

UNIVERSIDADE FEDERAL DO PARANÁ

JOSIEL QUENNEHEN DE OLIVEIRA

RELATÓRIO FINAL

(Período no qual esteve vinculado ao Programa 09/2018 a 09/2019)

PROGRAMA DE IC:

- ☐ PIBIC
- ☐ PIBIC Af
- ☐ PIBIC EM
- ☒ PIBITI

MODALIDADE:

- ☐ CNPq
- ☐ UFPR TN
- ☐ Fundação Araucária
- ☐ Voluntária

ANÁLISE DE SENTIMENTOS (MINERAÇÃO DE OPINIÕES) EM BASE EXTRAÍDA DE UMA REDE SOCIAL

Relatório apresentado à Coordenação de Iniciação Científica e Tecnológica da Universidade Federal do Paraná como requisito parcial da conclusão das atividades de Iniciação Científica ou Iniciação em desenvolvimento tecnológico e Inovação - Edital 2019
Orientador(a): Prof.(a) Dr(a) Denise Fukumi Tsunoda
Título do Projeto: Recuperação e análise de informações de base de dados operacionais e históricas / 2010024931

CURITIBA

(2019)

SUMÁRIO

RESUMO.....3

1 - INTRODUÇÃO.....3

2 - REVISÃO DA LITERATURA.....5

3 - MATERIAIS E MÉTODOS10

4 - RESULTADOS E DISCUSSÕES18

5 - CONCLUSÕES19

REFERÊNCIAS.....19

RESUMO

O presente trabalho refere-se a análise das postagens do mês de maio de 2019, na rede social *Twitter*, que mencionam o recém-eleito presidente brasileiro Jair M. Bolsonaro. A coleta de dados ocorreu através da plataforma online *Netlytic*, o processamento dos dados ocorreu através da linguagem de programação *Python* que foi utilizada através do software *Jupyter Notebook* integrante da plataforma *Anaconda*. Utilizou-se para a análise técnicas de Processamento de Linguagem Natural e de Análise de Sentimento aliadas à abordagem de Aprendizagem de Máquina Supervisionada. Foi sugerido como hipótese que as postagens refletiriam a polarização política atual da sociedade brasileira atual e de acordo com a proporção de postagens favoráveis (57%) e desfavoráveis (43%) obtidas ao final da pesquisa observou-se que de fato essa polarização se manifestou.

Palavras-chave: Análise de Sentimento. Sentiment Analysis. Opinion Mining. Mineração de Opinião. Processamento de Linguagem Natural. Natural Language Processing. Jair Bolsonaro. Twitter. Polarização. Política. Python. Anaconda. Jupyter Notebook. Scikit-learn. NLTK.

1 - INTRODUÇÃO

Segundo (D'ANCONA, 2018), entramos em uma nova era marcada por um fenômeno denominado de “Pós-verdade”. Algumas de suas características seriam um intenso combate político e intelectual onde a racionalidade estaria abalada pela emoção, a diversidade pelo nativismo e o surgimento de um movimento rumo a autocracia. Nesse processo a ciência estaria cada vez mais desacreditada, as instituições fragilizadas e as mídias sociais cada vez mais assumiriam o papel, antes efetuado pela mídia impressa, da transmissão das informações.

A realidade brasileira não foge muito às mencionadas constatações de D'Ancona. Atualmente o Brasil encontra-se em um cenário extremamente polarizado politicamente, e essa polarização acaba por se difundir para os mais diversos setores da sociedade.

As redes sociais são um lugar onde essa polarização é intensamente manifesta. De um lado políticos utilizam-se das redes sociais para entrarem em contato com seus eleitores e obter *feedback* sobre a repercussão que suas medidas tem perante a sociedade. Do outro lado eleitores abrem espaços de debate nas redes sociais para criticar as posições políticas alheias ou para defender as suas próprias posições. No meio disso existe uma guerra de desinformação com a disseminação de “*fake news*” tanto por parte dos políticos quanto por parte dos próprios eleitores. Em meio a esta “guerra” de informação e desinformação torna-se difícil saber qual lado diz a verdade. Torna-se perigosa, portanto, a decisão, por parte de pessoas que dependem da opinião pública, de ficarem alheias ao que ocorre nas redes sociais.

Considerando que as redes sociais se tornaram um meio de expressão política muito eficaz tanto para a disseminação de informações quanto para o combate à desinformação é interessante notar um sensível aumento na adesão das redes sociais por parte de políticos como uma nova e eficiente maneira de atingir tanto o seu público “fiel” quanto seus “adversários políticos”.

No Brasil é notório, e até alvo de críticas, a frequente utilização do *Twitter* pelo presidente da república Jair. M. Bolsonaro. Notícias como “Bolsonaro confunde a todos ao governar pelo Twittter”¹ e “Bolsonaro faz do Twitter seu palanque virtual”² são exemplos de como naturalizou-se a utilização por parte do presidente da república do *Twitter* como plataforma de interação. Através de seu perfil ele constantemente entra em contato com seu público tanto para expor medidas políticas como para expor determinadas visões particulares acerca de determinados assuntos. Essas postagens frequentemente se disseminam pela rede social através de compartilhamentos e comentários e acabam por gerar debates entre quem é adepto de suas ideias e entre quem não é.

No entanto esses debates não são gerados apenas pelas publicações do próprio presidente. Diversos usuários comentam, postam e compartilham notícias relacionadas ao presidente e suas medidas políticas e muitas vezes isso acaba por gera intensos debates no *Twitter*.

¹ GABRIEL, Ruan de S; LIBÓRIO, Bárbara. Época: Bolsonaro faz do Twitter seu palanque virtual. Disponível em: <<https://epoca.globo.com/bolsonaro-faz-do-twitter-seu-palanque-virtual-23572419>>. Acesso em: 09 ago. 2019.

² VIANNA, Luiz F. Época: Bolsonaro confunde a todos ao governar pelo Twitter. Disponível em: <<https://epoca.globo.com/bolsonaro-confunde-todos-ao-governar-pelo-twitter-artigo-23647155>>. Acesso em: 09 ago. 2019.

Através desses debates as pessoas acabam por expressar suas opiniões e seus sentimentos acerca de determinados fatos ou assunto. Essas opiniões se tornam extremamente importantes para se ter ideia do que os usuários daquela rede pensam acerca de determinado tópico e tendo o controle de quais são essas opiniões é possível direcionar ações para reforça-las ou para modificá-las.

Uma área que lida com essas opiniões e emoções expressas por determinada entidade é a área de “*Sentiment Analysis*”, ou Análise de Sentimento. Análise de sentimento pode ser definido como o processo de descobrir e classificar opiniões expressas em determinado texto.

Esta pesquisa se torna importante na medida em que visa abordar dois temas atuais, relevantes, e que estão intrinsicamente relacionados: a polarização política brasileira e a utilização das mídias sociais para a expressão de emoções e opiniões acerca do atual presidente brasileiro Jair M. Bolsonaro.

Desta forma assumimos a hipótese, e pretendemos verificar sua validade ao final dessa pesquisa, de que devido ao ambiente altamente polarizado da sociedade brasileira em relação à política, essa polarização também seria manifesta nas postagens relacionadas ao presidente Jair M. Bolsonaro, dividindo-as em proporções favoráveis e desfavoráveis praticamente equivalentes.

O objetivo desta pesquisa é, portanto, analisar as postagens efetuadas no *Twitter*, durante o mês de maio de 2019, que mencionam o presidente Jair M. Bolsonaro e verificar se as emoções expressas no conteúdo dos *tweets* possuem caráter favorável ou desfavorável ao mesmo.

2 - REVISÃO DA LITERATURA

A Análise de Sentimento é uma área derivada de uma outra denominada de “*Natural Language Processing*”, ou seja, Processamento de Linguagem Natural. Segundo Rosa (2014), apesar de o Processamento de Linguagem Natural poder ser definido de várias formas, todas elas envolvem a noção de armazenamento de dados em máquinas e a manipulação de dados linguísticos. Em resumo seria a habilidade de um computador processar a mesma linguagem que os seres humanos utilizam no dia a dia. No entanto, essa definição pode fazer essa tarefa de processamento parecer

bem mais simples do que realmente é. Existem diversos fatores que podem dificultar a realização eficaz desse processo.

De acordo com Luger (2004, p.509):

Comunicar-se através de linguagem natural, quer seja como texto ou como um ato de fala, depende enormemente do nosso conhecimento e expectativas dentro do domínio do discurso. A compreensão da linguagem não é meramente a transmissão de palavras: ela também requer inferências sobre o objetivo, conhecimento e suposições do locutor, bem como sobre o contexto da interação. A implementação de um programa para compreender linguagem natural requer que representemos conhecimento e expectativas do domínio e raciocinemos sobre eles. Precisamos considerar questões como não monotonia, revisão de crença, metáfora, planejamento, aprendizado e as complexidades práticas da interação humana.

Apesar da dificuldade de lidar com os atos de fala, conforme mencionado acima, atualmente há um crescente interesse nessa área de pesquisa. Empresas como *Google*, *Facebook* e *Amazon* cada vez mais desenvolvem pesquisas e produzem artefatos para lidar de forma aprimorada com dados de linguagem natural. Através dessas pesquisas é possível, por exemplo, melhorar o funcionamento de aplicativos que executam tarefas através de comandos de voz, aumentar a performance e a eficiência de mecanismos de busca, melhorar processos de transcrição de textos a partir de áudios, dentre outras tarefas.

O processamento de linguagem natural basicamente se dá sobre dados originalmente desestruturados. Dados desestruturados podem ser considerados como dados que não seguem um formato/padrão específico. São exemplos de dados desestruturados: imagens, vídeos, dados de mídia social, mensagens de texto enviadas por dispositivos móveis, etc. Após uma etapa de processamento ser realizada sobre os dados desestruturados é possível transformá-los em dados estruturados para adequá-los a técnica que será utilizada para analisá-los.

A análise, dependendo do objetivo do pesquisador, pode resultar na obtenção de informações como: os termos mais frequentes de um documento, as entidades (nomes de pessoas, empresas, organizações, datas, moedas, horários, etc) presentes, os relacionamentos entre as entidades (quem / o que / onde), os principais tópicos ou assunto de determinado texto, as opiniões ou emoções expressas em determinado documento ou sentença

A Análise de Sentimento (*Sentiment Analysis*) embora utilize-se de várias técnicas e ferramentas de Processamento de Linguagem Natural, possui um foco bem específico. Ela objetiva identificar opiniões ou sentimentos associados com determinado texto.

Segundo Benevenuto, Ribeiro & Araújo ([ca 2000], não paginado):

O principal objetivo da análise de sentimentos é definir técnicas automatizadas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimento, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.

Estendendo a análise para as mídias sociais e adequando-a para o foco desta pesquisa é possível obter informações tais como: o que está sendo falado sobre mim, o que eles gostam em mim, o que eles não gostam em mim, como sou comparado a outros políticos, quão leais são meus eleitores.

Ainda segundo Benevenuto, Ribeiro & Araújo ([ca 2000], não paginado), a área de análise de sentimentos é dividida em várias frentes de pesquisa de acordo com o nível de granularidade abordado. Seriam essas frentes:

- **Estado emocional:** é a análise do estado emocional expresso em um documento
- **Análise de sentimentos para comparação:** objetiva identificar sentenças que contém termos comparativos como “maior que”, “pior que” e extrair as respectivas entidades para cada termo.
- **Nível de documento:** visa extrair uma opinião única acerca de determinado documento
- **Nível de sentença:** analisa cada frase e determina qual o sentimento/opinião/polaridade predominante nelas
- **Nível de palavra ou dicionário:** visa construir ou otimizar léxicos de sentimentos
- **Nível de aspecto:** ao contrário do nível de documento assume-se a possibilidade de que em cada texto possam haver várias opiniões, sentimentos e entidades. O objetivo é estabelecer a relação entre a entidade e o sentimento externado

Atualmente os dois métodos mais comuns para a detecção do sentimento são os baseados em aprendizagem de máquina e os métodos léxicos. Os métodos baseados em aprendizagem de máquina geralmente usam um conjunto de dados com textos já classificados que são utilizados para gerar um modelo algorítmico que pode ser utilizado para a detecção automática de textos não rotulados. Essa abordagem é chamada de aprendizado supervisionado. Também existem os métodos não-supervisionados quando nenhuma rotulação prévia existe. Já os métodos léxicos,

segundo Benevenuto, Ribeiro & Araújo ([ca 2000], não paginado), “em geral, são baseadas em tratamentos léxicos de sentimentos que envolvem o cálculo da polaridade de um texto a partir de orientação semântica das palavras contidas nesse texto.”

Nesta pesquisa a abordagem seguirá a linha dos métodos supervisionados baseados em aprendizagem de máquina. Esta abordagem, apesar de ser muito comum apresenta uma série de dificuldades. Uma delas, de acordo com Benevenuto, Ribeiro & Araújo ([ca 2000], não paginado), é:

a alta subjetividade envolvida na tarefa e a demanda de tempo necessária para que especialistas definam a polaridade de muitas sentenças. Muitas sentenças são altamente ligadas a alguma situação ou evento específico e a definição de polaridade pode ser extremamente difícil por quem não esteja inserido no contexto em questão. (...) Situações com utilização de sarcasmo e ironia também tornam a avaliação uma tarefa complexa. Soma-se a isso a necessidade de que muitas sentenças sejam rotuladas, o que poderia demandar semanas de trabalho de um especialista.

Devido a essa dificuldade mencionada acima, para esta pesquisa foi adotada como método de classificação automática de textos uma abordagem denominada de “*Distant Supervisor*”. De acordo com essa técnica são utilizados padrões que ocorrem com frequência nos textos de forma a permitirem uma rotulação automática de qual categoria pertencem. Nessa pesquisa partiu-se do pressuposto de que os textos das *hashtags* eram um indício de que o conteúdo do texto era favorável ou não ao presidente Jair M. Bolsonaro. Na prática nem sempre essa abordagem produz êxito, com textos desfavoráveis sendo classificados como favoráveis e vice-versa, ocasionando ruídos para a geração de um modelo eficiente. No entanto a classificação de uma forma geral foi mais proveitosa do que prejudicial, levando-se em consideração a dificuldade de ficar rotulando continuamente uma grande quantidade de dados para o treinamento do modelo preditivo. É importante notar também tanto a abordagem de classificação manual, quanto a abordagem léxica foram testadas, sem obter, no entanto, um custo benefício tão elevado quanto com o “*Distant Supervisor*”. As maiores dificuldades foram o processo lento de classificação manual, embora mais apurado, e a ineficiência dos métodos léxicos que analisavam os textos de acordo com a polaridade das palavras disponíveis no mesmo.

Embora haja uma grande quantidade de métodos de processamento de linguagem natural, existem determinadas técnicas que são utilizadas com mais frequência. Algumas delas são:

Tokenization: É o processo de quebrar textos em sentenças, sentenças em palavras e palavras em caracteres. Qual tarefa dessas 3 tarefas mencionadas será executada depende do objetivo do pesquisador, e somente através de um objetivo claro é que é possível determinar qual a técnica mais adequada. Uma frase como “Esta é uma frase de exemplo”, após a tokenização se transformaria em um vetor de palavras independentes: [“Esta”, “é”, “uma”, “frase”, “de”, “exemplo”].

Removing Noise: Consiste em remover todas as informações que não são relevantes para o contexto do texto a ser analisado. Isso inclui a remoção de pontuações, números, espaços em branco e palavras comuns, também chamadas de “stop words”.

Part of Speech Tagging: É bastante utilizada em abordagens que se utilizam de dicionários léxicos. Consiste em associar cada palavra à sua função gramatical (sujeito, adjetivo, verbo, advérbio, etc) na sentença ou texto.

Stemming: É o processo de transformar uma palavra em sua forma raiz. Como exemplo podemos citar as palavras (perder, perdido, perdedor). Após o processo de stemming essas palavras seriam transformadas respectivamente em (perd, perd, perd). Essa transformação reduz o número de palavras únicas na base de dados, melhorando o consumo de recurso computacional e aumentando a performance dos modelos preditivos. Os algoritmos mais comuns para a realização desta tarefa são: *Porter*, *Lancaster* e *Snowball*.

Lemmatization: É o processo de transformar a palavra em sua forma mais simples. Seguindo com o exemplo dado acima as palavras (perder, perdido, perdedor) se transformariam respectivamente em (perder, perder, perder).

Bag of Words (BoW): Consiste em um método onde é formado um dicionário com todas as palavras que se tenha a intenção de verificar a ocorrência ou não delas em um outro texto a ser utilizado como entrada. Esse método apenas verifica a ocorrência ou não da palavra, e em alguns casos também pode verificar quantas vezes ela ocorre. Cada palavra do dicionário é percorrida e caso ela esteja presente no texto de entrada é assinalado na matriz a sua ocorrência (1) ou não (0).

N-grams: Consiste em criar um conjunto de palavras que ocorrem de forma sequencial. É bastante utilizada para que as palavras não percam muito do contexto em que estão inseridas dentro do texto completo. A sequência de palavras geralmente ocorre de 2 em duas ou de 3 em 3. Como exemplo, a frase (1) “Não, hoje tomei café da manhã” e esta outra frase (2) “Hoje não tomei café da manhã”, depois de um pré-

processamento poderiam ter o mesmo significado num modelo preditivo caso a técnica de *bag of words* seja utilizada. Isso pode confundir o modelo na hora de realizar a classificação de sentenças. Abaixo segue o exemplo com a técnica de *bag of words* simples e com a técnica de *bag of words* associada com a 2-gram. O número 1 nas células indica a presença na frase e o 0 indica ausência.

TABELA 1 – Representação de um *Bag of Words*

	nao	hoje	tomei	cafe	da	manha	classe
Frase(1)	1	1	1	1	1	1	x
Frase(2)	1	1	1	1	1	1	Y

FONTE: O autor (2019)

TABELA 2 – Representação de um *Bag of Words* associado a técnica de *N-Gram*

	nao hoje	hoje tomei	tomei cafe	cafe da	da manha	nao tomei	classe
Frase(1)	1	1	1	1	1	0	x
Frase(2)	0	1	1	1	1	1	y

FONTE: O autor (2019)

Term Frequency-Inverse Document Frequency (TF-IDF): É uma abordagem estatística que mede a relevância de um termo em um documento, ou em uma coleção deles. Segundo essa abordagem um termo que ocorre em muitos documentos não é um bom discriminador, portanto, esses termos mais comuns devem receber um peso menor do que o dado às características menos frequentes.

3 - MATERIAIS E MÉTODOS

Quanto ao nível desta pesquisa, ela pode ser considerada de caráter descritivo (GIL, 2002) na medida em que tem como objetivo descrever as características dos dados coletados assim como estabelecer relações entre suas variáveis.




Os dados utilizados neste estudo foram recolhidos através de uma ferramenta online (<https://netlytic.org>) que coleta automaticamente dados de redes sociais e que também possui recursos para análise de dados e visualização da informação.

De acordo com a sua própria definição em sua homepage:

Netlytic is a cloud-based text and social networks analyzer that can automatically summarize and discover communication networks from publicly available social media posts. It uses public APIs to collect posts from Twitter, Youtube, and Facebook (public pages). It also supports analysis of your own datasets.

No momento desta pesquisa ela possui 3 planos para sua utilização, sendo que dois (*Tier 1*, *Tier 2*) são gratuitos e um (*Tier 3*) é pago. Para a execução deste projeto foi utilizado o plano “*Tier 3*” já que pela quantidade de dados que deveriam ser armazenados isso não seria possível com os planos inferiores.

FIGURA 1 – Planos disponíveis para a utilização do *Netlytic*

Account Types			
<p>We are committed to maintain free access to this service for Tier 1 & 2 accounts.</p> <p>However, collecting and analyzing millions of data points from social media require a lot of computing power. To make sure that we have enough “juice” to keep Netlytic running smoothly, we rely on a commercial web hosting company to run this tool. If you like Netlytic, please support hosting of this project by upgrading to “Tier 3?”.</p> <p>Below is more information about each Tier.</p>			
			
	Tier 1 (Free)	Tier 2 (Free)	Tier 3 (Community-supported)
Max # of Datasets	3	5	300
Max # of Records/Dataset	2500	10000	100000
	Great for exploring what Netlytic can do!	Great for smaller projects and class assignments!	Great for larger research projects and brand management tasks!
	This is a default tier	Request a free upgrade by logging in to your account and clicking on the “My Account” page	
			UPGRADE

FONTE – O Autor (2019)

A rede social escolhida para que se realizasse a coleta foi o Twitter. Para realizar a coleta dos dados o Netlytic utiliza a própria API pública do Twitter e se faz necessário a vinculação de uma conta do Twitter do usuário à plataforma do Netlytic.

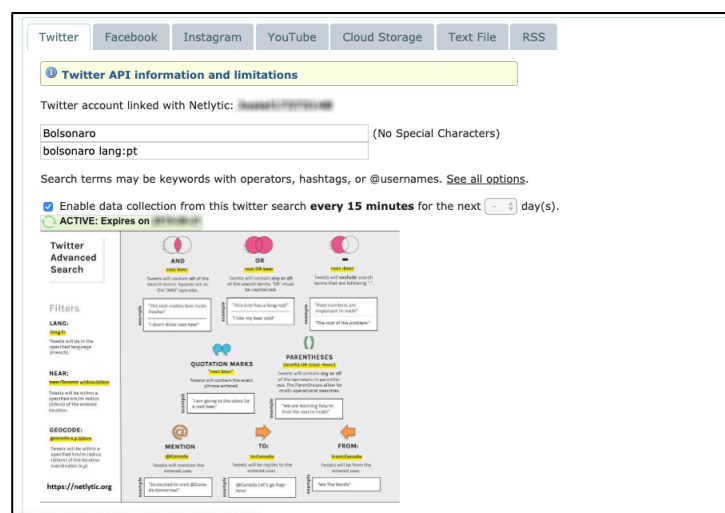
Para realizar as pesquisas a plataforma utiliza os métodos de busca e busca avançada (<https://twitter.com/search-advanced>) do próprio *Twitter*. É possível configurar o *Netlytic* para que ele realize pesquisas a cada 15 minutos e por uma quantidade de dias determinados que pode ir de 1 a 62. Quando o prazo de dias está para expirar o usuário é notificado e pode então renovar o prazo caso seja de seu interesse. Após expirado o prazo, e caso a renovação não tenha acontecido, a base de dados permanece disponível na plataforma, porém, sem coletar novos dados.

Cada pesquisa realizada pela plataforma, obedecendo uma limitação da própria *API* do *Twitter*, retorna até 1000 *tweets* e se limita a pesquisar somente os *tweets* dos últimos 7 dias. A cada pesquisa realizada a base de dados pré-existente é então incrementada com os novos *tweets* que ainda não estavam disponíveis. Nos planos “*Tier 3*” e “*Tier 2*” cada *dataset* pode ter um máximo de 100.000 registros. Quando esse limite é atingido é criado automaticamente um novo *dataset* com os novos dados que estão sendo coletados.

As bases de dados coletadas ficam disponíveis na própria plataforma e é possível realizar o download delas nos formatos ‘.xlsx’ ou ‘.csv’.

Para realizar a coleta de dados foi necessário a vinculação de uma conta do *Twitter* à plataforma do *Netlytic*. Após essa etapa é necessário criar um nome (no âmbito dessa pesquisa foi escolhido o nome de “Bolsonaro”) para o *dataset* que será coletado, assim como configurar os termos de busca (“bolsonaro lang:pt”) que serão executados de acordo com os parâmetros que a *API* de busca do *Twitter* aceita. Os termos utilizados para a busca determinam que seja executada a busca automática de todos os *tweets* que mencionem o termo “bolsonaro” e que estejam em língua portuguesa. A coleta foi então configurada para que se executasse a busca a cada 15 minutos durante o mês inteiro de maio de 2019.

FIGURA 2 – Como configurar a ferramenta para coletar dados do *Twitter* no *Netlytic*



FONTE – O Autor (2019)

Os dados coletados geraram 9 *datasets* com 100.000 *tweets* cada, 1 *dataset* com 99.981 e um último com 995.84, totalizando 1.099.565 *tweets*. Tanto no primeiro

dataset quanto no último foram registrados tweets dos meses de Abril e Junho respectivamente. Houve então a necessidade de removê-los já que o escopo do projeto definia que apenas o mês de Maio seria analisado. Para a remoção dos registros que não seriam utilizados juntamente com as demais etapas de pré-processamento que serão mencionadas a seguir foi utilizado o software *Jupyter Notebook*, incluído na plataforma *Anaconda*. No *Jupyter Notebook* foi utilizado *Python* como linguagem de programação para processar os dados.

Após a remoção dos *tweets* dos meses que não correspondiam à Maio, o primeiro *dataset* ficou com 76.001 *tweets*, com a remoção de 23.999 referentes ao mês de Abril, e o último *dataset* ficou com 20.993 *tweets* após a remoção de 79.007 *tweets* referentes ao mês de Junho.

Após a etapa mencionada acima todos os *datasets* foram agrupados originando um único conjunto de dados com um total de 996.559 registros.

Para se obter uma ideia geral do conjunto de dados foi realizada a Análise Exploratória dos Dados (*EDA*). Através dessa análise foi possível perceber que os *tweets* que tinham registrado a localização do usuário seguiam a seguinte distribuição em ordem decrescente por quantidade de *tweets* das 5 primeiras cidades:

TABELA 3 – Cidades com mais tweets durante o mês de maio de 2019 na base de dados analisada

Cidade	Quantidade de tweets
São Paulo	58517
Rio de Janeiro	52311
Brasília	16085
Belo Horizonte	15184
Curitiba	11165

FONTE – O Autor (2019)

Os 5 usuários que mais possuíam tweets no mês de Maio seguiam esta distribuição em ordem decrescente:

TABELA 4 – Usuários com mais tweets durante o mês de maio de 2019 na base de dados analisada

Usuários	Quantidade de tweets
Usuário A	965
Usuário B	679
Usuário C	640

Usuário D	551
Usuário E	528

FONTE – O Autor (2019)

A data de criação dos perfis dos usuários seguia esta distribuição para as 5 primeiras datas, também em ordem decrescente:

TABELA 5 – Data de criação dos perfis que postaram durante o mês de maio de 2019 na base de dados analisada


Data de criação do perfil	Quantidade de perfis
2018-12-18 18:11:05	965
2013-08-04 13:34:50	679
2011-02-28 06:54:39	640
2012-01-22 08:51:44	551
2009-07-06 06:32:38	528

FONTE – O Autor (2019)

A média de amigos para cada perfil de usuário nos dados coletados foi de 1.037. A mediana foi de 367. A média de seguidores foi de 6.294 para cada perfil enquanto que a mediana foi de 242.

Existiam diversos perfis que possuíam a mesma descrição. A descrição dos perfis dos usuários teve a seguinte distribuição:

TABELA 6 – Perfis com descrições repetidas que postaram durante o mês de maio de 2019 na base de dados analisada

Descrição do perfil	Quantidade com a mesma descrição
	1314
#LulaLivre	1259
...	946
.	835
sem tempo irmão	730

FONTE – O Autor (2019)

Os aparelhos que foram utilizados para a realização dos tweets foram os seguintes para as 5 plataformas mais utilizadas:

TABELA 7 – Plataformas mais utilizadas nas postagens durante o mês de maio de 2019 na base de dados analisada

Plataformas	Qtd. tweets por plataforma
Twitter for Android	543559
Twitter for iPhone	218025
Twitter Web App	92567
Twitter Web Client	89932
Facebook	27248

FONTE – O Autor (2019)

Após essa etapa (EDA) foram eliminados todos os *retweets*. Isso foi necessário para diminuir a quantidade de recurso computacional exigido para a criação do modelo preditivo e também para evitar que os *retweets* tivessem um peso maior no resultado final da pesquisa. A eliminação foi feita excluindo todos os *tweets* cujos títulos contivessem a palavra “RT”.

Após a eliminação restaram 227.821 registros, sendo eliminados, portanto, 768.738 registros.

Mesmo com essa eliminação foi verificado que ainda existiam *tweets* em que o título não identificava que eles eram *retweets*, mas a mensagem do *tweet* ainda se repetia. Esses *tweets* (1.048 registros) também foram eliminados resultando num conjunto de dados final de 226.773 registros.

Após o conjunto de dados a ser analisado ter sido formado, teve início da classificação de *tweets* em favoráveis ou desfavoráveis à entidade pesquisada, ou seja, Jair Bolsonaro. A técnica utilizada para a classificação foi a já mencionada “*Distant Supervisor*”. Esta foi a técnica escolhida devido ao fato de estar-se utilizando a abordagem de aprendizado de máquina supervisionado e a decorrente dificuldade de estar classificando manualmente a quantidade de dados necessários para se treinar um modelo eficientemente. Outras abordagens foram testadas, como por exemplo a baseada na polaridade das palavras constantes na sentença, mas o resultado não foi satisfatório.

A classificação seguiu as seguintes etapas:

- 1) Através do uso de expressão regular foram retornadas as *hashtags* mais frequentes em cada dia de maio.

2) Baseado no texto das *hashtags* mais frequentes, quando o texto dava margem para a classificação, classificava-se o *tweet* em favorável ou desfavorável. Como exemplo de *hashtags* favoráveis temos: #EstouComBolsonaro, #BolsonaroOrgulhodoBrasil e #elesim. Como exemplo de *hashtags* desfavoráveis temos: #BolsonaroEnvergonhaOBrasil, #ForaBolsonaro e #EleNao. Também foram procurados termos que geralmente davam indícios de que o *tweet* era favorável ou desfavorável. Como exemplo dos termos desfavoráveis temos: minion, bozo, marielle, queiroz, etc. Como exemplo dos termos favoráveis temos: mortadela e cuba. Seguem abaixo as *hashtags* mais frequentes nos *tweets* coletados, por dia, no mês de maio:

TABELA 8 – Distribuição das *hashtags* mais frequentes por dia, durante o mês de maio de 2019, na base de dados analisada

Dia	Hashtag	Quantidade
1	#OsPingosNosIs	492
2	#CancelBolsonaro	3247
3	#CancelBolsonaro	1985
4	#BoicoteBurgerKing	853
5	#BolsoSilvioVemAi	985
6	#BolsoNoSBT	135
7	#SomosTodosLeticia	237
8	#TexasCancelBolsonaro	81
9	#ArmasMatam	55
10	#VamosInvadirBrasilia	478
11	#VamosInvadirBrasilia	308
12	#MoroVendido	631
13	#MoroVendido	1001
14	#TsunamiDaEducação	982
15	#TsunamiDaEducação	3864
16	#TsunamiDaEducação	640
17	#BolsonaroNossoPresidente	2076
18	#BolsonaroNossoPresidente	2796
19	#Dia26EuVou	2400
20	#Dia26EuVou	1905
21	#NordesteCancelaBolsonaro	2109

22	#NordesteComBolsonaro	926
23	#Dia26ReformasJa	1143
24	#DomingoPeloBrasil	2540
25	#DomingoPeloBrasil	1748
26	#BrasilNasRuas	3910
27	#Brasil	955
28	#QuemMandouMatarBolsonaro	847
29	#Tsunami30M	1604
30	#30MpelaEducacao	3461
31	#BolsonaroNoTheNoite	1404

FONTE – O Autor (2019)

- 3) Como resultado obteve-se um conjunto de dados com sentenças desfavoráveis com 11.007 registros e um outro conjunto de dados com 10.271 registros, totalizando 21.277 registros.

Nestes dois conjuntos de dados gerados foi realizada uma etapa de pré-processamento que buscou corrigir palavras grafadas incorretamente ou abreviadas, corrigir pontuações, eliminar espaços em branco desnecessários, substituir o endereço de links pela palavra “*link*”, remoção de menções a usuários e a “stemmização” utilizando o algoritmo *RSLPStemmer*, algoritmo esse adequado para o processamento de textos de língua portuguesa. Esse algoritmo faz parte do pacote da biblioteca *NLTK* da linguagem de programação *Python*.

Posteriormente foi utilizada a biblioteca *Scikit-learn* para separar os dados classificados em dados de treinamento e dados de teste na proporção de 90% e 10% respectivamente. Após essa etapa o conjunto de dados foi transformado numa matriz “*bag of words*” associada a já mencionada técnica 2-gram. Foi gerada então uma matriz com 106.458 conjuntos de palavras de 2 em 2.

Para o treinamento foi utilizado o algoritmo “*MultinomialNB*” disponível na biblioteca *Scikit-learn*. O modelo gerado através desse treinamento alcançou uma acurácia de 89% nos dados de treinamento.

Após a utilização deste modelo algorítmico gerado nos 226.773 registros que são o foco deste estudo, foi obtido o seguinte resultado:

TABELA 9 – Resultado final da análise de sentimento nas postagens do mês de maio de 2019 na base de dados analisada

	Total	%
Tweets desfavoráveis	128.336	57%
Tweets favoráveis	98437	43%
Total de Tweets	226.773	100%

FONTE – O Autor (2019)

É interessante notar que existe uma margem tolerável para erros já que *tweets* são catalogados de forma errônea para ambos lados. Também existe o fato de que muitos *tweets* são links de notícias, por vezes claramente identificáveis com uma tendência favorável, por outras com uma tendência desfavorável. Também existem as notícias que são simplesmente relatos ou links, que pendem mais para a neutralidade. Acrescenta-se ao que já foi mencionado que existem diversos *tweets* que pendem para a ironia ou para a piada, tornando difícil sua classificação até mesmo quando feita de forma manual.

4 - RESULTADOS E DISCUSSÕES

Conforme mencionado no início do trabalho, partiu-se da hipótese de que devido ao momento atual da sociedade brasileira, com uma crescente polarização devido em grande parte a adesão pela sociedade de posições políticas antagônicas, essas posições polarizadas também se manifestariam através das postagens nas redes sociais. Conforme ficou demonstrado ao final da pesquisa houve uma divisão de certa forma igualitária entre as postagens favoráveis e desfavoráveis ao atual presidente. Deve-se ressaltar que isso reflete o resultado obtido com uma população específica em um momento específico. Apesar dessa especificidade o resultado acaba por espelhar um cenário condizente com o que ocorre na sociedade atualmente.

Observando-se a distribuição das *hashtags* através dos dias de maio é possível observar muitas vezes uma ação de uma posição política e já no dia seguinte uma reação da outra ala. Isso demonstra uma espécie de “guerra” para se obter a hegemonia do discurso, ou da “verdade”, nas redes sociais.

5 - CONCLUSÕES

Iniciamos esta pesquisa com uma hipótese em mente: a polarização política existente atualmente na sociedade brasileira provavelmente seria refletida nas postagens das redes sociais, especificamente no Twitter. Escolheu-se trabalhar com as postagens que mencionassem o recém-eleito presidente do Brasil Jair M. Bolsonaro, um frequente usuário do Twitter. Após a coleta efetuada com a ferramenta online Netlytic, foi feito o pré-processamento através da linguagem de programação Python, utilizando os softwares Anaconda e Jupyter Notebook e após a aplicação de técnicas de Processamento de Linguagem Natural e de Análise de Sentimento verificou-se que a hipótese sustentada inicialmente foi validada pelo resultado final da pesquisa, com proporções semelhantes de postagens favoráveis e contrárias ao presidente. Como artefato originário da pesquisa tanto os códigos quanto as bases de dados ficarão disponibilizados no Github (<https://github.com/josieloliveira99/iniciacao-cientifica-analise-sentimento-bolsonaro>).

REFERÊNCIAS

BENEVENUTO, Fabrício; Ribeiro, Filipe; Mathews Araújo. Métodos para análise de sentimentos em redes sociais. Disponível em: <<https://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>>. Acesso em 09 ago. 2019.

D'ANCONNA, Matthew. Pós-verdade: a nova guerra contra os fatos em tempos de fake News. Barueri: Faro Editorial, 2018.

GABRIEL, Ruan de S; LIBÓRIO, Bárbara. Época: Bolsonaro faz do Twitter seu palanque virtual. Disponível em: <<https://epoca.globo.com/bolsonaro-faz-do-twitter-seu-palanque-virtual-23572419>>. Acesso em: 09 ago. 2019.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. São Paulo: Atlas, 2002.

LUGER, George F. Inteligência Artificial: estruturas e estratégias para a solução de problemas complexos. Porto Alegre: Bookman, 2004.

ROSA, João Luís Garcia. Fundamentos da inteligência artificial. Rio de Janeiro: LTC, 2014.

VIANNA, Luiz F. Época: Bolsonaro confunde a todos ao governar pelo Twitter. Disponível em: <<https://epoca.globo.com/bolsonaro-confunde-todos-ao-governar-pelo-twitter-artigo-23647155>>. Acesso em: 09 ago. 2019.