os retweets.

A remoção dos duplicados é necessária mesmo que a extração tenha sido feita em tempo real para que nenhum tweet seja contabilizado mais de uma vez por conta da redundância, atrapalhando assim os resultados. Os retweets por sua vez são contabilizados como novas mensagens porque pode-se dizer que quem as compartilhou integralmente e sem acréscimos, compactua com a mensagem antes dita revelando o mesmo sentimento que o autor original.

O pré-processamento se iniciou com a verificação dos duplicados para garantir que na base não haveria *tweets* redundantes. E após a varredura constatou-se a unicidade deles.

[3] https://cloud.google.com/translate/

[4] https://aws.amazon.com/pt/translate/

Como a extração foi feita por meio da API *Streamming* em que as mensagens são captadas em tempo real, não deveria mesmo haver mensagens repetidas. Porém, trabalhando com a hipótese de ocorrer falhas, tanto na captura como na gravação em arquivo, preferiu-se garantir via software a integridade da base.

As mensagens obtidas são como as mostradas na figura 10, abaixo. Há nelas a presença de elementos próprios da ferramenta Twitter como: o símbolo RT no início da frase indicando um retweet, o @ precedido das menções e o # das hashtags. Porém, além deste elementos é comum encontrar também emojis (imagens que representam emoções), abreviações, palavras escritas erroneamente, gírias e expressões próprias da internet.

Figura 10 – Exemplo de mensagens compartilhadas na rede social

```
RT @cirogomes: O Brasil não aguenta mais essa polarização. Garotos se agredindo na internet, famílias
se dividindo. Em todas as projeções eu venço no segundo turno, mas, para isso, eu preciso, de fato
chegar lá. Por isso peço seu voto para o número 12. #DebateSBT #UOLnasUrnas #Ciro12

As mulheres são a maioria da população brasileira e irão tirar o Brasil do atraso. Mais da metade da
população de mulheres do país rejeita Bolsonaro e estarão no próximo sábado em um grande ato por
#EleNão!Por @designersunivos #DesignAtivista #DebateSBT https://t.co/wEMYjptH6u
```

Fonte: Autora

A presença das letras indicadoras e do nome do usuário que a publicou originalmente além de não agregarem sentimento atrapalham o algoritmo classificador, portanto, através de expressões regulares toda a formatação 'RT @usuario :' foi removida de todos os textos.

4.3.2 Remoção de Links

No Twitter, os usuários podem anexar em suas postagens endereços virtuais que indicam o caminho para um objeto na rede, esses caminhos são conhecidos como *Uniform Resource Locator* (URL). Ao clicar nesses links o usuário é direcionado para imagens, gifs, sites ou vídeos. Porém, estes caminhos adicionados na mensagem não carregam consigo um conteúdo útil, tornando assim para esta pesquisa, irrelevante. A presença deles pode ainda atrapalhar o algoritmo classificador diminuindo a sua acurácia. Para identificar as URLs, as expressões regulares também foram utilizadas para a remoção de palavras que têm seu início por www, http, https. Assim, o resultado é o que pode ser visto a seguir:

Figura 11 – Mensagens após remoção de indentificadores de Retweet e Links

```
O Brasil não aguenta mais essa polarização. Garotos se agredindo na internet, famílias se dividindo. Em todas as
projeções eu venço no segundo turno, mas, para isso, eu preciso, de fato chegar lá. Por isso peço seu voto para o
número 12. #DebateSBT #UOLnasUrnas #Ciro12

As mulheres são a maioria da população brasileira e irão tirar o Brasil do atraso. Mais da metade da população de
mulheres do país rejeita Bolsonaro e estarão no próximo sábado em um grande ato por #EleNão!Por @designersunivos
#DesignAtivista #DebateSBT
```

Fonte: Autora

4.3.3 Tokenização e Remoção de Stopwords

A tokenização teve início com toda a base de dados já traduzida para o inglês. Como a tradução seria necessária, optou-se por traduzir as mensagens antes da remoção das *stopwords*,

já que a ausência delas poderia acarretar em traduções incorretas. Além da tradução, os dados passaram por uma normalização em que foram removidos os caracteres especiais, espaços duplos, tabulação e quebras de linha. O texto também foi padronizado em letras minúsculas.

Após a normalização o processo de tokenização têm início. Nele uma frase é separada em pequenos fragmentos chamados *tokens*, através dos limitadores. Nesta pesquisa, o limitador utilizado foi o caracter 'espaço' já que o português-brasileiro é uma língua delimitada por espaçamento.

Com toda a base de dados transformada em tokens o próximo passo foi remover as palavras *stopwords* das frases. Esta pesquisa reuniu uma lista com 174 palavras inglesas consideradas stopwords pelo site Ranks NL[5] (um website de ferramentas online recomendado por grandes empresas como IBM e MOZ). Algumas destas palavras podem ser vistas no quadro 7, abaixo. O algoritmo JAVA desenvolvido buscou e eliminou os tokens correspondentes as *stopwords* presentes na lista.

Quadro 7 – Fragmento da lista de stopwords

Stopwords					
about	above	after	again	from	here
both	too	am	an	further	down
and	any	are	aren't	how	out
as	at	be	because	same	that

Fonte: Autora

Como resultado da etapa de pré-processamento, têm-se as mensagens abaixo ilustradas pelo quadro 8.

Quadro 8 – Amostra de resultado do pré-processamento

Mensagens
debatenaglobo without cabodaciolo fraud vote51
night debate among presidential candidates relevant tweets realtime update political analysis memes oplease follow debatenaglobo elections2018
geraldo deserves hug pretty frustrating elections debatenaglobo
loved ciro feat marina start debate making voters aware really need current situation brazil debatenaglobo
debatenaglobo we going recover public finances want cut social rights serious cut right worker organize finances
thought standup comedy just debatenaglobo even got giggles audience background
bozo always running away debates showing everyone competence debate proposals vdd going shout debate debatenaglobo
see ciro marina debating comes hurt heart preparation enormous competence valued brazilians debatenaglobo
man meirelles little guy comedy kkk debatenaglobo
boulos sending sincere message brazilians seem think want risk return dictatorship really want dictatorship cironaglobo debatenaglobo

Fonte: Autora

[5] https://www.ranks.nl/stopwords

4.4 Indexação

As informações coletadas do Twitter estão armazenadas em quatro documentos JSON referentes a cada um dos eventos monitorados. Após a etapa de mineração já com os resultados da classificação das mensagens, todos os documentos foram armazenados em um banco de dados para melhor recuperação da informação e análise dos resultados obtidos.

4.5 Mineração

Para a realizar a etapa de mineração, esta pesquisa optou por utilizar uma aplicação desenvolvida especificamente para o Twitter que têm por objetivo classificar os tweets nas polaridades desejadas por esta pesquisa: O Sentiment140[6].

O Sentiment140 é uma aplicação online que mediante o login no Twitter permite descobrir o sentimento do momento (positivo, negativo ou neutro) sobre qualquer marca ou produto (SENTIMENT140, 2018). Ele nasceu a partir de um projeto de classe da Universidade de Stanford e está disponível gratuitamente para toda a comunidade. A aplicação já possui uma base de treinamento preparada para se realizar a análise de sentimentos através de algoritmos de máxima entropia. Esta é uma aplicação que opera na lingua inglesa e além disso, disponibiliza uma API que permite utilizar o classificador desenvolvido pelos alunos em outra ferramenta. Para usufruir é necessário um registro com um email válido e o aceite de alguns termos de uso. Posteriormente, deve-se escolher entre os três serviços oferecidos:

1. Simple Classification Service: Este serviço faz uma classificação simples de uma mensagem e para isso requer os parâmetros texto e linguagem. Podem ser inseridos opcionalmente: a função de retorno de chamada e o sujeito da mensagem para uma classificação direcionada. A resposta contém, o texto e o sujeito originais além da polaridade encontrada.

2. Bulk Classification Service (JSON): Os serviços em Bulk fazem uma classificação em lotes, dessa forma milhares de mensagens podem ser classificadas de uma só vez. Neste em especial deve ser passado um array JSON dentro de um objeto JSON com todas as mensagens. Como resposta, esse objeto será retornado ao requisitante com um campo adicionado em cada array, chamado *polarity* contendo a classificação obtida.

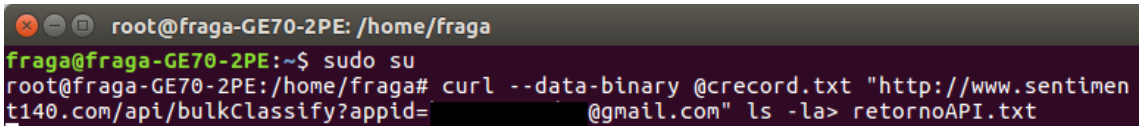
3. Bulk Classification Service (CSV): Da mesma forma que o anterior, porém mais simplório, este serviço requer um arquivo do tipo texto contendo uma mensagem por linha. Seu retorno é dado em um arquivo do tipo *Comma-separated values* (CSV) em que seus campos são separados por vírgulas. De forma que o primeiro campo é a classificação encontrada e o segundo, o texto original.

Devido à quantidade de *tweets* da base de dados desta pesquisa o terceiro serviço Bulk Classification Service (CSV) foi escolhido por conseguir processar até 10.000 mensagens por

[6] http://www.sentiment140.com

vez. Após executar o comando via terminal para enviar o arquivo .txt à aplicação (Figura 12), o retorno é um arquivo com apenas dois campos: a polaridade e o texto. A polaridade aparece representada por valores numéricos inteiros: 0 - Negativo, 2 - Neutro, 4 - Positivo. Enquanto o texto é replicado integralmente, como pode ser observado abaixo no quadro 9.

Figura 12 – Comando de envio de dados à aplicação Sentiment140



Fonte: Autora

Quadro 9 – Exemplo de retorno da aplicação Sentiment140

Retorno
"2","even alvaro dias wants call meireles kkkkkk debatenaglobo"
"2","lula haddad manunojaburu brazil comes development education eusoulula"
"2","day 07 17 junteseaobolsonaro"

Fonte: Autora

4.5.1 Tradução

O aplicativo Sentiment140 funciona somente na lingua inglesa portanto, toda a base de dados captada durante os eventos foi traduzida para que fosse submetida a classificação. Para a tradução do texto de Português-Brasileiro para o Inglês foi utilizado o Google Tradutor [1]. A ferramenta do Google permite o envio de arquivos para tradução como mostra a imagem 13 na opção *"traduza um documento"*.

Como o website do Google Tradutor não informa o tamanho limite do arquivo que pode ser submetido a tradução, foi realizado uma série de testes com arquivos de tamanhos diferentes e avaliado o resultado da tradução. Assim foram enviados arquivos do tipo texto (.txt) de 100, 200, 250, 300 e 400 linhas. Descobriu-se que os arquivos com 250 linhas ou mais, tiveram somente parte do documento traduzido, devolvendo assim, as últimas linhas na lingua original (Português-BR). Diante deste resultado, esta pesquisa dividiu toda a base em documentos com no máximo 200 linhas, o que permitiu a tradução de apenas 200 tweets por vez, já que cada linha representava um tweet.

[1] https://translate.google.com/

Figura 13 – Ferramenta de tradução gratuita

Fonte: (GOOGLE, 2018)

4.6 Análise

A fim de verificar a classificação automática feita pela aplicação Sentiment140, nesta etapa ocorreu uma interpretação humana dos dados. Assim é possível fazer uma comparação entre os resultados obtidos por ambas formas. Para isso, foram criados quatro novas bases de dados que correspondem a frações das bases já classificadas pertencentes à cada evento desta pesquisa. Para compor as bases foram selecionados *tweets* aleatórios até um limite de 100 mensagens por evento. Este limite foi estabelecido por esta ser uma tarefa penosa já que deve ser levada em conta a interpretação e subjetividade de cada mensagem em seu devido contexto.

Como as bases de classificação manual seriam classificadas por uma pessoa, elas permaneceram da forma publicada por seus autores e não foram submetidas a processos de pré-processamento e normalização, preservando assim a legibilidade. Com a base montada, a autora classifica independentemente os *tweets* nas categorias positivo, negativo e neutro. Após a classificação manual houve a comparação através dos números identificadores dos *tweets* selecionados, com os *tweets* classificados automaticamente pelo Sentiment140 a fim de verificar se sentimentos obtidos foram semelhantes ou divergentes.

5 Resultados e Discussões

O capítulo que se inicia expõe os principais resultados obtidos durante o desenvolvimento da pesquisa com a aplicação da abordagem proposta na base selecionada. Além disso, serão feitas discussões e análises diante destes dados.

A etapa de mineração obteve como resultado, a classificação automática dos *tweets* realizado através da aplicação Sentiment140, capaz de classificar mensagens textuais e retornar os sentimentos positivo, negativo e neutro. O quadro 10, exibe alguns exemplos dessas mensagens obtidas e também a sua polarização. Observa-se nas mensagens que os usuários argumentam sobre as propostas dos candidatos, citam as falas dos presidenciáveis ditas durante o debate, fazem comparações entre eles bem como demonstram apoio e rejeição.

Quadro 10 – Mensagens polarizadas

Mensagem	**Polaridade**
o Ciro é o ÚNICO que explica tudo direitinho sobre as propostas dele, falando que vai investir nisso,que vai fazer assim e tal, ELE EXPLICA. Os outros só falam que vão fazer, mas não explicam COMO vão fazer. Ciro, te amo lindão #DebateSBT	Positivo
Vamos fazer o Brasil ser feliz de novo, com duas palavras mágicas: trabalho e educação - @Haddad_Fernando #HaddadÉLulaNoSBT #DebateSBT #JilmarTatto #Tatto132	Positivo
Daciolo faz declaração de amor em debate: 'Mãe, te amo, varoa, te amo'#DebateSBTDaciolo. É assim que se destrói um mito. #CariuaSemArrudeios	Positivo
Gente, tô amando ver Boulos e Haddad metendo o pau no Bolsonazi juntos. VAMOS UNIR AS FORÇAS CONTRA ESSE LIXO, PELAMOR DE DEUS! #DebateNaGlobo	Positivo
terei que votar nele infelizmente queria o haddad mas é impossive	Negativo
Alkimin falando em governar para pobre é a maior piada de 2018 #DebateNaGlobo	Negativo
O EleNao descontextualizou a realidade, ignorou a rejeição do PT, subestimou a mulher e se ferrou!	Negativo
Bolsonaro foi cerceado de fazer campanha depois daquela tentativa de assassinato. Ele corre atrás com a internet e entrevistas gravadas.Chega a ser triste e desesperador ver 7 candidatos atacando covardemente as únicas formas que Bolsonaro tem para se expressar. #DebateNaGlobo	Negativo
Para de falar de Bolsonaro, Haddad... #DebateNaGlobo	Neutro
#Ciro12 vamos levar o @cirogomes para o segundo turno. #debatenarecord#Vote12	Neutro
Haddad @Haddad_Fernando no #DebateSBT: Imposto de Renda Justo Entenda a proposta do Imposto de Renda Justo,que isenta quem ganha até 5 salários mínimos de pagar imposto de renda.	Neutro
CIRO atenta sobre a necessidade de investimento estratégico! Meirelles concorda! O Meirelles vai chamar o CIRO? #TsunamiCiro12 #ciro12 #DebateNaGlobo	Neutro

Fonte: Autora

No quadro 11, estão apresentados a quantidade total de *tweets* analisados da base de dados correspondente à cada evento, juntamente com a quantidade e percentual respectivo de cada polaridade retornada pela análise. Ao total foram analisados 20000 *tweets*, e destes, 84,99% foram classificados como neutros, 8,18% como negativos e 6,84% como positivos. Assim, as mensagens que continham um sentimento explícito útil (positivo ou negativo) representaram 15% da base total analisada.

Quadro 11 – Resultado geral da classificação automática dos tweets

Evento	**Positivo**	**%**	**Negativo**	**%**	**Neutro**	**%**	**Total**
Candidatura Haddad	628	12.56	366	7.32	4006	80.12	5000
Debate do SBT	350	7	809	16.18	3841	76.82	5000
Debate da Record	127	2.54	9	0.18	4864	97.28	5000
Debate da Globo	262	5.24	451	9.02	4287	85.74	5000
Total	1367	6.84	1635	8.18	16998	84.99	20000

Fonte: Autora

Como descrito na seção 4.6, uma classificação humana dos dados foi realizada com 100 *tweets* publicados para cada um dos eventos, totalizando 400. O quadro 12 abaixo, compara os dois tipos de classificação: humana e automática. Através dessa comparação, pôde-se avaliar o algoritmo utilizado, Sentiment140, para que houvesse uma referência quanto ao seu desempenho. O algoritmo é desenvolvido para trabalhar com grandes volumes de informação, por isso, esta comparação utilizou uma amostra limitada com apenas 400 *tweets*. Devido a grandeza das informações se tornaria inviável fazer a classificação manual da base inteira.

Nesta comparação, percebeu-se uma diferença maior na quantidade de *tweets* classificados como positivos do que dos negativos. Das mensagens positivas compartilhadas, o algoritmo classificou apenas 5.75% como positivas enquanto a classificação humana considerou 13.75%, uma diferença de 8%. Quanto à totalidade de *tweets* negativos, o algoritmo também classificou um percentual menor, 7.75% enquanto a humana 13%, revelando uma diferença de 5.25% na classificação do sentimento negativo.

Quadro 12 – Resultado das classificações manual e automática

	Classificação Automática			**Classificação Manual**		
Evento	Positivo	Negativo	Neutro	Positivo	Negativo	Neutro
Candidatura Haddad	11	6	83	13	12	75
Debate do SBT	6	15	79	17	16	67
Debate da Record	3	0	97	6	6	88
Debate da Globo	3	10	87	19	18	63
Total	23	31	346	55	52	293
Total (%)	5.75	7.75	86.5	13.75	13	73.25

Fonte: Autora

A política é um tema que gera muita discussão e que muitas vezes pode ser acalorado. As pessoas defendem os seus pontos de vista, levantam os prós e contras tanto das propostas como das qualidades de seus candidatos. Oferecem o seu apoio ao favorito e distribuem críticas aos adversários. Fazem das palavras a sua maior arma para que a mensagem propagada se dissemine e consiga convencer os demais. Porém, o resultado das classificações mostrou que durante os debates e o evento de candidatura do Fernando Haddad, a maioria das mensagens publicadas eram fatos ou opiniões neutras, que não contêm sentimentos explícitos, contrariando a expectativa de se encontrar uma grande quantidade de sentimentos presentes nas mensagens.

A polaridade neutra prevaleceu com resultado acima dos 70% em ambas as classificações. Uma mensagem pode ser considerada neutra, devido à ausência de sentimentos expressos, ou também à presença equilibrada de sentimentos positivos e negativos. E em alguns casos, a polaridade neutra é atribuída a mensagens que não conseguiram ser classificadas nos outros sentimentos.

A classificação manual realizada pela autora, conseguiu encontrar 13,25% mais mensagens com sentimentos úteis do que o algoritmo. Esta diferença se deve à alguns fatores. O primeiro deles é a dificuldade de encontrar sentimentos em textos que segundo Liu (2010), é uma tarefa complexa.

A atualização do Twitter que permitiu a extensão das mensagens de 140 para 280 caracteres, possibilitou a publicação de mensagens longas. Essas mensagens longas, podem reunir variadas frases com sujeitos e sentimentos distintos. Assim, a mineração de texto fica mais complexa porque aglomera sentimentos que se referem a atores distintos em uma só mensagem. Nestes casos, algoritmos como o Sentiment140 faz uma análise do sentimento predominante, porém, se eles estiverem equilibrados a frase é caracterizada como neutra. A publicação à seguir de 202 caracteres, foi classificada como neutra e reúne diferentes orações e sujeitos:

"Dia 7, o Brasil tem uma chance de dar um passo rumo ao desenvolvimento, sem a esquerda no poder! Verás que um filho teu não foge à luta! A nossa bandeira jamais será vermelha! #DebateSBT #Bolsonaro17"

Outro fator que impacta na classificação dos sentimentos é a forma de escrita da internet. Ela dificulta ainda mais o processo, devido à linguagem predominantemente informal, uso de gírias, expressões populares e abreviações. Dessa forma, algumas palavras não foram traduzidas pelo tradutor e nem reconhecidas pelo classificador, interferindo no resultado. No quadro 13, estão dispostas algumas mensagens para exemplificação da linguagem informal comumente utilizada no ambiente online. Nelas podem ser observadas repetição de letras para causar sensação de intensidade, risadas (pela presença das repetições 'kkk'), abreviações (tô, tou, pq, fds, tá, qr, ein), expressões e gírias (birra, asco, pagando mico, pipoco, ta que ta) como também erros gramaticais (como 'maos' no lugar de 'mais').

Quadro 13 – Mensagens extraidas do Twitter

Mensagem
FOGOOOOO NO PARQUINHOOOOOO #DebateNaGlobo #Haddad13 #Boulos50
Kkkkkkk tô desse jeito! #HaddadÉ13
Cara, tou pegando birra desse Álvaro Dias. Daqui a pouco vira as asco. #DebateNaGlobo
O botox só pagando mico. Bota ele na linha Haddad. #DebateNaGlobo
Armas são ruins certo... pq todo politico anda com segurança armado ??? Pipoco na cara dos outros é refresco #DebateNaGlobo #BolsonaroNaRecord #Eleicoes2018
Cara, o Boulos ta que ta hj ein? #DebateNaGlobo
Rapaz o Hadad nao se reelegeu em São Paulo e qr ser presidente e prometendo mundos e fds, e olha q ele foi prefeito da cidade maos rica do país kkk #DebateNaGlobo @jairbolsonaro

Fonte: Autora

A diferença encontrada entre as classificações esta também relacionada com a curva

de aprendizado do algoritmo utilizado, que como qualquer outro, possui uma performance quanto à acertos e erros. A curva, é uma indicação da evolução do comportamento dos erros e da generalização, alcançados pelo algoritmo ao longo do treinamento da base. Como a taxa de erro existe, é esperado que existam mensagens classificadas errôneamente.

A tradução das mensagens para uma outra língua, também influenciou na divergência entre os resultados da classificação manual e automática, pois em uma tradução, corre-se o risco de haver perda do significado e sentido das palavras. O português do Brasil, possui um vocabulário muito extenso com uma grande variação das palavras que nem sempre conseguirão ser correspondidas na língua de destino. O quadro 14, apresenta alguns exemplos de mensagens classificadas erroneamente pelo algoritmo.

Quadro 14 – Mensagens classificadas erroneamente

Mensagem	**Class. Encontrada**	**Class. Verdadeira**
A hipocrisia e a arrogância petista é tão grande que o Andrade não citou a proposta indecente que o São Lula fez ao Ciro dias antes do registro das chapas! #Ciro12 #DebateSBT #CiroNoSBT	Positivo	Negativo
#DebateNaGlobo Quando o até o Meirelles é melhor que o bozo e o mamulengo do Lula.	Neutro	Positivo
o ciro é o candidato mais preparado, só não vê quem não quer #DebateNaGlobo	Neutro	Positivo
se a desculpa da galera é votar em alguém que defende valores cristãos, era pra votar no daciolo. mas a gente sabe que na verdade você escolhem o bololo por um razão: ele é a personificação dos preconceitos que vocês também tem! #DebateSBT	Positivo	Neutro
#DebateNaGlobonão sei se assisto o debate ou aprecio a surra de memes de Twitter to tipo:	Positivo	Neutro
Gostaria de fazer essa pergunta ao @jairbolsonaro, que está de alta, deu uma entrevista longa e fugiu do debate. Você acha correto que um homem que fugiu do debate tem capacidade de governar o Brasil? , pergunta Ciro Gomes ao candidato Henrique Meirelles no #DebateNaGlobo	Positivo	Negativo
O #DebateSBT foi a prova que o Bonoro não faz falta nos debates, quem realmente faz falta e toda a diferença é o Cabo Daciolo.Certeza absoluta do nosso entretenimento memeal.Te amo, varão.	Negativo	Positivo

Fonte: Autora

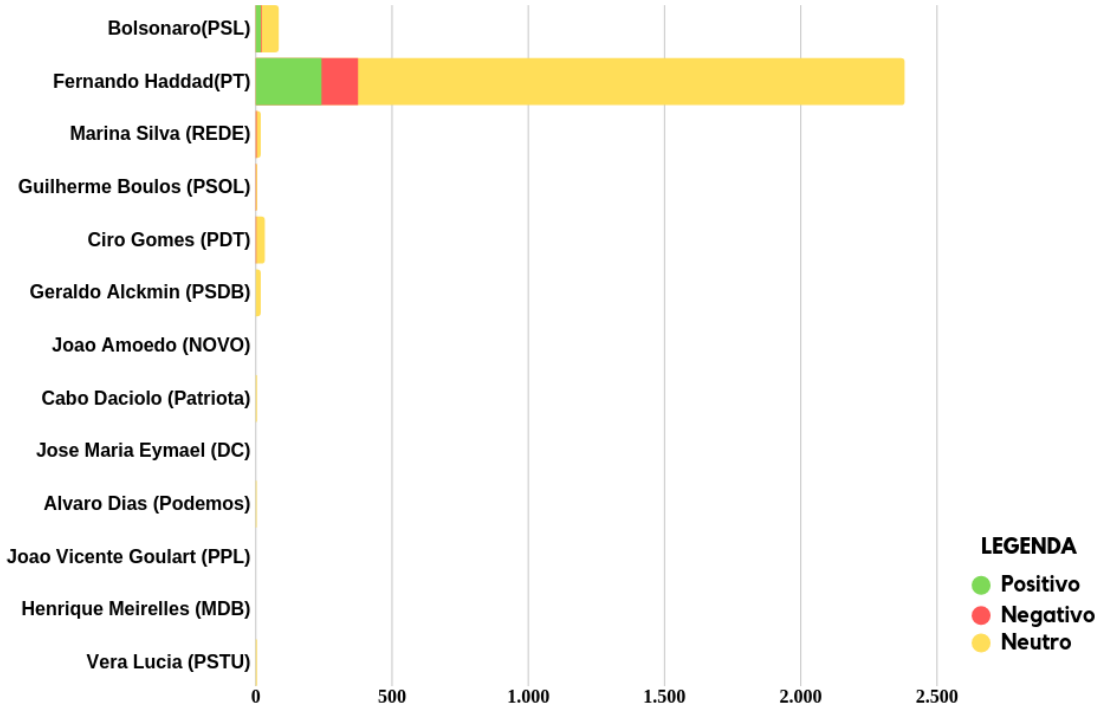
Foram encontradas também frases com a presença de ironia como a frase abaixo, classificada como neutra pelo algoritmo:

'De acordo com o PT, na época do bandido Lula o pobre andava de BMW e viajava para Dubai, o Nordeste era praticamente uma Dinamarca com calor e a Itaipava tinha gosto de Heineken.'

A ironia é um dos desafios da mineração de texto e é bem dificil para um algoritmo captá-la. No caso desta frase, apesar de todo esse sarcasmo, ela foi classificada corretamente. Porém, há casos em que a frase é classificada de forma literal e lhe é atribuído sentimento oposto.

A fim de identificar os sentimentos que estavam ao redor de cada presidenciável, foi elaborada uma imagem que representa o retrato do evento. Para isso, as mensagens foram separadas por sentimento e deveriam citar apenas um candidato para que fossem contabilizadas. Assim, as mensagens foram agrupadas por candidato e por sentimento, sendo representado pela cor verde o sentimento positivo, vermelho o negativo e amarelo o sentimento neutro. Os valores apresentados na imagem se referem à quantidade de *tweets*.

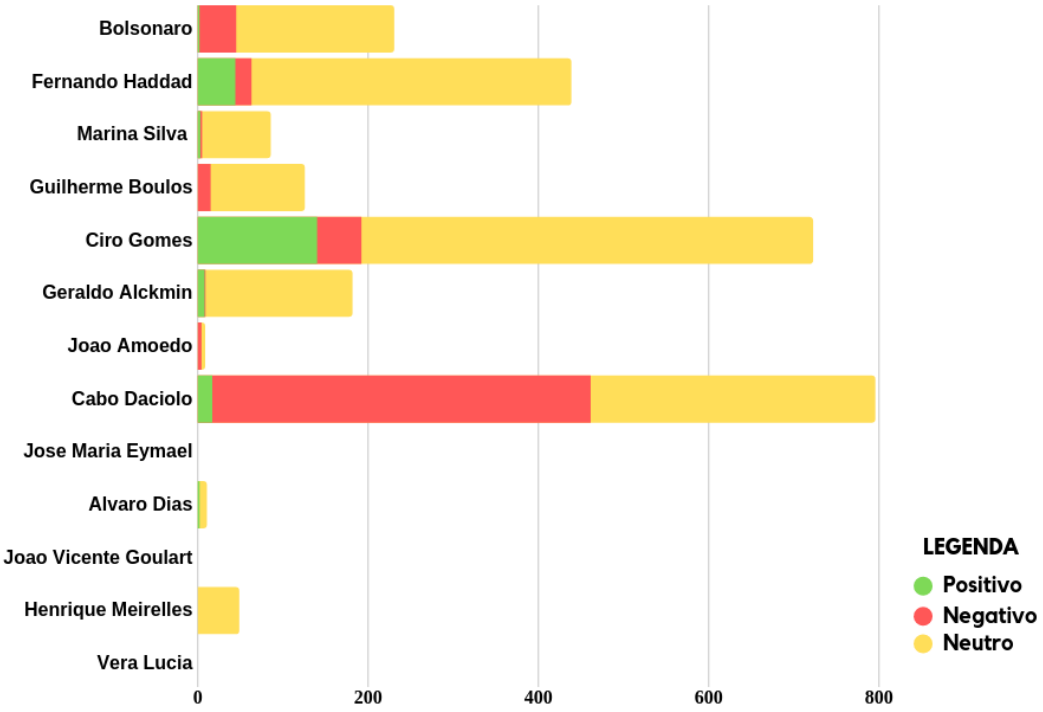
Figura 14 – Retrato da candidatura de Fernando Haddad



Fonte: Autora

Na Figura 14, é possível verificar que por se tratar do anúncio oficial da candidatura de Fernando Haddad como presidente pelo Partido dos Trabalhadores, as citações ao candidato Haddad foi majoritária recebendo mais de 2.400 *tweets*. Durante este dia, os *tweets* classificados como positivos que citavam o candidato, se sobressaíram aos *tweets* negativos. Houveram também citações menos expressivas à outros presidenciáveis como Bolsonaro, segundo candidato mais citado com pouco mais de 100 *tweets*, Marina Silva, Guilherme Boulos, Ciro Gomes, Geraldo Alckmin, Daciolo, Álvaro Dias e Vera Lúcia.

Figura 15 – Retrato do Debate do SBT



Fonte: Autora

Os candidatos mais citados durante o debate do SBT, foram Cabo Daciolo e Ciro Gomes. E os candidatos que mais receberam comentários positivos, foram respectivamente: Ciro Gomes, Haddad e Cabo Daciolo.

Daciolo, foi citado precisamente 795 vezes e ganhou grande repercussão neste dia após uma de suas declarações no debate, que gerou a mensagem mais retweetada:

'Uma pessoa que diz que vai tirar o FIES, o Bolsa Família, o ProUni, essa pessoa nunca passou necessidade', afirma Cabo Daciolo, no #DebateSBT. 'Com certeza sou favorável às cotas, e vou fazer um trabalho mais amplo. As pessoas tratam o ser humano como se fosse nada.'

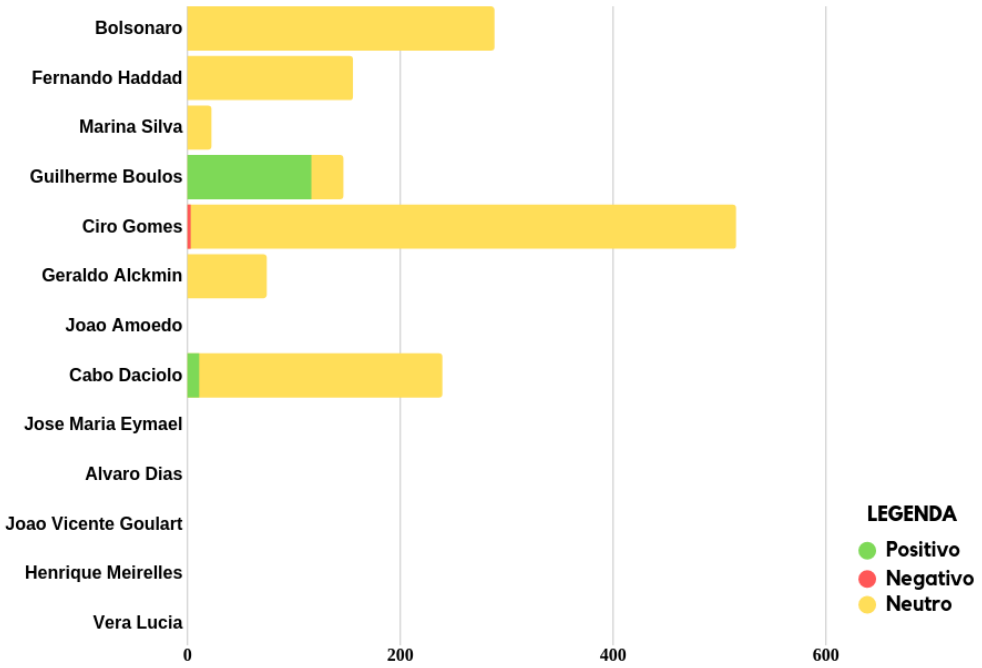
Esta mensagem, que cita Daciolo como autor da frase, recebeu sentimento negativo pelo classificador. Por este motivo, Daciolo possui neste retrato uma grande barra negativa que se refere quase em sua totalidade à este *tweet*. E apesar da classificação da mensagem ser negativa, a declaração do candidato foi muito bem vista pelo público.

Ciro Gomes também teve muitos *tweets* positivos (118 ao total), como mostra a figura, em decorrência da sua performance e desenvoltura durante o debate, explicando as suas propostas de forma clara e objetiva.

Bolsonaro aparece como quarto candidato mais citado e as mensagens negativas que o citaram sobressaíram às positivas.

Ao total, o debate do SBT recebeu oito candidatos: Haddad, Marina Silva, Guilherme Boulos, Ciro Gomes, Geraldo Alckmin, Daciolo, Álvaro Dias e Henrique Meirelles.

Figura 16 – Retrato do Debate da Record



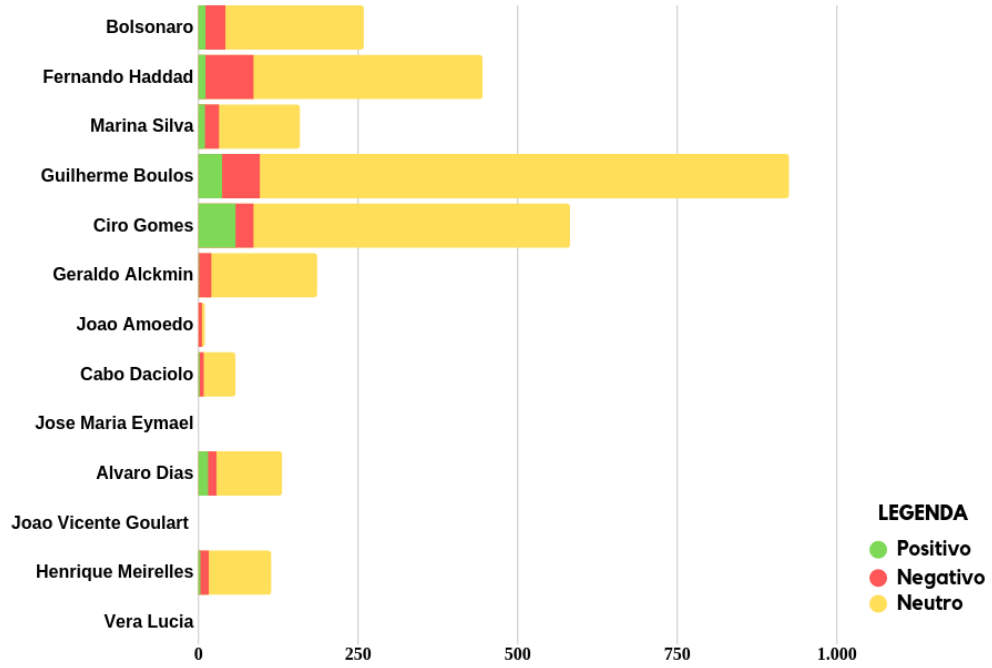
Fonte: Autora

O debate da Record dentre os demais debates foi o que apresentou o menor número de citações aos candidatos e o que apresentou a menor quantidade de sentimentos classificados. Os candidatos mais citados neste evento foram Ciro Gomes, Bolsonaro e Cabo Daciolo. Ciro Gomes, porém não recebeu nenhuma citação positiva neste debate só *tweets* negativos e neutros. Bolsonaro recebeu apenas *tweets* neutros e apesar de não ter comparecido ao debate foi o segundo mais comentado no Twitter.

Guilherme Boulos se destacou positivamente pois 80% dos *tweets* que o citavam eram positivos. E Cabo Daciolo recebeu apenas *tweets* neutros e positivos.

O debate da Record recebeu novamente os oito candidatos: Haddad, Marina Silva, Guilherme Boulos, Ciro Gomes, Geraldo Alckmin, Daciolo, Álvaro Dias e Henrique Meirelles.

Figura 17 – Retrato do Debate da Globo



Fonte: Autora

O Debate da Globo apresentou os sentimentos explícitos de forma mais distribuída, diferentemente do que aconteceu no debate da Record com Guilherme Boulos ou como o Cabo Daciolo no debate do SBT. Todos os candidatos que foram citados possuíram sentimentos positivos e negativos.

Os candidatos mais citados neste debate foram Guilherme Boulos e Ciro Gomes. O candidato mais citado em *tweets* positivos foi novamente Ciro e o mais citado em *tweets* negativos foi o Fernando Haddad.

Guilherme Boulos repercutiu muito no Twitter com a sua fala sobre a ditadura provocando este destaque e superando os demais quanto às citações:

'Não dá para fingir que está tudo bem. Nós saímos de uma ditadura em que muita gente morreu. Quando eu nasci, estava numa Ditadura, e não quero que minhas filhas cresçam numa. Ditadura nunca mais', diz Boulos #Eleições2018 #DebateNaGlobo

O debate da Globo recebeu apenas sete dos treze candidatos: Haddad, Marina Silva, Guilherme Boulos, Ciro Gomes, Geraldo Alckmin, Álvaro Dias e Henrique Meirelles.

5.1 Considerações Finais

Quando se analisou o resultado do monitoramento dos debates percebeu-se que o Cabo Daciolo foi muito citado no debate do SBT e na Record, porém, o mesmo não foi convidado para o debate da Globo provocando uma queda drástica com o número de suas citações únicas nas mensagens. Entretanto, o mesmo não ocorreu com Bolsonaro. O candidato do PSL esteve ausente em todos os debates, porém, a quantidade de mensagens que o citavam foram sempre significativas quando comparadas aos candidatos presentes. Durante os debates, os adversários fizeram menções e criticas a Bolsonaro, o que acabou trazendo ele para o debate refletindo nas mensagens compartilhadas pelos usuários e espectadores. Já Daciolo, como não estava presente fisicamente e ninguém o mencionou, acabou não sendo tão lembrado como Bolsonaro.

Através deste estudo foi possível se confirmar que o resultado obtido através das mensagens da rede social, se referem aos sentimentos dos usuários no momento dos eventos, já que foi possível captar a reação instantânea do público às frases e discursos dos candidatos durante o debate. Havia uma expectativa de se conseguir uma quantidade maior do que de apenas 15% de sentimentos explícitos. Devido à esta taxa, a base composta por 20.000 *tweets* conseguiu um número bem limitado de sentimentos úteis para análise, mas que ainda foram suficientes para tal.

A primeira vista, a idéia de se extrair sentimentos de textos pareceu bem complexa e desafiadora. Ao longo deste estudo, porém, foram encontradas metodologias e técnicas que resgataram os ânimos e tornaram o objetivo atingível. Um tema muito tecnológico como esse envolve também desafios quanto à inovação. Durante o processo, a área de desenvolvedores do Twitter passou por modificações e foi quase totalmente alterada, desde o *layout* externo do site como as regras de acesso da API. As informações presentes nas redes sociais estão muito visadas porque passaram a ter valor comercial e as plataformas estão se adaptando à esta nova realidade. Dessa forma, o acesso aos dados têm se tornado cada vez mais difícil e restrito, concedidos somente mediante pagamento. Por outro lado, houve uma maior atenção quanto a documentação e instrução para os desenvolvedores. A princípio o site contava apenas com as informações essenciais que ainda deixavam algumas dúvidas quanto à execução, após a atualização ele ficou muito mais completo com tutoriais e guias, para melhor auxiliar os desenvolvedores em seus novos projetos.

Mesmo com toda a inovação e com as mudanças constantes na área tecnológica, os seres humanos continuarão a se expressar por meio de palavras e a área de análise de sentimentos têm ainda muito à evoluir.

6 Conclusão

Este estudo explorou os temas de redes sociais, extração de dados e mineração de textos para descobrir como extrair sentimentos de textos publicados em uma rede social. Para o seu desenvolvimento eram necessárias a escolha de uma rede social e um problema prático para que pudesse ser extraída informações úteis. A rede social escolhida foi o Twitter, por possuir aspectos favoráveis à análise de sentimentos e a política estava entre os temas mais relevantes para uma aplicação prática devido ao ano de 2018 ser um ano eleitoral.

Dessa forma, o objetivo desse estudo foi de extrair sentimentos de textos publicados por usuários de uma rede social. E para isso, foram realizadas as etapas de mineração de textos: coleta de dados, pré-processamento, indexação, mineração e análise. O objetivo proposto foi realizado com sucesso. A extração de dados efetuada, permitiu a criação de uma base de dados volumosa que reuniu mais de 650.000 *tweets*. A base de dados final composta por 20.000 *tweets*, passou por tratamentos de pré-processamento e tradução sendo posteriormente classificada em sentimentos através da ferramenta de análise de sentimentos Sentiment140 e os resultados retornados pelo algoritmo de classificação automática foi comparado com uma classificação manual realizada pela autora.

Como resultado, este estudo obteve uma classificação terciária em que as mensagens publicadas pelos usuários de uma rede social foram polarizadas nos sentimentos positivo, negativo e neutro. E os sentimentos úteis, positivos e negativos, foram identificados em 15% do total de mensagens classificadas. A classificação automática divergiu da classificação manual em 8% para menos nas mensagens positivas e 5.25% para menos nas mensagens negativas. Em contrapartida para as neutras a divergência foi de 13.25% a mais para a classificação automática. Levando em conta a dificuldade existente no processo de trabalhar com textos, a taxa de erro existente nos algoritmos de AS, a influência da tradução, da linguagem informal e da normalização da base de dados, pode-se dizer que a classificação automática teve uma performance satisfatória quando comparada à classificação manual realizada pela autora.

Acredita-se que este estudo pode ser considerado um alicerce para futuros trabalhos da análise de sentimentos, favorecendo uma maior exploração e evolução desta área no meio acadêmico já que apresentou um método viável de se conseguir obter sentimentos de textos publicados em redes sociais. E a partir da experiência relatada, os trabalhos podem aprimorar as etapas de mineração de textos e também explorar outros algoritmos de aprendizado de máquina utilizados na análise de sentimentos.

Foram encontradas durante a implementação deste estudo algumas limitações. Apesar de se ter conseguido extrair um volume considerável de publicações da rede para análise, o tamanho da base necessitou ser reduzido para 20.000 *tweets* devido à etapa de tradução. A princípio a tradução iria ser executada em toda a base de dados de forma automática, porém o serviço a ser utilizado foi descontinuado inviabilizando o processo. Então, a base precisou ser reduzida para ser dividida em pequenos documentos e ser enviada de forma manual à

tradução através da plataforma do Google Tradutor.

A classificação manual também pode ser considerada uma etapa limitante, pois a autora teve que classificar os *tweets* nas polaridades positiva, negativa e neutra. E por requerer uma avaliação da subjetividade das mensagens e isso demandar muito esforço, foi determinada uma quantidade máxima de 100 *tweets* por evento.

7 Trabalhos Futuros

O presente estudo conseguiu através do método escolhido, atingir o seu objetivo de classificar mensagens de uma rede social em sentimentos positivos, negativos e neutros. Assim, mostrou-se a viabilidade de se conseguir obter sentimentos através de textos. Portanto, para trabalhos futuros aplicações mais objetivas que utilizem análise de sentimentos podem ser definidas para se descobrir ou obter um panorama sobre a opinião dos usuários sobre determinado tema ou assunto e posteriormente conectá-los com fatos ou dados reais.

Algumas melhorias para esta continuidade ainda podem ser implementadas como:

- Realizar a análise de sentimentos utilizando um algoritmo de aprendizado de máquina como *Naive Bayes* ou *Support Vector Machine* utilizando uma base de treinamento na língua de origem como o português brasileiro, para que não haja perda de acurácia com traduções.
- Aprimorar a performance do algoritmo incrementando tratamentos léxicos na fase de pré-processamento. Normalizando a base quanto à presença de repetição de caracteres, diminutivo, aumentativo e derivações, presença de gírias e palavras escritas errôneamente.
- Expandir a pesquisa para outras ferramentas como o Facebook.

Referências

ANTHONY, M.; BARTLETT, P. L. *Neural network learning: Theoretical foundations*. [S.l.]: cambridge university press, 2009. Nenhuma citação no texto.

ARANHA, C. N.; VELLASCO, M.; PASSOS, E. Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. *Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ*, 2007. Citado nas páginas 33, 34, 35, 36, 37 e 39.

ARAÚJO, M. et al. Métodos para análise de sentimentos no twitter. In: *Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)*. [S.l.: s.n.], 2013. Citado na página 19.

ATTUX, R. C. Predição dos resultados das eleições 2014 para presidente do brasil usando dados do twitter. Universidade Federal de Uberlândia, 2017. Citado na página 19.

BALAZS, J. A.; VELáSQUEZ, J. D. Opinion Mining and Information Fusion: A survey. *Information Fusion*, Elsevier B.V., v. 27, p. 95–110, 2016. ISSN 15662535. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2015.06.002>. Citado na página 12.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. *Porto Alegre: Sociedade Brasileira de Computação*, 2011. Citado nas páginas 22 e 23.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. *Métodos para análise de sentimentos em mídias sociais*. 2015. Citado nas páginas 16, 17, 18 e 38.

BERMINGHAM, A.; SMEATON, A. On using twitter to monitor political sentiment and predict election results. In: *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*. [S.l.: s.n.], 2011. p. 2–10. Citado na página 18.

BOYD, d. m.; ELLISON, N. B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, Blackwell Publishing Inc, v. 13, n. 1, p. 210–230, 2007. ISSN 1083-6101. Disponível em: <http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x>. Citado nas páginas 21 e 22.

BUCCAFURRI, F. et al. Comparing Twitter and Facebook user behavior: Privacy and other aspects. *Computers in Human Behavior*, Elsevier Ltd, v. 52, p. 87–95, 2015. ISSN 07475632. Disponível em: <http://dx.doi.org/10.1016/j.chb.2015.05.045>. Citado nas páginas 27 e 32.

CAVALIN, P. R. et al. Real-time sentiment analysis in social media streams: The 2013 confederation cup case. *Proceedings of BRACIS/ENIAC*, 2014. Citado na página 18.

CHAMLERTWAT, W. et al. Discovering consumer insight from twitter via sentiment analysis. *J. UCS*, v. 18, n. 8, p. 973–992, 2012. Citado na página 16.

CHEE, B.; KARAHALIOS, K. G.; SCHATZ, B. Social visualization of health messages. In: IEEE. *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. [S.l.], 2009. p. 1–10. Citado na página 16.

CHEW, C.; EYSENBACH, G. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, Public Library of Science, v. 5, n. 11, p. e14118, 2010. Citado na página 16.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 16.

DIFFEN. *Facebook vs. Twitter*. 2016. <https://www.diffen.com/difference/Facebook_vs_Twitter>. Acessado em 20 de Novembro de 2018. Citado na página 27.

DIJCK, J. V. *The culture of connectivity: A critical history of social media*. [S.l.]: Oxford University Press, 2013. Citado na página 23.

ELLISON, N. B.; STEINFIELD, C.; LAMPE, C. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of computer-mediated communication*, Oxford University Press Oxford, UK, v. 12, n. 4, p. 1143–1168, 2007. Citado na página 21.

EVANGELISTA, T. R.; PADILHA, T. P. P. Monitoramento de posts sobre empresas de e-commerce em redes sociais utilizando análise de sentimentos. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2013. Citado na página 16.

FACEBOOK. *Graph API*. 2018. <https://developers.facebook.com/docs/graph-api/>. Acessado em 8 de Dezembro de 2018. Citado na página 31.

FERREIRA, G. C. Redes sociais de informação: uma história e um estudo de caso. *Perspectivas em Ciência da Informação*, SciELO Brasil, v. 16, n. 3, p. 208–231, 2011. Citado nas páginas 12 e 21.

FRANÇA, T. C. et al. Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. _____ *Tópicos em Gerenciamento de Dados e Informações*, p. 8–45, 2014. Citado na página 28.

GHIASSI, M.; SKINNER, J.; ZIMBRA, D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, Elsevier, v. 40, n. 16, p. 6266–6282, 2013. Citado na página 16.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, v. 1, n. 12, 2009. Citado na página 19.

GONÇALVES, P. et al. Panas-t: Uma escala psicométrica para mediç ao de sentimentos no twitter. *Departamento de Computao, Universidade Federal de Ouro Preto (UFOP). Ouro Preto, MG, Brasil*, 2012. Citado na página 19.

GOOGLE. *Tradutor*. 2018. <https://translate.google.com/>. Acessado em 13 de Outubro de 2018. Citado na página 49.

GOOGLE. *Veja o que o mundo está pesquisando*. 2019. <https://trends.google.com.br/trends/?geo=BR>. Acessado em 5 Fevereiro 2019. Citado na página 14.

GOSLATE. *Project description*. 2016. <https://pypi.org/project/goslate/>. Acessado em 31 de Outubro de 2018. Citado na página 44.

GRASSI, M. et al. Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation*, Springer, v. 3, n. 3, p. 480–489, 2011. Citado na página 16.

HARDT, D. *The OAuth 2.0 authorization framework*. [S.l.], 2012. Citado na página 30.

HASSAN, A.; QAZVINIAN, V.; RADEV, D. What's with the attitude?: identifying sentences with attitude in online discussions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2010. p. 1245–1255. Citado na página 16.

HEARST, M. What is text mining. *SIMS, UC Berkeley*, 2003. Citado nas páginas 32 e 33.

HEARST, M. A. Untangling text data mining. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. [S.l.], 1999. p. 3–10. Citado na página 32.

INDURKHYA, N.; DAMERAU, F. J. *Handbook of natural language processing*. [S.l.]: CRC Press, 2010. v. 2. Citado na página 35.

JACOBSON, D.; BRAIL, G.; WOODS, D. *APIs: A Strategy Guide: Creating Channels with Application Programming Interfaces*. O'Reilly Media, 2011. ISBN 9781449321642. Disponível em: <https://books.google.com.br/books?id=om5tNwKW4xkC>. Citado nas páginas 27 e 28.

JSON. *Introducing JSON*. 2018. <https://www.json.org/>. Acessado em 8 de Dezembro de 2018. Citado na página 30.

JUNIOR, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. *Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro*, 2007. Citado nas páginas 36 e 37.

KANTARDZIC, M. *Data mining: concepts, models, methods, and algorithms*. [S.l.]: John Wiley & Sons, 2011. Citado na página 32.

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, v. 53, n. 1, p. 59–68, 2010. ISSN 00076813. Citado nas páginas 12 e 22.

KORB, K. B.; NICHOLSON, A. E. *Bayesian artificial intelligence*. [S.l.]: CRC press, 2010. Citado na página 17.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised leaning. *International Journal of Computer Science*, Citeseer, v. 1, n. 2, p. 111–117, 2006. Citado nas páginas 34 e 35.

LEIBA, B. Oauth web authorization protocol. *IEEE Internet Computing*, IEEE Computer Society, Los Alamitos, CA, USA, v. 16, p. 74–77, 2012. ISSN 1089-7801. Citado na página 29.

LIU, B. Sentiment analysis and subjectivity. *Handbook of natural language processing*, v. 2, p. 627–666, 2010. Citado nas páginas 13, 14, 18, 37, 38, 40 e 52.

MAKICE, K. *Twitter API: Up and Running: Learn How to Build Applications with the Twitter API*. O'Reilly Media, 2009. ISBN 9780596555511. Disponível em: <https://books.google.com.br/books?id=kBXXZNcRCGcC>. Citado na página 27.

MALHEIROS, Y.; LIMA, G. Uma ferramenta para análise de sentimentos em redes sociais utilizando o senticnet. 2013. Citado nas páginas 15 e 16.

MALINI, F.; CIARELLI, P.; MEDEIROS, J. O sentimento político em redes sociais: big data, algoritmos e as emoções nos tweets sobre o impeachment de dilma rousseff| political sentiment in social networks: big data, algorithims and emotions in tweets about the impeachment of dilma rousseff. *Liinc em Revista*, v. 13, n. 2, 2017. Citado na página 16.

MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999. Citado na página 35.

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. Citado nas páginas 16, 17, 37 e 38.

MENCZER, F.; PANT, G.; SRINIVASAN, P. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, ACM, v. 4, n. 4, p. 378–419, 2004. Citado na página 34.

MURPHY, K. P. Machine learning: A probabilistic perspective. The MIT Press, 2012. Citado na página 16.

NASCIMENTO, P. et al. Análise de sentimento de tweets com foco em notícias. In: *Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2012. Citado na página 16.

NIGAM, K.; LAFFERTY, J.; MCCALLUM, A. Using maximum entropy for text classification. In: *IJCAI-99 workshop on machine learning for information filtering*. [S.l.: s.n.], 1999. v. 1, p. 61–67. Citado na página 17.

OAUTH. *OAuth Community Site*. 2017. <https://oauth.net/>. Acessado em 24 de Julho de 2017. Citado na página 29.

PAL, S. K.; TALWAR, V.; MITRA, P. Web mining in soft computing framework: relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, IEEE, v. 13, n. 5, p. 1163–1177, 2002. Citado na página 36.

PANG, B.; LEE, L. et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008. Citado nas páginas 14 e 16.

PANT, G.; SRINIVASAN, P.; MENCZER, F. Crawling the web. In: *Web Dynamics*. [S.l.]: Springer, 2004. p. 153–177. Citado na página 34.

PARROTT, W. G. *Emotions in social psychology: Essential readings*. [S.l.]: Psychology Press, 2001. Citado na página 18.

PIOVESAN, A.; TEMPORINI, E. R. Pesquisa exploratória: procedimento metodológico para o estudo de fatores humanos no campo da saúde pública. *Saúde Pública*, Departamento de Prática de Saúde Pública da Faculdade de Saúde Pública - USP, 1995. Citado na página 39.

RUSSELL, M. A. *Mining the Social Web*. [S.l.: s.n.], 2013. 444 p. ISSN 16130073. ISBN 978-1449367619. Citado nas páginas 12 e 29.

SCHMIDT, D. C.; BUSCHMANN, F. Patterns, frameworks, and middleware: their synergistic relationships. In: IEEE. *Software Engineering, 2003. Proceedings. 25th International Conference on*. [S.l.], 2003. p. 694–704. Citado na página 28.

SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. *Introduction to information retrieval*. [S.l.]: Cambridge University Press, 2008. v. 39. Citado na página 36.

SENTIMENT140. *For Academics*. 2018. <http://help.sentiment140.com/for-students>. Acessado em 13 de Outubro de 2018. Citado na página 47.

SIGNIFICADOS. *Significado de Emoji*. 2018. <https://www.significados.com.br/emoji/>. Acessado em 01 de Novembro de 2018. Citado na página 19.

SOARES, F. de A. *Categorização Automática de Textos Baseada em Mineração de Textos*. 2013. Tese (Doutorado) — PUC-Rio, 2013. Citado na página 36.

SOCIAL, W. A.; HOOTSUIT. *Digital in 2018 in southern America*. 2018. <https://pt.slideshare.net/wearesocial/digital-in-2018-in-southern-america-part-1-north-86863727>. Acessado em 01 de Março de 2018. Citado nas páginas 14, 24, 31 e 39.

SOUZA, A. O. Identificação e classificação de usuários influentes no twitter por áreas de influência. Centro Federal de Educação Tecnológica de Minas Gerais, 2015. Citado na página 18.

SOUZA, Q. R.; QUANDT, C. O. Metodologia de análise de redes sociais. *O Tempo das Redes*, Perspectiva, p. 31–63, 2008. Citado na página 12.

STIEGLITZ, S.; DANG-XUAN, L. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In: IEEE. *System Science (HICSS), 2012 45th Hawaii International Conference on*. [S.l.], 2012. p. 3500–3509. Citado na página 16.

TAN, A.-H. et al. Text mining: The state of the art and the challenges. In: SN. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. [S.l.], 1999. v. 8, p. 65–70. Citado na página 32.

TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, v. 2011, n. 8, p. 1–13, 2011. ISSN 16469895. Citado nas páginas 19 e 22.

TWITTER. *Getting started*. 2018. <https://developer.twitter.com/en/docs/basics/things-every-developer-should-know/>. Acessado em 8 de Dezembro de 2018. Citado na página 32.

TWITTER. *our values*. 2018. <https://about.twitter.com/en_us/values.html>. Acessado em 20 de Novembro de 2018. Citado na página 25.

TWITTER. *Twitter*. 2018. <https://twitter.com/>. Acessado em 26 de Outubro de 2018. Citado na página 42.

VERGARA, S. C. Tipos de pesquisa em administração. Escola Brasileira de Administração Pública da FGV, 1990. Citado na página 39.

VYDISWARAN, V.; ZHAI, C.; ROTH, D. Gauging the internet doctor: ranking medical claims based on community knowledge. In: ACM. *Proceedings of the 2011 workshop on Data mining for medicine and healthcare*. [S.l.], 2011. p. 42–51. Citado na página 16.

WANG, H. et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL 2012 System Demonstrations*. [S.l.], 2012. p. 115–120. Citado na página 18.

WEBTEGRITY. *WHAT'S THE DIFFERENCE BETWEEN FACEBOOK AND TWITTER?* 2014. <https://webtegrity.com/our-blog/social-media-marketing/whats-difference-between-facebook-and-twitter/>. Acessado em 20 de Novembro de 2018. Citado na página 26.

WEISS, S. M. et al. *Text mining: predictive methods for analyzing unstructured information*. [S.l.]: Springer Science & Business Media, 2010. Citado na página 33.

WILBUR, W. J.; SIROTKIN, K. The automatic identification of stop words. *Journal of information science*, Sage Publications Sage CA: Thousand Oaks, CA, v. 18, n. 1, p. 45–55, 1992. Citado na página 36.

YOUTUBE. *Youtube*. 2019. <https://www.youtube.com/intl/pt-BR/yt/about/>. Acessado em 5 Fevereiro 2019. Citado na página 24.

ZHANG, H. The optimality of naive bayes. *AA*, v. 1, n. 2, p. 3, 2004. Citado na página 17.