UNIVERSIDADE FEDERAL DE ALAGOAS

INSTITUTO DE COMPUTAÇÃO

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

GABRIEL DE MENEZES SILVA

MICHELLE BERNARDINO DA SILVA

UMA FERRAMENTA PARA ANÁLISE DE SENTIMENTO PARA O TWITTER

Arapiraca – AL

2020

GABRIEL DE MENEZES SILVA

MICHELLE BERNARDINO DA SILVA

UMA FERRAMENTA PARA ANÁLISE DE SENTIMENTO PARA O TWITTER

Trabalho de Conclusão de Curso submetido ao Curso de Sistemas de Informação do Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do Grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Rodolfo Carneiro Cavalcante

Arapiraca – AL

2020

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Valter dos Santos Andrade

S586a Silva, Gabriel de Menezes.
Uma ferramenta para análise de sentimento para o Twitter / Gabriel de Menezes Silva, Michelle Bernardino da Silva, Maceió – 2022.
42 f. : il.

Orientador: Rodolfo Carneiro Cavalcante.
Monografia (Trabalho de Conclusão de Curso em Sistemas de Informação) – Universidade Federal de Alagoas, Instituto de Computação, Maceió, 2022.

Bibliografia: f. 40-42.

1. Mineração de textos. 2. Mineração de Opinião. 3. Aprendizado de Máquina. 4. Processamento de linguagem natural. 5. Twitter (Rede social on-line). 6. Análise de sentimento. I. Título.

CDU: 004

DEDICATÓRIA

Dedico este trabalho aos meus pais, Cícero André e Verônica Menezes, familiares e amigos que sempre estiveram presentes em minhas batalhas diárias, incentivando-me a conseguir alcançar meus objetivos e me permitindo fazer sonhar com grandes conquistas. Não poderia deixar de dedicar este trabalho ao Prof. Rodolfo Carneiro que teve fundamental importância na elaboração do mesmo, sempre disposto a nos ajudar.

Gabriel de Menezes Silva

Dedico este trabalho aos meus pais, Damião Pedro e Maria Hélia, por todos os esforços que tiveram e por tudo que fizeram por mim ao longo de minha vida. A minha irmã, Danielle, por ter me incentivado a não desistir diante das dificuldades. Por fim , não menos importante, ao meu orientador, Prof. Rodolfo Carneiro, por aceitar ser meu orientador e ter tido paciência, competência e dedicação comigo.

Michelle Bernardino da Silva

AGRADECIMENTOS

Agradecemos a Deus pelo dom de nossas vidas, por todas pessoas especiais que tiveram fundamental importância, para que este trabalho tenha sido feito, que ele tem nos mostrado.

Agradecemos aos nossos pais que sempre nos apoiaram e nos deram condições para que pudéssemos lutar pelos nossos objetivos.

Queremos agradecer aos nossos amigos discentes que sempre nos auxiliaram e nos motivaram a continuar estudando mesmo durante tantas adversidades que os contratempos têm nos mostrado.

Eu, Gabriel de Menezes Silva, agradeço a Deus pelo maior presente que ele poderia me ofertar, meu filho, Pedro Gabriel, e também a minha esposa que me dão coragem a cada dia de lutar por uma vida melhor.

Agradecemos aos tutores, em especial, a Kelciane Cavalcante de Lima e a Nayara Rosy Félix da Silva, que sempre foram muito solícitas e nos ajudaram a conseguir este êxito de concluir a graduação.

Agradecemos ao professor Rodolfo Carneiro Cavalcante que sempre nos ajudou, em todas as etapas deste trabalho, com sua experiência e boa vontade em servir ao próximo. Que Deus abençoe ao senhor, Professor, somos eternamente gratos pela sua contribuição.

A toda coordenação do curso que nunca nos abandonou e fez com que este momento tão esperado virasse realidade.

EPíGRAFE

“A maior recompensa para o trabalho do homem não é o que ele ganha com isso, mas o que ele se torna com isso.”

(John Ruskin)

RESUMO

O crescimento das redes sociais proporcionou algumas vantagens, dentre as quais podemos destacar o compartilhamento de informações entre os usuários em um curto espaço de tempo. A postagem de opiniões sobre diferentes eventos que estão em alta no momento é comum entre os usuários das redes sociais, tendo como consequência um volume crescente de dados diariamente. No entanto, toda essa informação gerada por meio destas postagens pode ser de grande utilidade, se a mesma for tratada corretamente gerando assim um conhecimento que poderá ser utilizado em diversas áreas. Este trabalho apresenta um estudo sobre análise de sentimento, mais especificamente, tendo como base de estudo a rede social Twitter, por esta conter recursos que facilitam a coleta de dados. Para a composição do trabalho utilizou-se um dataset (conjunto de dados) contendo informações que foram coletadas do Twitter por pessoas que trabalham para o governo de Minas Gerais, ou seja, as postagens dos usuários, conhecidas como tweets. Esses dados passaram por um processamento de linguagem natural a fim de remover *stopwords* (palavras irrelevantes para o sentido de determinada informação) e caracteres indesejados. Foi utilizado o algoritmo de aprendizagem de máquina Naive Bayes para classificação de sentimento em postagens. A ferramenta utilizada apresentou, apesar de ser utilizada uma base de teste simples, uma acurácia satisfatória de 86,85% na fase de teste. Posteriormente a ferramenta foi utilizada nos tweets recuperados e também mostrou-se bastante usual, conseguindo fazer a predição de todos os tweets recuperados.

Palavras-chave: Mineração de Textos; Mineração de Opinião; Aprendizado de Máquina; Processamento de Linguagem Natural; Twitter; Redes Sociais.

ABSTRACT

The growth of social networks has provided some advantages among which we can highlight the sharing of information among users in a short space of time. The posting of opinions on different events that are currently on the rise and common among users of social networks results in a growing volume of data on a daily basis. However, this information generated through these posts can be very useful if they are treated correctly, thus generating knowledge that can be used in different areas. This article presents a study on sentiment analysis, more specifically based on the social network Twitter because it contains resources that facilitate data collection. For the composition of the work, a dataset was used with data collected from Twitter by people who work for the government of Minas Gerais, that is, the users' posts, known as tweets, that data went through a natural language processing in order to remove stopwords and unwanted characters. The Naive Bayes machine learning algorithm was used to classify sentiment in posts. The tool used presented, despite being used a simple test base, a satisfactory accuracy of 86.85% in the test phase. Subsequently, the tool was used in retrieved tweets and also proved to be quite usual, managing to predict all retrieved tweets.

Keywords: Text mining; Opinion mining; machine learning; natural language processing; Twitter; Social Networks.

LISTA DE FIGURAS

LISTA DE SIGLAS

API - Interfaces de Programação de Aplicativo

PLN - Processamento de Linguagem Natural

NLTK - Natural Language Toolkit

SVM - Support Vector Machine

KNN - K- Nearest Neighbor

CGU – Controladoria Geral da União

TSV - Tab-Separeted Value

BOW - Bag-of-Words

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO

1.1. Contexto

As redes sociais conectam pessoas e também propiciam o compartilhamento de informações entre os seus usuários, independentemente do caráter, seja pessoal, profissional ou comercial. Com a popularização das tecnologias de informação e comunicação, as redes sociais têm adquirido cada vez mais usuários. No entanto, esse crescimento no número de usuários tem acarretado um enorme fluxo de dados.

Redes sociais têm proporcionado que pessoas, grupos e organizações possam se conectar a partir de interesses ou valores em comum, interagindo entre si. Essas ferramentas de comunicação, como o Twitter, permitem que milhões de usuários troquem e compartilhem informações todos os dias com sua opinião a respeito de pessoas, produtos ou serviços em geral. À medida que esse tipo de conteúdo é publicado, consumidores e empresas têm à sua disposição uma grande quantidade de opiniões, que influenciam de alguma forma as decisões dos consumidores em relação às suas compras ou das empresas em relação ao lançamento de um novo produto ou serviço (CARVALHO, 2014).

Para Benevenuto *et al.* (2015), as redes sociais representam uma revolução digital. Elas permitem a expressão e difusão das emoções e opiniões através da rede, uma vez que são locais onde as pessoas discutem sobre tudo expressando opiniões políticas, religiosas ou mesmo sobre marcas, produtos e serviços.

Para Araújo *et al.* (2013), as redes sociais têm se tornado uma importante plataforma de comunicação onde são reunidos os mais variados tipos de informação, dentre as quais destacam-se as opiniões e sentimentos expressos pelos usuários em conversas simples ou por meio de mensagens. A quantidade de usuários nessas redes cresce consideravelmente todos os dias, gerando um grande volume de dados. O autor destaca que vários estudos foram realizados nas redes sociais para a identificação e monitoramento de polaridade em mensagens compartilhadas, pois uma parcela das mensagens postadas estaria relacionada ao humor e as emoções expressas pelos usuários. Apesar das inúmeras aplicações que podem ser feitas com a análise de polaridade em mensagens, o grande fluxo de dados torna essa tarefa muito complicada.

As opiniões contidas nas redes sociais são carregadas de emoções e com uma carga de sentimentos por trás de cada palavra escrita. No entanto, a enorme circulação de opiniões por meio das redes sociais fazem com que as informações geradas sobre um determinado assunto se tornem difíceis de se identificar manualmente em razão do grande fluxo de dados e da ambiguidade existente, o que pode tornar a tarefa impraticável (SANTOS, 2010).

Portanto, com a crescente popularização das redes sociais juntamente com a internet, surgiu essa necessidade de explorar os dados gerados a fim de extrair conhecimento de forma automática dada a importância que essas mensagens possuem.

1.2. Justificativa.

Redes sociais têm se tornado ferramentas de comunicação muito popular entre os usuários da internet. Diariamente, milhões de usuários trocam/compartilham informações contendo opiniões a respeito de política, religião, marcas de produtos e serviços. Diante desse cenário acelerado de compartilhamento de informações, surge a necessidade de identificar as opiniões por trás desses textos, o que se torna uma tarefa impraticável para se realizar manualmente, devido a grande quantidade de textos publicados. Portanto, é necessário a criação de processos automáticos para realizar a análise das informações contidas nesses textos.

1.3. Objetivos.

1.3.1. Objetivo Geral

Este trabalho tem por objetivo o estudo e o desenvolvimento de uma ferramenta capaz de analisar o sentimento contido em dados disponíveis no Twitter, tendo como fonte de dados os tweets disponíveis pela rede social Twitter através da API (Interfaces de Programação de Aplicativo) REST desta rede social.

1.3.2. Objetivos Específicos.

Abordar os campos mais relevantes da mineração de dados e análise de sentimentos e implementar uma ferramenta para realizar análise de sentimentos sobre dados coletados do Twitter a respeito de determinado tema, utilizando para isso aprendizado de máquina e conceitos de processamento de linguagem natural (PLN).

1.4. Metodologia

Para a elaboração deste trabalho, foi realizada uma abordagem bibliográfica a qual nos auxiliou no capítulo da fundamentação teórica. Foram utilizados artigos de periódicos, teses e dissertações para se obter conceitos relacionados à análise de sentimentos, aos métodos de classificação de aprendizado de máquina, ao processamento de linguagem natural, bem como aos trabalhos já desenvolvidos na mesma linha de estudo.

Esta pesquisa também se enquadra como quantitativa na qual pretende-se realizar um estudo de caso, utilizando uma ferramenta desenvolvida na linguagem de programação Python juntamente com a rede social Twitter para a coleta de dados. As bibliotecas Python foram utilizadas na busca, leitura, pré-processamento e análise de sentimento, bem como a API REST disponibilizada pelo Twitter para ter acesso aos dados disponíveis para desenvolvedores.

CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresentará algumas definições que são importantes para o entendimento dos processos que serão realizados e apresentados ao longo deste trabalho.

O processamento de linguagem natural (PLN) tem por objetivo capacitar o computador/máquina a se comunicar, mesmo que não seja de maneira direta, por meio da linguagem humana. O PLN está diretamente ligado à inteligência artificial, tendo como um de seus pressupostos a facilitação na compreensão de textos expressos em linguagem humana natural para as máquinas, auxiliando-as na extração de sentido semântico dos dados em linguagem natural que lhes foram apresentados[1].

Quando se fala em PLN não podemos deixar de mencionar a biblioteca NLTK[2] (Natural Language Toolkit), que é uma biblioteca que foi desenvolvida na linguagem de programação Python, tendo como objetivo a implementação de softwares que buscam trabalhar com a linguagem humana, servindo como suporte no estudo do processamento de linguagem natural (WEIAND, 2018).

Segundo Becker e Tumitan (2013, p.02), a análise de sentimentos é descrita como “qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual.” A análise de sentimentos está diretamente relacionada à capacidade de identificação, extração e polarização do sentimento que está expresso em textos, emoticons, vídeos ou imagem, sentimento este que pode ser negativo, positivo ou neutro. Esta análise é feita graças ao processamento de linguagem natural que foi citado anteriormente e algoritmos que utilizam o aprendizado de máquina como um de seus pressupostos.

[1] Processamento de Linguagem Natural, o que é e sua importância, **SAS**. Disponível em:< https://www.sas.com/pt_br/insights/analytics/processamento-de-linguagem-natural.html>. Acesso em 04 de outubro de 2020.
[2] https://www.nltk.org/

2.1. Processamento de Linguagem Natural

Segundo Vieira e Lopes (2010, p. 185), “o Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais”. Podemos nos fazer valia desses processos de várias formas, como, por exemplo, na análise de aceitação de algum produto de determinada empresa, nas possíveis intenções de votos numa eleição, na visualização dos feedbacks de clientes sobre determinada empresa, dentre outras formas.

Podemos demonstrar alguns objetivos usuais em PLN que Liddy (2003) destaca, são eles: “a recuperação de informação a partir de textos, tradução automática, interpretação de textos e realização de inferências a partir de textos” (*apud* VIEIRA; LOPES 2010, p. 185).

De acordo com Gonzalez e Lima (2003), “o PLN trata computacionalmente os diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos”. Gonzalez e Lima (2003) ainda discorrem sobre o PLN numa forma mais ampla, defendendo que este objetiva fazer com que uma máquina possa se comunicar por meio da linguagem humana, ainda que não seja por meio de todos os níveis de entendimento, basta que haja compreensão.

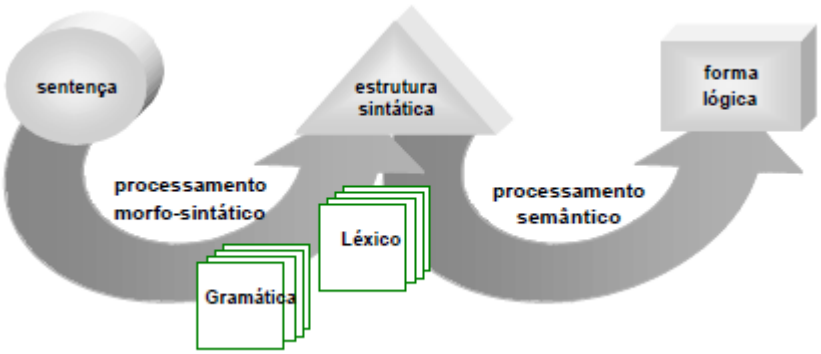
Os níveis são mostrados a seguir de maneira individualizada e com seus respectivos conceitos (GONZALEZ; LIMA, 2003, p.3):

> - fonético e fonológico: do relacionamento das palavras com os sons que produzem;
> - morfológico: da construção das palavras a partir unidades de significado primitivas e de como classificá-las em categorias morfológicas;
> - sintático: do relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças;
> - semântico: do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças; e
> - pragmático: do uso de frases e sentenças em diferentes contextos, afetando o significado.

Podemos observar que Gonzalez e Lima (2003) buscam representar as sentenças de maneira gramatical, eles citam que o significado de uma sentença é obtida por meio da sua forma lógica. Na parte lógica é codificado os sentidos que as palavras podem transmitir , assim como os relacionamentos semânticos entre as

orações são identificados (ALLEN, 1995; FRANCONI, 2001, *apud* GONZALEZ; LIMA, 2003, p.3), como é mostrado na Figura 1.

Figura 1. Transformações da sentença na estrutura sintática e na forma lógica.



Fonte: Gonzalez e Lima (2003, p. 3)

A estruturação de uma frase ou oração pode ser auferida quando conseguimos fazer o seu processamento morfossintático, observando, é claro, as regras gramaticais. Já as categorias da parte morfológica são explicitadas em léxicos, parte esta que é necessária para a etapa do processamento morfossintático Gonzalez e Lima (2003).

Graças ao mapeamento semântico podemos estruturar uma frase na forma sintática para sua forma lógica. A gramática e os léxicos são extremamente importantes para se transformar uma estrutura sentencial numa estrutura lógica, visto que, estes nos dão os significados dos itens lexicais e aqueles nos orientam quanto às classificações e estruturações que as frases/sentenças podem exercer num contexto, como bem observam Gonzalez e Lima (2003).

A biblioteca NLTK foi desenvolvida na linguagem de programação Python e tem como objetivo a implementação de softwares que buscam trabalhar com a linguagem humana. Possui ainda vários léxicos, bibliotecas de processamento de textos, tokenização, marcação e raciocínio semântico, além de disponibilizar recursos por meio de sua interface gráfica simples (NLTK, c2020). A NLTK nos ajuda no desenvolvimento e implementação do algoritmo de Naive Bayes[3], assim como na

[3] “Modelo probabilístico simples com base na regra de Bayes com seleção de recursos independentes.” (WEIAND, 2018, p.9),

utilização de outros recursos, por exemplo, stemming e treino de palavras (WEIAND, 2018, p. 11).

2.2. Análise de Sentimento

A mineração de opiniões ou análise de sentimentos é o estudo por meio computacional das opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, que foram expressos em forma de texto. Para se ter a análise de sentimentos, é preciso verificar as opiniões de um texto a partir de suas partes, independente de tamanho e formato, como páginas web, posts, comentários, tweets, revisões de produtos, dentre outros (BECKER; TUMITAN, 2013).

Tendo como base o que Becker e Tumitan (2013) explicitaram, podemos afirmar que toda opinião é composta de pelo menos dois elementos chave: um alvo e um sentimento sobre este alvo. Um alvo pode ser uma entidade, aspecto de uma entidade, ou tópico representando produto, pessoa, organização, marca ou evento. O sentimento é a opinião demonstrada a partir da emoção que o autor tem a respeito do alvo.

Ao longo de seu estudo, Becker e Tumitan (2013) mostraram uma forma de facilitar a análise de sentimento, que se dá através da divisão da análise em diferentes granularidades. Destacamos que a decisão do nível está atrelada ao conteúdo que a aplicação demonstra. Segundo Becker e Tumitan (2013), os níveis são:

> - Documento: o objetivo é classificar um documento como um todo, para sabermos se este reproduz um sentimento positivo ou negativo, é mais eficaz quando o documento trata de apenas uma única entidade(alvo);
> - Sentença: neste nível, o objetivo é mais específico, analisa-se uma sentença isolada de um documento, bastante útil quando o documento trata de mais de uma entidade(alvo). Além de permitir a identificar a distinção entre sentenças objetivas(fatos) e subjetivas(opiniões);
> - Entidade e Aspecto: o foco deste nível é a opinião expressa, podendo haver várias opiniões sobre diferentes alvos em diferentes sentenças, documentos e orações. Exemplo: “Adoro esse bolo porque seu recheio é muito saboroso. Infelizmente seu preço é bastante alto”, observamos três opiniões sobre o bolo e sobre dois aspectos do bolo (recheio e preço). Este nível se torna mais complexa a análise, por conta desta variedade de opiniões sobre diferentes aspectos de determinada opinião.

Para se detectar um sentimento presente num texto, é muito comum acreditarmos que é necessário o uso de palavras sentimentais, que expressem opinião, como, por exemplo, os adjetivos, entretanto o uso destes tipos de palavras não é condição necessária, nem suficiente. Tendo em vista que palavras de opiniões

podem ter diferentes sentidos a depender do contexto em que foram utilizadas, destaca Becker e Tumitan (2013).

Becker e Tumitan (2013) destacam ainda que existem alguns problemas que podem atrapalhar de certa forma a nossa análise de sentimento, como a ironia e o sarcasmo. Como veremos na Figura 2.

Figura 2. Exemplo de Ironia e Sarcasmo.

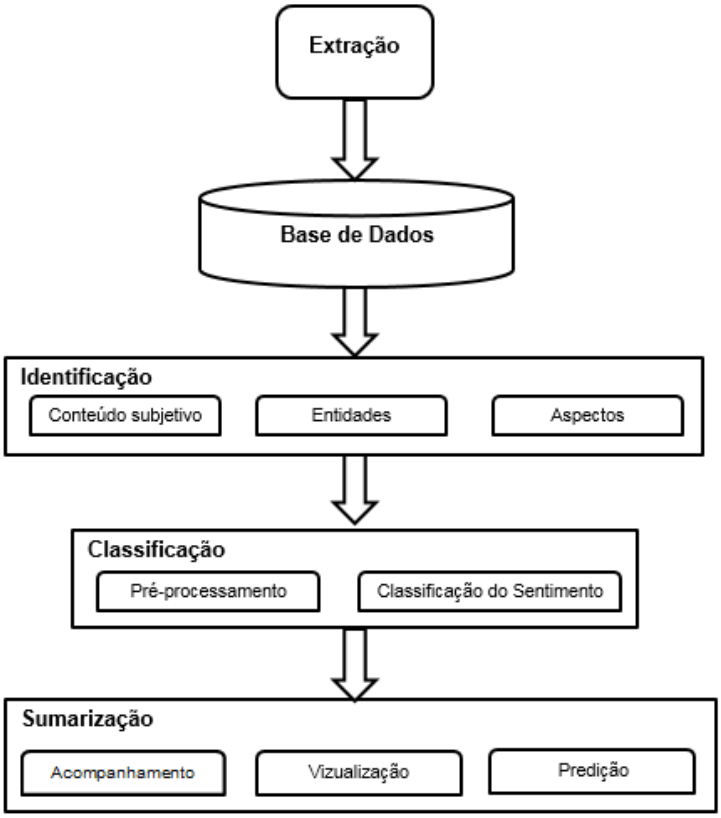
Fonte: Adaptada de Toda Matéria (2020).[4]

Podemos também citar os casos dos diferentes significados que uma palavra pode assumir. Como por exemplo, seguem as orações “Você é uma pessoa caríssima” e “Esta jaqueta é caríssima”. Observamos que na primeira oração o adjetivo caríssima tem um sentido positivo e na segunda tem um sentido negativo. Ademais, nem todo sentimento é expressado através de palavras de sentimento, existe a negação que pode mudar completamente o sentido de qualquer palavra ou adjetivo o/a qual ela esteja modificando.

Retomando à análise de sentimento em si, podemos dizer que a mesma pode ser feita por meio de três tipos de tarefas genéricas: **identificar** as opiniões expressas sobre determinado assunto ou alvo em um conjunto de documentos, **classificar** a orientação ou polaridade desta opinião, isto é, se tende a positiva ou negativa e **apresentar** os resultados de forma agregada e sumarizada conforme menciona Becker e Tumitan (2013), conforme a Figura 3 nos mostra.

[4] DIANA, Daniela. Diferença entre Sarcasmo e Ironia. **Toda Matéria**. Disponível em:<https://www.todamateria.com.br/sarcasmo-e-ironia/>. Acesso em 25 de Novembro de 2020.

Figura 3. Etapas da Análise de Sentimentos.



Fonte: Adaptado de Becker e Tumitan (2013, p. 6)

Existem quatro grandes grupos de abordagens de classificação de polaridade, que são divididas em **léxicas**, **estatísticas**, **semânticas** e **aprendizado de máquina** e que serão melhor detalhadas na próxima seção. É importante frisar que essas técnicas podem ser combinadas para se obter um resultado melhor, porém, destacam-se as abordagens léxicas e aprendizagem de máquina, na questão de predominância. Não havendo qualquer relação com o desempenho das demais técnicas, ou seja, nenhuma técnica se sobressai a outra em questão de desempenho (BECKER e TUMITAN, 2013).

2.3. Aprendizagem de Máquina

Segundo Castro e Ferrari (2016, p.14), “aprendizado de máquina é a área que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio de experiência” (*apud* SANTOS, 2017, p. 26). De acordo com Michalski, Carbonell, Mitchel (1983), ”aprendizagem de máquina é uma subárea

de pesquisa em inteligência artificial que estuda métodos computacionais para adquirir comportamento inteligente e simular o aprendizado humano a partir de computadores por meio de indução", (*apud* SOUSA, 2012, p.31). Bishop (2003) destaca que "Aprendizado de máquina é uma área de Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo de aprendizado" (*apud* SOUSA, 2012, p.31).

Com o objetivo de classificar uma opinião quanto à sua polaridade, muitos trabalhos utilizam técnicas baseadas em algoritmos tradicionais de aprendizado de máquina (CARVALHO, 2014). O aprendizado de máquina pode ser considerado um dos fundamentos da tecnologia da informação e, portanto, de fundamental importância para os dias atuais devido ao grande fluxo de dados gerados todos os dias, onde suas técnicas podem ser aplicadas a fim de realizar o reconhecimento de padrões existentes neles.

Em geral, o aprendizado de máquina é subdividido em supervisionado, não supervisionado e por reforço. A seguir é apresentado os conceitos relacionados a cada um desses métodos, dando ênfase ao aprendizado supervisionado que foi utilizado neste trabalho:

Segundo Trevisan (2015), o aprendizado supervisionado é feito a partir de um grupo de dados já rotulados. Esses dados são manipulados pela figura de um supervisor, que faz com que o sistema já conheça as entradas e suas respectivas saídas e seja capaz de fazer o mesmo com rótulos desconhecidos. No aprendizado não supervisionado as entradas são conhecidas, mas as saídas não. Por isso, a máquina busca identificar grupos ou padrões para só depois rotular os dados. O terceiro método de aprendizado de máquina é chamado de aprendizado por reforço, e nele a máquina tenta aprender qual a melhor ação a ser tomada para uma sequência de decisões.

No aprendizado supervisionado tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: entrada/saída desejada. O algoritmo de aprendizado de máquina extrai a representação do conhecimento a partir desses exemplos. O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente. No aprendizado não-supervisionado não há a presença de um professor, ou seja, não existem exemplos rotulados. O algoritmo de

aprendizado de máquina aprende a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade. Essas técnicas são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados (LORENA e CARVALHO, 2007).

Para Lorena e Carvalho (2007), um requisito importante para as técnicas de aprendizado de máquina é que elas sejam capazes de lidar com dados imperfeitos, denominados ruídos. Esses tipos de dados são encontrados em textos a serem analisados, pois é comum a presença de dados com rótulos e/ou atributos incorretos.

O tipo de aprendizado utilizado neste trabalho é baseado em aprendizado supervisionado e tem como finalidade determinar a polaridade das opiniões contidas em avaliações de mensagens do Twitter, utilizando o algoritmo de aprendizado de máquina Naive Bayes.

Para Santos (2017), o *Naives Bayes* é utilizado para a modelagem de previsões exploratórias, sendo uma técnica para construir classificadores. Esses classificadores atribuem rótulos de classes para as instâncias de algum problema, e esses rótulos são extraídos de algum conjunto de dados. Amaral (2016, p.41) complementa que:

> [...] é um algoritmo *bayesiano,* baseado na teoria das probabilidades e que supõe que os atributos vão influenciar a classe de forma independente. Na criação do modelo, este classificador vai construir uma tabela mostrando o quanto cada categoria de cada atributos contribui para cada classe. Uma vez montado o modelo, ao submetermos uma nova instância para o classificador, ele vai olhar os pesos nesta tabela, somá-los e ver qual classe teve um peso maior, que sairá como “vitorioso”.

O algoritmo é baseado em torno da regra de Bayes, uma maneira de olhar para as probabilidades condicionais que permite que você alterne em torno da condição de uma forma conveniente. A condicional é um evento que provavelmente irá ocorrer X, dada a evidência Y. Isso é normalmente escrito P (X | Y). A regra de Bayes nos permite determinar essa probabilidade, quando tudo o que temos é a probabilidade de o resultado oposto e dos dois componentes individualmente (WEIAND, 2018):

Figura 4. Fórmula do Teorema de Bayes.

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

Fonte: Weiand (2018, p. 10).

- P(X/Y): probabilidade de X acontecer dado que Y ocorreu
- P(Y/X): probabilidade de Y acontecer dado que X ocorreu
- P(X): probabilidade de X ocorrer
- P(Y): probabilidade de Y ocorrer

Portanto, o algoritmo Naive Bayes funciona como um classificador que utiliza a probabilidade para determinar a ocorrência de um evento.

As SVMs (Support Vector Machine) surgiram no final dos anos 70. Segundo Weston *et al.*(1998), “desde então têm sido cada vez mais utilizadas para resolver problemas de diversas naturezas como reconhecimento de dígitos manuscritos, reconhecimento de sons e imagens, estimativa de densidade, classificação de textos, dentre outras aplicações” (*apud* SOUSA, 2012, p.32). Elas são máquinas de vetores de suporte com uma entrada/saída, onde um usuário é capaz de inserir uma entrada, e de acordo com o modelo de treinamento desenvolvido, ela devolverá uma saída. Esse modelo é muito utilizado em diversas tarefas de reconhecimento de padrões.

CAPÍTULO 3 - TRABALHOS RELACIONADOS

A partir do levantamento feito através das pesquisas realizadas na área de mineração de opinião em textos, abrangendo problemas semelhantes a esse estudo, foram encontrados alguns trabalhos relacionados. Os principais estão listados abaixo.

3.1. Mineração de textos no Twitter.

O trabalho proposto por Trevisan (2015) descreve a criação de uma ferramenta para análise de opiniões, utilizando para isso dados coletados da rede social Twitter. Para a coleta de tweets e consequente criação de um dataset utilizou-se a API Search do Twitter e o algoritmo não supervisionado k-means para o agrupamento de dados. Posteriormente, são oferecidas duas opções de classificadores de sentimentos, a saber SVM (*Support Vector Machine*) e Naive Bayes. No entanto, em uma primeira etapa, com o uso do SVM foram obtidos os melhores resultados, mas logo em seguida, quando melhorada a medida testada anteriormente o SVM apresentou resultados piores, enquanto que o Naive Bayes mostrou-se sem nenhuma alteração mesmo na utilização de diferentes ponderações.

Portanto, o artigo mostra o desenvolvimento de uma ferramenta para análise de um evento real, passando por várias etapas de processamento de dados, desde sua coleta até sua análise. A utilização de uma ferramenta como essa se torna muito útil para o acompanhamento de um evento que está ocorrendo, embora sua criação requer um dataset extenso para o treinamento dos classificadores de sentimentos utilizados e consequentemente para resultados mais precisos.

3.2. Análise de opiniões utilizando técnicas de mineração de dados em redes sociais. Estudo de caso: Twitter.

O trabalho proposto por Santos (2017) teve como objetivo analisar as técnicas de mineração de dados, utilizando para isso a rede social Twitter para realizar a classificação de opiniões, pois esta mostrou uma melhor disponibilidade de dados. A autora analisou o sentimento da população brasileira sobre a absolvição do presidente Michel Temer, onde a coleta dos dados foi baseada no seguinte acontecimento: a Câmara dos Deputados rejeita a denúncia da Procuradoria Geral da República, por

crime de corrupção passiva contra o presidente Michel Temer e o livra de responder ao processo no Supremo Tribunal Federal (SENADO NOTÍCIAS, 2017), criando desse modo um dataset com essas informações. Nessa pesquisa foram utilizados três tipos de algoritmos SVM (Support Vector Machine), KNN (K- Nearest Neighbor) e Naive Bayes a fim de verificar qual deles apresentaria melhores resultados. Os três tipos de algoritmos, após serem testados, demonstraram como resultado uma carga de sentimento negativo em relação ao acontecimento. No SVM, 92% dos tweets foram classificados como negativos; o KNN apresentou uma taxa de 86% das opiniões como negativas e por fim o Naive Bayes, que apresentou um resultado de 88% das opiniões classificadas em negativas. Embora o Naive Bayes tenha mostrado resultado inferior quando comparados aos demais, ainda assim os seus resultados foram bem próximos entre si, o que os colocaria em um mesmo patamar.

3.3. Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na CGU.

Andrade (2015) aborda, nesse estudo, algoritmos de classificação utilizados atualmente para fazer a análise de documentos textuais provenientes do setor de triagem de denúncias na CGU (Controladoria Geral da União) e com isso gerar uma execução automática de classificação desses documentos. Os dados textuais utilizados foram extraídos do banco de dados SQL Server e usados com algoritmos já conhecidos para classificação de textos (Random Forest, Decision Tree, Naive Bayes e SVM), a fim de verificar qual deles apresenta melhor desempenho. Como nenhum deles apresentou resultado satisfatório, optou-se pelo classificador CAH + MDL baseado na Árvore de Huffman, que apresentou uma taxa de acerto de 84%.

3.4. Mineração de textos do twitter utilizando técnicas de classificação

O trabalho proposto por Leite (2015) utilizou técnicas do algoritmo de classificação Naive Bayes para gerar um modelo de classificação de tweets, sendo o principal objetivo deste a distribuição de tweets por categorias. Primeiramente foi criado um script em Python para comunicar-se com a API do Twitter e fazer a coleta dos tweets, e seu posterior armazenamento em um arquivo TSV (*Tab-Separeted Value*). Após essa primeira etapa é realizada uma limpeza desses dados para a

remoção de termos irrelevantes e para uma melhor organização dos mesmos, utilizando também a plataforma NLTK (Natural Language Toolkit), que trabalha com ferramentas para PLN. O autor dividiu entre categorias pré-definidas – esporte, economia, religião, política e outros – os tweets utilizados para criar o conjunto de treino, obtendo como resultado uma taxa de 83.6761% de precisão. O resultado mostra que, a partir de categorias pré-definidas pelo pesquisador, é possível classificar novos tweets, com um nível de precisão aceitável utilizando o algoritmo de classificação Naive Bayes.

3.5. Tweetmining: análise de opinião contida em textos extraídos do Twitter

O trabalho proposto por Sousa (2012) busca a implementação de uma aplicação para classificar automaticamente o sentimento contido em tweets. Os tweets são classificados como positivos, negativos ou neutros, baseado no que os usuários do Twitter expressam seus sentimentos sobre a cidade de Campinas. Para a criação do dataset foram utilizadas APIs disponibilizadas pelo Twitter e os resultados armazenados em arquivos de textos no formato JSON. Foi utilizado o algoritmo de treinamento SVM, que alcançou uma taxa de 84.2903% de acerto. A fim de obter resultados satisfatórios é necessário ter uma boa base de dados, visto que os tweets apresentam problemas como ambiguidades em seus textos, má formação de frases (o que causa ruídos), abreviaturas, entre outros que dificultam sua classificação.

3.6. Twitter, análise de sentimento e desenvolvimento de produtos: Quanto os usuários estão expressando suas opiniões?

No trabalho proposto por Santos *et a*l. (2010), os autores procuram investigar a rede social Twitter como uma ferramenta capaz de interligar clientes com a empresa. Utilizando o método aprendizado de máquina supervisionado SVM, treinou-se o algoritmo a fim de classificar as mensagens como opinativas ou neutras em relação ao produto em estudo, nesse caso o Windows 7. A coleta dos tweets foi feita por meio da Twitter Streaming API, e seguidamente optou-se por filtrar as mensagens buscando apenas os textos em inglês, realizando-se uma classificação manual. Um teste realizado teve como resultado uma taxa de acerto de 80%, tornando o classificador apto para realizar a classificação das outras mensagens

automaticamente. No trabalho, os autores recomendam uma análise de sentimento no período de lançamento de um certo produto, visto que as pessoas tendem a comentar sobre assuntos atuais no Twitter e não sobre algo que já estão usando há algum tempo.

3.7. Análise de sentimento de insegurança através do Twitter

O trabalho proposto por Nascimento e Reis (2018) tem como principal objetivo identificar, por meio de mensagens postadas no Twitter, o sentimento de insegurança gerado no cidadão brasileiro. Os pesquisadores utilizaram a API REST do Twitter para a coleta dos dados, cerca de 400 tweets que geraram um dataset com 1821 palavras e que foram utilizadas para treinamento dos algoritmos de classificação. Foram utilizados os algoritmos de aprendizado de máquina Floresta Aleatória, SVM e Regressão Logística, que tiveram em média uma taxa de acerto de 66,5%, sendo apresentadas as matrizes de confusão que explicitam esses resultados.

3.8. Análise de sentimentos do Twitter com Naive Bayes e NLTK

O trabalho proposto por Weiand (2018) tem como objetivo a construção de um algoritmo capaz de realizar uma análise do sentimento dos usuários do Twitter, utilizando para isso um dataset de tweets coletados com o auxílio da classe desenvolvida por Sanders (2011), que contém um arquivo em formato “csv” com tweets que possuem dois tipos de classificação de sentimentos diferentes. Esse arquivo conta ainda com uma classe em Python, que possibilita a obtenção dos textos destes tweets diretamente do Twitter, porém o autor não disponibiliza os mesmo já categorizados, pois há direitos autorais. Os tweets opinativos são classificados em opiniões positivas e negativas. Foi utilizado o algoritmo de classificação Naive Bayes em conjunto com as bibliotecas NLTK e ScikitLearn. Os dados foram separados para o desenvolvimento do treino, validação cruzada e teste. O dataset foi separado aleatoriamente em três grupos: um com 80% dos dados, os quais foram utilizados para o treinamento do algoritmo; 10% dos dados para a validação cruzada e os 10% restantes para o teste, obtendo como resultado um número de acerto alto 91%.

Pode-se observar em todos os estudos realizados nos textos que compõem esse capítulo, que existem grandes dificuldades para os processos de análise de sentimento em textos e a consequente extração de opinião.

Este trabalho de pesquisa baseou-se, se não em sua totalidade, mas em boa parte, na elaboração do trabalho proposto por Weiand (2018). As semelhanças vão desde o uso de um conjunto de dados com tweets já classificados quanto ao sentimento, da escolha do algoritmo para classificação Naive Bayes, bem como das bibliotecas que foram utilizadas (NLTK e SciKItLearn) até mesmo a maneira escolhida para a divisão dos dados utilizados para o treino do algoritmo.

CAPÍTULO 4 - ESTUDO DE CASO

No estudo de caso, será apresentada uma ferramenta elaborada em python com auxílio das bibliotecas Twitter Search[5], Pandas[6], SKlearn[7] e NLTK, em que fará a busca por tweets de determinados assuntos através da API do Twitter e em seguida aplicará a análise de sentimento nos tweets recuperados, utilizando os métodos da mineração de dados. Para ser feito o treino do algoritmo utilizou-se o algoritmo Multinomial de Naive Bayes, para posteriormente realizar-se a análise de sentimentos nos tweets recuperados com o modelo treinado. Após o treino do modelo, serão feitos os testes com os tweets que serão recuperados.

4.1. Twitter e API do Twitter

O Twitter é uma rede social bastante conhecida no mundo, usada por milhares de pessoas. Ele possibilita uma interação com os outros usuários desta rede, assim como com outras organizações. É possível interagir, nesta rede social, por meio de “tweets”, que são pequenas publicações que contêm uma quantidade máxima de 280 caracteres, após uma mudança em 2017.

As APIs do Twitter[8] dão acesso a partes dos serviços disponibilizados pela rede social, para que desenvolvedores possam criar softwares que tenham como pressuposto a interação com o Twitter. Neste trabalho, utilizou-se a API REST que permite a autenticação através de mecanismos baseados no padrão OAuth (Open Authentication), um padrão aberto para autorizar aplicações a acessar dados em nome do usuário. O próprio site menciona que este acesso possibilita uma ajuda para as empresas no que diz respeito a medição de opiniões dos clientes sobre ela, ou seja, pode auxiliar no feedback do cliente a respeito de marcas, empresas, tendências, dentre outras coisas.

[5] https://twittersearch.readthedocs.io/en/latest/
[6] https://pandas.pydata.org/docs/
[7] https://scikit-learn.org/stable/
[8] Sobre as APIs do Twitter ,**Help Twitter**. Disponível em: <https://help.twitter.com/pt/rules-and-policies/twitter-api>. Acesso em 06 de set. de 2020.

4.2. Tendo acesso à API do Twitter

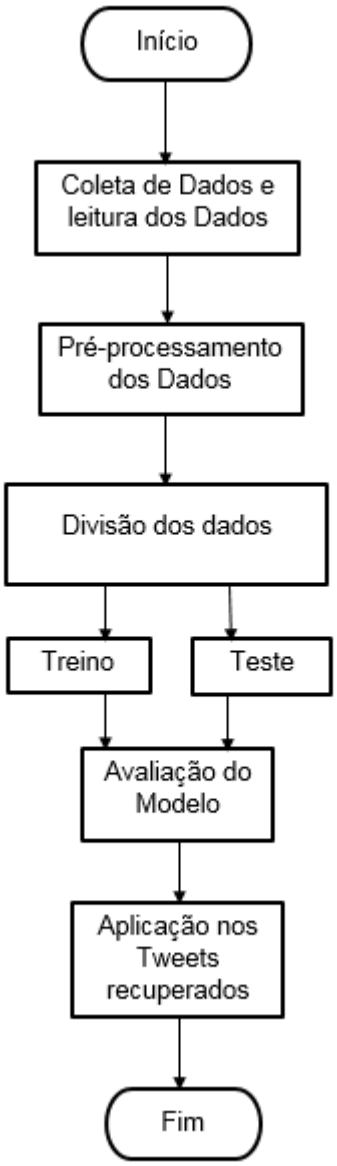
Inicialmente é preciso ter um registro no site https://apps.twitter.com/, pois isso permitirá ao Twitter identificar a aplicação. Após ser feito o login no site e o devido preenchimento dos campos solicitados, têm-se acesso às chaves de segurança (Consumer Key, Consumer Secret, Access Token e Access Token Secret) para a utilização da API REST.

Após a finalização do processo de criação da aplicação e liberação das chaves, o desenvolvedor torna-se habilitado a fazer o uso da API REST, sendo assim possível acessar os dados que o Twitter disponibiliza para os desenvolvedores.

4.3. Desenvolvimento da ferramenta

Na Figura 5, as fases de desenvolvimento da ferramenta podem ser visualizadas. Na fase de coleta de dados, os dados que estão disponibilizados no Dataset são coletados. No pré-processamento, a limpeza dos mesmos é feita. Na separação dos dados de teste e treino, indica-se ao algoritmo, que os dados do dataset são os dados de treino e a lista que foi criada com 10 sentenças serão os dados de teste. Durante a avaliação do modelo, é realizado o teste e a verificação da acurácia, para que se tenha uma visão geral da usabilidade da ferramenta. Após a realização das fases descritas acima, aplica-se a ferramenta nos tweets recuperados.

Figura 5. Fluxograma do Algoritmo.



Fonte: Elaborado pelos Autores.

4.3.1 Importação das bibliotecas

Inicialmente, as bibliotecas Twitter Search, NLTK, Pandas e Sklearn, que serão utilizadas no algoritmo, são importadas. A biblioteca Twitter Search é responsável pela busca na API do Twitter do tema pesquisado, a NLTK auxilia no processamento de linguagem natural, a Pandas ajuda na análise dos dados que serão exibidos por meio de gráficos, e a Sklearn fará com que o algoritmo consiga implementar o aprendizado de máquina através do Multinomial de Naive Bayes.

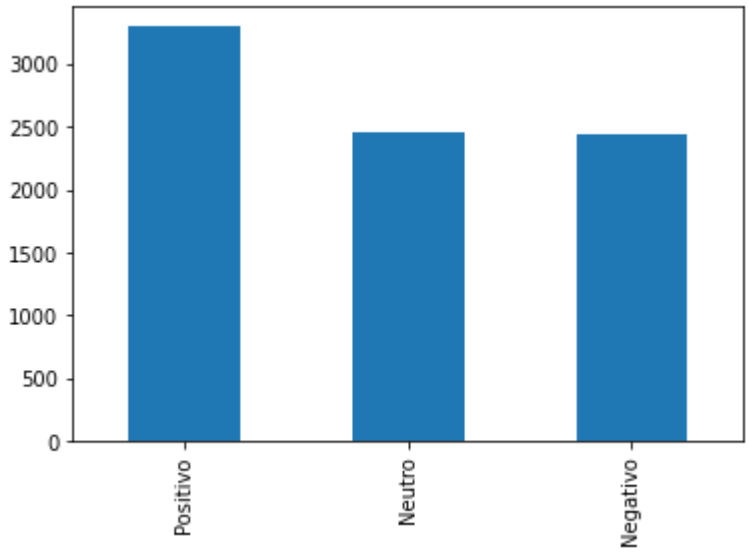
4.3.2. Coleta e leitura dos dados

Utilizou-se, como base de treino do algoritmo, o dataset elaborado pelo governo de Minas Gerais, ao qual pôde ter seus dados carregados e acessados após a importação dos módulos, com o auxílio da biblioteca Pandas. A escolha deste dataset se deve ao fato de o mesmo encontrar-se com seus tweets rotulados quanto à sua polaridade. Para um melhor estudo, serão utilizados como dados, apenas os tweets e seus respectivos sentimentos, assim como a contabilização quanto às suas polaridades classificadas como positivas, neutras e negativas, ou seja, a quantidade de tweets de acordo com sua classificação.

Nos trabalhos pesquisados no capítulo da fundamentação teórica, viu-se que a utilização de um dataset supervisionado teria uma grande importância, visto que o mesmo já entregaria os tweets com suas respectivas polaridades/sentimentos, facilitando no quesito aprendizagem de máquina.

Na Figura 6 é ilustrada a distribuição das polaridades, contendo 3300 tweets positivos, 2453 neutros e 2446 negativos, porém ainda sem o pré processamento.

Figura 6. Quantidade de tweets e polaridades do Dataset sem o pré-processamento.



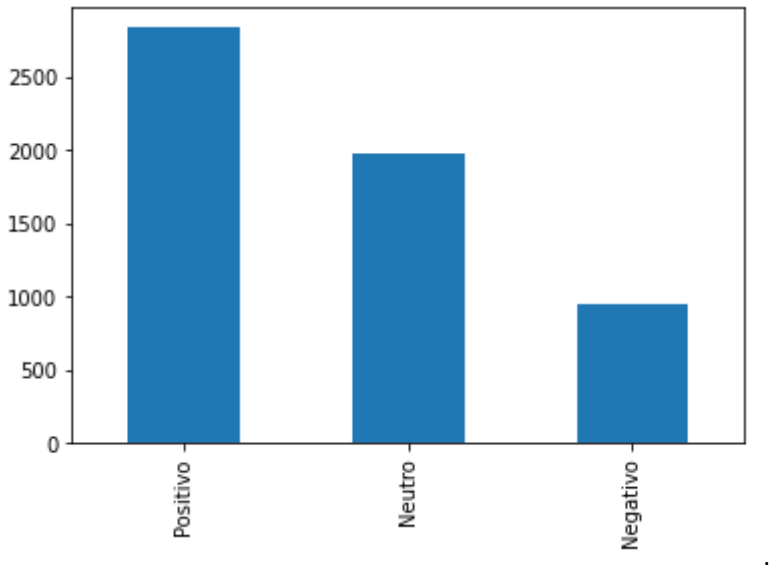
Fonte: Elaborada pelos Autores.

4.3.3. Pré-Processamento dos dados

Para se ter uma análise de sentimento mais confiável é preciso fazer o tratamento dos dados. Por esta razão, os tweets e seus respectivos sentimentos foram classificados, retirando as stopwords[9], caracteres indesejados, assim como foram removidas as linhas duplicadas do Dataset.

Quando realizado o pré-processamento, foi possível observar que o Dataset diminuiu de tamanho, a quantidade de dados rotulados como positivos passou de 3300 para 2840, de neutros 2453 para 1974 e de 2446 negativos para 951. Tudo isso graças à remoção de linhas duplicadas. A Figura 7 mostra a nova distribuição de tweets e suas respectivas polaridades após a etapa de pré-processamento.

Figura 7. Quantidade de tweets e polaridades do Dataset após o pré-processamento.



Fonte: Elaborada pelos Autores.

4.3.4. Divisão dos Dados de Treino e dados de Teste

Os tweets disponíveis no dataset de Minas Gerais já estão rotulados com suas polarizações positivas, negativas ou neutras, justificando assim o seu uso como dados de treino. Após instanciar o objeto que faz a vetorização dos dados é feita a aplicação nos dados de texto do dataset por meio do vetorizador CountVectorizer. É

[9] Palavras que não agregam para o sentido da frase, como por exemplo: artigos, pronomes, preposições, etc.

imprescindível destacar que, quando se está trabalhando com texto, faz-se necessário transformar as palavras em dados binários, a fim de que o algoritmo de aprendizado de máquina possa compreendê-las, já que o mesmo não consegue trabalhar diretamente com as palavras. A solução para este problema é utilizar a metodologia BOW (Bag-of-Words), criando uma matriz de presença, conforme Santana (2020), por meio do vetorizador citado anteriormente. Assim, terminado o processo de vetorização, é necessário transformá-los em dados binários, para que o computador possa compreender. Para isso utilizou-se o método vectorizer.fit_transform(tweets), passando os tweets que são dados na forma de texto como parâmetro.

Como dados de teste utilizou-se uma lista com frases relacionadas ao dataset. Serão utilizadas 10 frases para verificar se o modelo está conseguindo fazer a predição das polaridades das frases. Da mesma forma que foi citado acima, é preferível fazer a vetorização dos dados de teste com o CountVectorizer. Depois da vetorização, ocorre a transformação desses dados em dados binários, para que o algoritmo consiga compreendê-los, porém antes disso é preciso fazer o seu pré-processamento.

4.1.5. Treino do modelo de aprendizado de máquina

Depois de todo processo de limpeza de dados (pré-processamento), vetorização e identificação dos dados de treino e teste, é feito o treino do modelo. Nos dados de treino utilizou-se o algoritmo Multinomial de Naive Bayes, passando como parâmetro os 5765 tweets vetorizados, juntamente com as suas polarizações, para que ele aprenda como representar as frases e os seus respectivos sentimentos, ou seja, é apresentado um modelo supervisionado. Foi utilizado o Multinomial de Naive Bayes, que faz a classificação com recursos discretos (contagem de palavras para se classificar o texto). A distribuição multinomial requer contagens de recursos inteiros, porém, na prática, as contagens também podem ser feitas de forma fracionada.

4.4. Resultados

4.4.1. Avaliação do Modelo

Após o treino do modelo, é preciso verificar se o mesmo está entregando a predição das frases de teste de forma satisfatória. Para visualização das predições, utilizou-se o método predict, chamando as frases da lista e o modelo treinado, assim obtém-se para cada frase a sua classificação. Aqui é feita a análise de sentimento.

É importante fazer a validação do modelo, para se obter confiança na sua utilização. Para isso utilizou-se a validação cruzada, onde foram separados os dados de teste (lista com 10 frases) dos dados de treino em partes distintas. A validação cruzada irá auxiliar no teste do modelo.

Depois da vetorização, anteriormente citada, foi realizada a divisão do Dataset em 5 partes chamadas de folds. Através da técnica Cross Validation, são obtidas 5 partes, destas uma será para teste e as restantes serão para treino. Esta tarefa é feita por diversas vezes até que o modelo esteja treinado e com todas as partes treinadas, conforme Santana (2017). O Cross Validation auxilia no teste do modelo, fazendo com que todos os dados disponíveis no dataset sejam testados.

Finalizando a parte do algoritmo, a acurácia do modelo pode ser verificada, passando as classificações e os resultados da validação cruzada para o método metrics.accuracy_score. O modelo apresentou 86,85% de acurácia, porém destaca-se que o teste foi feito com uma base simples de apenas 10 frases. Esse teste é feito para que se possa saber se a ferramenta está entregando os resultados esperados. A quantidade de frases foi razoavelmente pequena, para que se verificasse cada predição feita pela ferramenta, visto que se fosse utilizado um número muito grande de frases na fase de teste, não seria possível identificar cada predição manualmente, para posteriormente utilizar-se em grande quantidade de dados.

4.4.2. Aplicação do modelo Treinado em novos tweets recuperados

Depois dos testes e avaliação do modelo, foi feita a aplicação do modelo nos dados do Tweet. Esses tweets são buscados graças à biblioteca Twitter Search que foi citada anteriormente. Primeiramente utilizou-se as chaves de acesso, para ter acesso aos dados que a API do Twitter disponibiliza para desenvolvedores. Em

seguida foi passado o termo a ser buscado e o idioma da pesquisa e, após esse processo, recebe-se os dados de quem tweetou e o texto do tweet. Na busca pelo termo “cidades de Minas Gerais” foram recuperados, para o período em que foi feita a pesquisa, 175 tweets entre tweets e retweets, os quais foram capturados numa lista.

Da mesma forma que se realizou com o Dataset, os tweets capturados passam pelo processo de limpeza, removendo as stopwords e termos indesejados, em seguida, faz-se a vetorização dos tweets e a transformação em dados binários. Utilizando o método predict como citado anteriormente e chamando os tweets junto com o modelo treinado tem-se para cada tweet recuperado a sua classificação/polarização. Aqui ocorre a análise de sentimento em si, onde há a busca por um termo recuperado, ou seja, o tweet e sua classificação é apresentada de acordo com sentimento expresso seja ele positivo, negativo ou neutro.

4.4.3. Discussão

Neste estudo de caso, foi abordado a análise de sentimentos, utilizando dados da rede Social Twitter, por meio de sua Interface de Programação de Aplicativo (API), disponibilizada para desenvolvedores, com o auxílio da linguagem de programação Python que disponibiliza várias bibliotecas as quais destacam-se: Twitter Search, NLTK, Pandas e Sklearn que auxiliaram no processo de análise de sentimentos dos dados coletados do Twitter.

Foi apresentado o fluxograma de um sistema de informação que busca os tweets de usuários sobre um determinado assunto, depois da busca, é feita a captura desses tweets numa lista. Para auxiliar na classificação/polarização dos mesmos, foi utilizada a abordagem Aprendizado de Máquina com um modelo supervisionado, que é um dataset de treinamento disponibilizado pelo governo de Minas Gerais, utilizando-o como a base de treino para nosso modelo com o amparo do algoritmo Multinomial de Naive Bayes que é utilizado pela biblioteca SKlearn. Após treinar o algoritmo e testar com a base de teste composta por uma lista contendo 10 frases, é feito um cross validation. Em seguida é verificada a acurácia e então inicia-se a classificação e análise de sentimento dos tweets recuperados que pode ser Positiva, Negativa ou Neutra.

Nos testes, observou-se que a ferramenta está entregando o que foi proposto, que é a análise de sentimentos. É importante salientar que esta análise, se feita com

um dataset com maior quantidade de sentenças polarizadas e dados bem estruturados, os resultados poderiam ser ainda melhores, visto que o dataset utilizado era um pouco limitado e seus dados não estavam bem estruturados, já que foram retirados de tweets sobre o estado de Minas Gerais e não foram pré-processados antes de serem utilizados na ferramenta.

CAPÍTULO – 5 CONSIDERAÇÕES FINAIS

5.1. Conclusões

O presente trabalho buscou explorar o campo da mineração de dados junto com a análise de sentimentos, abordando os temas mais relevantes desta área, os quais foram bem citados nos trabalhos relacionados. Neste trabalho foi proposto a implementação de um algoritmo básico que faz a análise de sentimentos de tweets pesquisados sobre determinado assunto.

Na fundamentação teórica discutiu-se as etapas e os principais métodos para a mineração de dados. Um tema bem relevante que foi o processamento de linguagem natural é uma área que anda junto com a análise de sentimentos e aprendizagem de máquina, em algoritmos que busquem fazer a polarização de textos em linguagem natural necessitam estar interligados a estas abordagens.

A ferramenta proposta neste trabalho foi um algoritmo em Python que busca fazer a classificação/polarização dos Tweets pesquisados. Foram utilizadas várias bibliotecas: Twitter Search, NLTK, Pandas e SKlearn. Também utilizou-se a abordagem de Aprendizado de Máquina e o modelo supervisionado que usou um dataset disponibilizado pelo estado de Minas Gerais, auxiliado pelo classificador multinomial de Naive Bayes.

O modelo consegue ler o dataset, vetorizar os dados de texto que são os tweets, fazer a relação entre os dados de texto do dataset e suas respectivas polaridades, foi feito o pré-processamento de ambos, removendo caracteres indesejados e linhas duplicadas. O algoritmo busca os tweets e os vetoriza assim como faz com o dataset, após essa vetorização foi realizada a predição comparando os textos rotulados no dataset aos tweets capturados, em seguida, é feita a acurácia do modelo.

Observa-se que o algoritmo é funcional, faz a vetorização, classificação e mede a acurácia do modelo, nota-se também que um dataset de qualidade auxiliaria na melhor acurácia e também no melhor treino para o modelo de Aprendizado de Máquina.

5.2. Trabalhos futuros

Visto que é um campo em constante crescente e de grande volume de publicações, os temas aqui abordados não se esgotam neste trabalho. Portanto, para os trabalhos futuros, indica-se a utilização de uma base de dados com estruturas e polarizações bem definidas para auxiliar no aprendizado do algoritmo, assim como o estudo mais aprofundado das técnicas de mineração de dados.

Por fim, é interessante salientar que o campo da mineração de dados está numa crescente ascensão e é extremamente relevante no contexto social, empresarial e educacional. Sugere-se que modelos de análise de opiniões, como o apresentado neste trabalho, sejam mais implementados por organizações, escolas e governos, para que se tenham melhores feedbacks das opiniões das pessoas, a fim de que se melhorem os processos e/ou serviços disponibilizados à comunidade.

REFERÊNCIAS

ANDRADE, Patrícia Helena Maia Alves de. **Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na CGU**. 2015. xi, 54 f., il. Dissertação (Mestrado Profissional em Computação Aplicada) - Universidade de Brasília, Brasília, 2015. Disponível em: < http://dx.doi.org/10.26512/2015.09.D.21004>. Acesso em: 14 de Agosto de 2020.

ARAÚJO, Matheus *et al*. **Métodos para análise de sentimentos no twitter**. In: Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13). 2013. Acesso em: 05 de Julho de 2020.

BECKER, Karin; TUMITAN, Diego. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. **Simpósio brasileiro de banco de dados**, v. 75, 2013. Disponível em: < https://sbbd2013.cin.ufpe.br/Proceedings/artigos/pdfs/sbbd_min_02.pdf>. Acesso em 07 de Julho de 2020.

BENEVENUTO, Fabrício *et al*. **Métodos para análise de sentimentos em mídias sociais**. In: Brazilian Symposium on Multimedia and the Web (Webmedia), Manaus, Brasil. 2015. Disponível em: <http://www.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>. Acesso em: 23 de Novembro de 2020.

CARVALHO, Jonnathan dos Santos. **Uma estratégia estatística e evolutiva para mineração de opiniões em tweets**. 2014. Dissertação de Mestrado – Universidade Federal Fluminense, Niterói, 2104. Disponível em < https://www.researchgate.net/profile/Jonnathan_Carvalho/publication/281407847_Uma_Estrategia_Estatistica_e_Evolutiva_para_Mineracao_de_Opinioes/links/55e5bb7408aebdc0f58b7442.pdf>. Acesso em: 05 de Julho de 2020.

DIANA, Daniela. Diferença entre Sarcasmo e Ironia. **Toda Matéria**, 2020. Disponível em:<https://www.todamateria.com.br/sarcasmo-e-ironia/>. Acesso em 25 de Novembro de 2020. il. color.

GONZALEZ, Marco; LIMA, Vera Lúcia Strube. Recuperação de informação e processamento da linguagem natural. In: **XXIII Congresso da Sociedade Brasileira de Computaçao**. 2003. p. 347-395.

JOHN Ruskin: A maior recompensa para o trabalho. **Pensador**, c2005. Disponivel em: <https://www.pensador.com/frase/NDMzMjg0/#:~:text=John%20Ruskin%3A%20A%20maior%20recompensa,ele%20se%20torna%20com%20isso>. Acesso em: 25 de Novembro de 2020.

LEITE, João Lucas Araújo. **Mineração de textos do Twitter utilizando técnicas de classificação**. 2015. TCC (Graduação) - Engenharia de Software, Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2015. Disponível em: < http://www.repositorio.ufc.br/handle/riufc/25016>. Acesso em: 05 de Novembro de 2020.

LORENA, Ana Carolina *et al*. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, Porto Alegre, RS, v. 14, n. 2, p. 43-67, dec. 2007. ISSN 21752745. Disponível em: <https://www.seer.ufrgs.br/rita/article/view/5690>. Doi: https://doi.org/10.22456/2175-2745.5690. Acessado em: 01 de Dezembro de 2020.

MENEZES, G. Algoritmo TCC. **Google Colab**, 2020. Disponível em: <https://colab.research.google.com/drive/1rHxW7-av0SWKHzAa8AB_hmLbzPc8wI4b?usp=sharing>. Acesso em: 19 de Novembro de 2020.

NATURAL Language Toolkit. **NLTK 3.5 documentation**, 2020. Disponível em: <https://www.nltk.org/>. Acesso em: 10 de Julho de 2020.

NASCIMENTO, André; REIS, Marcelo. **Análise de Sentimento de Insegurança Através do Twitter**. *In*: ESCOLA REGIONAL DE COMPUTAÇÃO BAHIA, ALAGOAS E SERGIPE (ERBASE), 18., 2018, Aracaju. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2018. p. 199-208. Disponível em: < https://sol.sbc.org.br/index.php/erbase/article/view/8542>. Acesso em: 23 de Outubro de 2020.

SANTANA, R. Análise de Sentimentos - Aprenda de uma vez por todas como funciona utilizando dados do Twitter. **Minerando Dados**, 2017. Disponível em: <https://minerandodados.com.br/analise-de-sentimentos-twitter-como-fazer/>. Acesso em: 07 de Julho de 2020.

SANTANA, R. Como preparar Dados de Texto para Machine Learning. **Minerando Dados**, 2020. Disponível em: <https://minerandodados.com.br/como-preparar-dados-de-texto-para-machine-learning/>. Acesso em: 14 de Novembro de 2020.

SANTOS, Leandro Matioli. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o Twitter. **Monografia. Departamento de Ciência da Computação, Universidade Federal de Lavras**, 2010.

SANTOS, Leandro Matioli et al. Twitter, análise de sentimento e desenvolvimento de produtos: Quanto os usuários estão expressando suas opiniões? **Revista Prisma. com**, n. 13, p. 159-170, 2010. ISSN1646 -3153. Disponível em: < https://ojs.letras.up.pt/index.php/prismacom/article/view/2005>. Acesso em: 28 Setembro de 2020.

SANTOS, Tatiane Gomes dos. **Análise de opiniões utilizando técnicas de mineração de dados em redes sociais. Estudo de caso: twitter**. 2017. Disponível em: <http://repositorio.aee.edu.br/jspui/handle/aee/53>. Acesso em: 15 de Julho de 2020.

SOBRE as APIs do Twitter. **Help Twitter**. Disponível em: <https://help.twitter.com/pt/rules-and-policies/twitter-api>. Acesso em: 13 de Julho de 2020.

SOUSA, Giulia Luan Santos de. **Tweetmining: análise de opinião contida em textos extraídos do Twitter**. Monografia (Graduação) - Bacharel em Sistemas de Informação, Universidade Federal de Lavras, Lavras, 2012. Disponível em: http://www.bsi.ufla.br/wp-content/uploads/2013/09/TWEETMINING-AN%C3%81LISE-DE-OPINI%C3%83O-CONTIDA-EM-TEXTOS-EXTRA%C3%8DDOS-DO-TWITTER-.pdf>. Acesso em: 22 de Setembro de 2020.

TREVISAN, Allan Caminha. **Mineração de textos no Twitter**. 2015.TCC (Graduação) - Curso de Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná, Curitiba, 2015. Disponível em:< http://repositorio.roca.utfpr.edu.br/jspui/bitstream/1/6659/1/CT_COSIS_2015_1_01.pdf>. Acesso em: 19 de Setembro de 2020.

VIEIRA, Renata; LOPES, Lucelene. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. **EM CORPORA**, p. 183, 2010.

WEIAND, Augusto. **Análise de sentimentos do Twitter com Naive Bayes e NLTK**. **Trajetória Multicursos**, [S.l.], v. 7, n. 2, p. 3 - 18, jan. 2018. ISSN 2178-4485. Disponível em: <http://sys.facos.edu.br/ojs/index.php/trajetoria/article/view/135>. Acesso em: 09 de Junho de 2020.