A Sentiment Analysis of Brazilian Elections Tweets

André L. Cristiani[1], Douglas D. Lieira[2,3], Heloisa A. Camargo[1]

[1] Universidade Federal de São Carlos, UFSCar, Brazil
andre.cristiani@estudante.ufscar.br, heloisa@dc.ufscar.br
[2] Universidade Estadual Paulista, UNESP, Brazil
douglas.lieira@unesp.br
[3] Instituto Federal de São Paulo, IFSP, Brazil

Abstract. The internet connection is present in people's lives all the time, through smartphones, tablets, computers, among others. The use of social networks is increasingly common around the world. Many companies and people use them to spread products and services and publish opinions, facts that have turned the social networks into powerful sources of information on various topics. Identifying these feelings is a great strategy for many types of decision making. Thus, the purpose of this paper is to collect messages from a specific social network, in this case Twitter, referring to the 2018 Brazilian presidential elections and classify them as: positive, negative and neutral, in order to discover a possible relationship between opinions of social network users and the final outcome of the elections. For this, a corpus was built, preprocessed and evaluated by two different machine learning approaches: Naive Bayes and SVM (Support Vector Machine). The results showed that this social network is a good source of information to perform sentiment analysis and that the number of tweets classified as positive have a strong relationship with the researchers and the final result of the 2018 elections.

CCS Concepts: • **Applied computing**;

Keywords: Social Network, Twitter, Brazilian presidential elections, Sentiment analysis

1. INTRODUÇÃO

As redes sociais são plataformas *online* com o intuito de criar rede de conexões que os usuários utilizem para as mais variadas finalidades, como: vender ou divulgar produtos, buscar empregos, compartilhar notícias, publicar opiniões sobre diversos acontecimentos, entre outros [Teixeira and Azevedo 2011]. Segundo uma pesquisa divulgada pelo site *eMarketer*[1], no ano de 2017, o Brasil é o país que mais acessa redes sociais na América Latina, tendo cerca de 93,2 milhões de usuários mensais ativos.

A análise de sentimentos é um ramo da mineração de dados focado na classificação de textos através de sentimentos e opiniões, pelo qual se pode descobrir o que as pessoas pensam e opinam sobre um determinado assunto, evento, acontecimento ou produto. Consequentemente, com o aumento do uso intensivo das redes sociais, a análise de sentimentos de usuários vem ganhando notoriedade e cada vez mais importância na descoberta de conhecimento [França et al. 2014].

O *Twitter* é uma rede social fundada em 2006, com o intuito de permitir que seus usuários publiquem mensagens de texto de até no máximo 140 caracteres. O *Twitter* é uma das redes sociais mais populares do mundo, tendo cerca de 319 milhões de usuários ativos mensalmente [Jianqiang et al. 2018].

No ano de 2018, ocorreram as eleições presidenciais do Brasil, na qual milhões de brasileiros foram às urnas para eleger o novo líder do país. Este evento de âmbito mundial gerou várias pesquisas envolvendo milhares de pessoas. Jornais, revistas e sites divulgaram diversas estatísticas contendo a

[1]Disponível em: https://www.emarketer.com/Chart/Social-Network-User-Penetration-Latin-America-by-Country-July-2017-of-internet-users/215406

opinião da população de diferentes regiões do país. Este acontecimento também gerou milhares de publicações em redes sociais, contendo opiniões dos usuários, sendo essas uma boa fonte de pesquisa.

Diversos autores utilizam o *Twitter* como fonte de dados para estudos sobre análise de sentimentos, acerca dos mais variados temas, por ser uma ótima fonte de opinião dos internautas [Pak and Paroubek 2010]. Segundo o site G1[2], as eleições brasileiras do ano de 2014 geraram quase 40 milhões de *tweets*, o que faz deste um tema relevante para aplicar técnicas de análise de sentimentos e descobrir se existe alguma relação entre a opinião dos usuários da rede social e o resultado final das eleições.

O objetivo deste trabalho é aplicar técnicas de análise de sentimentos para avaliar se existe uma relação entre a opinião dos usuários do Twitter e o resultado final das eleições. Com base nisso, foram realizadas coletas de *tweets*, durante alguns eventos relacionados às eleições, como debates, entrevistas e os domingos eleitorais. Em seguida, uma pequena parte dos *tweets* coletados foram rotulados manualmente pelos autores deste trabalho em três polaridades: positivo, negativo e neutro, de acordo com o sentimento expresso pelo usuário na mensagem. Todos os *tweets* foram submetidos a um processo de pré-processamento, com a finalidade de normalizar os dados, tornando-os mais adequados para a etapa de aprendizagem. Por fim, foi realizado o treinamento do classificador, utilizando dois algoritmos de aprendizado de máquina diferentes, *Naive Bayes* e SVM (*Support Vector Machine*), para gerar o modelo de classificação. O restante dos *tweets* não rotulados foram classificados automaticamente. Após a classificação, os usuários da rede social foram divididos em grupos específicos de interesse e os resultados foram comparados com pesquisas e o resultado final das eleições.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta a metodologia adotada para a realização do trabalho. A metodologia de avaliação e os resultados obtidos são discutidos na Seção 4. Por fim, a Seção 5 conclui este trabalho e lista trabalhos futuros.

2. TRABALHOS RELACIONADOS

Existem diversos trabalhos na literatura que aplicam algoritmos de aprendizado de máquina, com o intuito de analisar conteúdos textuais de redes sociais, afim de identificar o sentimento dos usuários, porém pouco destes trabalhos lidam com textos escritos na língua portuguesa; sendo que grande parte deles trabalham com textos escritos na língua inglesa. Esta seção apresenta uma análise sobre os trabalhos recentementes publicados que aplicam análise de sentimentos em dados de redes sociais.

Correa *et al.* [Correa et al. 2018] apresentam um estudo referente à análise de sentimentos de *tweets* relacionados aos filmes indicados ao prêmio de melhor filme, afim de buscar uma correlação entre o sentimento expresso pelos usuários do Twitter e o resultado da cerimônia do Oscar de 2017. Os autores realizaram a coleta durante todo o período de divulgação, até o dia anterior ao evento. Após a coleta, 3.235 *tweets* foram classificados manualmente e utilizados no treinamento de classificadores com três algoritmos: Naive Bayes, TextBlob e Sentiment140. Após os testes, o Naive Bayes obteve melhor acurácia (74,1%) e foi selecionado para realizar a classificação da base de dados completa. Com base no resultado da classificação, os autores identificaram que a metodologia proposta é útil para descoberta de informações sobre os filmes indicados ao Oscar, porém, poucas associações foram encontradas entre o *ranking* obtido como resultado do trabalho e o *ranking* final do Oscar de 2017.

Silva, Caseli e Teixeira [Faria Silva et al. 2017] apresentam uma comparação entre as técnicas de aprendizado supervisionado Naive Bayes e SVM, na utilização de um modelo de filtro de *tweets* relevantes para uso de aplicativos voltados à televisão social. Para tal, os autores realizaram a coleta de *tweets* referentes ao *talent show* culinário *MasterChef* Brasil. Após a coleta, 450 *tweets* foram anotados como relevante e não relevante e submetidos a fase de pré-processamento. Após a etapa de pré-processamento, os *tweets* foram submetidos para o treinamento e teste dos classificadores, onde o SVM se saiu melhor, obtendo uma acurácia de 86,9% e o Naive Bayes obtendo 80,6%.

[2]Disponível em: http://g1.globo.com/politica/eleicoes/2014/noticia/2014/10/eleicoes-brasileiras-geraram-quase-40-milhoes-de-tuites-diz-twitter.html

Öztürk e Ayvaz [Öztürk and Ayvaz 2018] investigaram as opiniões e os sentimentos dos usuários do Twitter em relação à crise dos refugiados sírios, comparando a opinião dos usuários da língua inglesa e turca. Os *tweets* turcos continham informações para analisar a percepção pública de um país anfitrião de refugiados, visto que a Turquia foi o país que acolheu o maior número de refugiados sírios. Após a análise dos sentimentos, os autores observaram que os sentimentos sobre sírios e refugiados era de uma proporção maior no turco (35%) do que no inglês (12%). Os autores também ressaltam que os *tweets* turcos foram distribuídos, quase uniformemente, entre positivos, neutros e negativos, enquanto os *tweets* ingleses eram, em grande parte, compostos por neutros e negativos.

Budiharto e Meiliana [Budiharto and Meiliana 2018] apresentam uma análise da previsão do resultado das eleições presidenciais da Indonésia de 2019, por meio de mensagens do Twitter no idioma indonésio. Neste trabalho não foram apresentados resultados de precisão obtidos com o algoritmo, porém, de acordo com os autores, o Twitter provou ser uma ferramenta válida para extração de opinião. Os resultados obtidos se mostraram confiáveis e correspondentes às pesquisas divulgadas por quatro institutos da Indonésia.

Diferente dos trabalhos apresentados, os quais grande parte realizam as análises de textos escritos na língua inglesa, este estudo busca trabalhar com textos escritos na língua portuguesa, sendo tais mensagens extraídas da rede social Twitter, com o conteúdo relacionado às eleições presidenciais no Brasil em 2018, aplicando técnicas de análise de sentimentos e buscando obter uma relação entre a opinião dos usuários da rede social e o resultado final das eleições.

3. METODOLOGIA

Algumas etapas foram percorridas a fim de alcançar o objetivo proposto neste trabalho. Essas etapas encontram-se detalhadas na seguinte subseção. O conjunto de dados e as codificações produzidas para a realização deste trabalho podem ser encontrados no GitHub[3].

3.1 Coleta dos dados

O Twitter foi utilizado como fonte de dados para esse trabalho. Para realizar a captura das mensagens, foi utilizada a API (Application Programming Interface) pública oficial disponibilizada pela rede social, que retorna apenas o conteúdo textual das mensagens publicadas pelos usuários. A coleta dos dados foi realizada em diversos eventos durante a campanha eleitoral do ano de 2018. A Tabela III representa todos os eventos em que foram realizadas coletas, juntamente com a data em que o evento aconteceu, o número total de tweets coletados e as palavras-chaves utilizadas para as buscas. Todas as coletas iniciaram uma hora antes do evento e finalizaram uma hora após o término do evento. Esta etapa resultou na coleta de 903.518 *tweets* que possuem alguma relação com as eleições presidenciais brasileiras de 2018.

3.2 Anotação dos dados

Após a etapa de coleta dos dados, deu-se início ao processo de anotação manual dos tweets, afim de obter um conjunto de dados para treinamento e testes para utilizá-los na criação do modelo de aprendizado supervisionado. Para isso, uma parte dos tweets foram selecionados, aleatoriamente, e classificados em três polaridades, de acordo com o seu conteúdo. As polaridades consideradas foram: (a) Positivo: quando o usuário publica uma mensagem demonstrando apoio ao candidato; (b) Negativo: quando o usuário publica uma mensagem demonstrando rejeição ao candidato e (c) Neutro: quando o usuário publica uma mensagem sobre um ou mais candidatos, mas não demonstra sua opinião sobre nenhum deles.

No total, 600 *tweets* foram anotados manualmente, entre as três polaridades consideradas. Para isso, foi levado em consideração o conteúdo da mensagem e o candidato que essa fazia referência. Sendo

[3]Disponível em: https://github.com/andrecristiani/analise-de-sentimentos-eleicoes-2018

Table I: Eventos em que foram realizadas coletas.

Evento	**Data**	**Palavras-chave**
Debate na Band	08/08/2018	"#DebateNaBand", "debate"
Entrevista com Jair Bolsonaro no Jornal Nacional	28/08/2018	"#BolsonaroNoJN"
Entrevista com Geraldo Alckmin no Jornal Nacional	29/08/2018	"#AlckminNoJN"
Entrevista com Marina Silva no Jornal Nacional	30/08/2018	"#MarinaNoJN"
Entrevista com João Amoêdo no Jornal Nacional	31/08/2018	"#JoaoNoJN"
Debate na Rede Globo	04/10/2018	"#DebateNaGlobo" e "debate"
Primeiro domingo eleitoral	07/10/2018	"#Eleição2018", "#Eleições2018", "eleicao", "eleição", "eleições" e "eleiçoes"
Segundo domingo eleitoral	27/10/2018	"#Eleição2018", "#Eleições2018„ "eleição","eleições", "eleiçoes", “#ViraVirouHaddad” e “#ObrasilVota17”

assim, se a mensagem faz referência para um determinado candidato, essa mensagem será positiva, negativa ou neutra para aquele candidato.

3.3 Agrupamento por candidatos

Após a anotação dos dados, para os eventos os quais havia a participação de dois ou mais candidatos, houve a necessidade de dividir os dados em grupos específicos, remetendo cada *tweet* para o candidato cujo conteúdo da mensagem fazia referência. Para tal, após uma longa análise dos dados coletados, foram selecionadas algumas palavras para fazer parte de um dicionário de palavras, que foi desenvolvido por meio de um script em python, com o intuito de realizar a divisão e agrupamento de *tweets* para seus respectivos candidatos. A Tabela II representa o dicionário de palavras criado para realizar o agrupamento dos *tweets* por candidatos.

Table II: Dicionário de palavras utilizadas no agrupamento de candidatos.

Candidato	**Palavras**
Álvaro Dias	"alvaro", "álvaro", "19"
Cabo Daciolo	"cabo", "daciolo", "deus", "deux", "deuxx", "51"
Ciro Gomes	"ciro", "viraviraclr0", "ciro12", "viraviraciro12", "12"
Fernando Haddad	"haddad", "hadad", "andrade", "haddadpresidente", "haddad13", "agoraéhaddad", "haddaptando", "13neles", "ptsim", "viravirouhaddad", "13"
Guilherme Boulos	"bulos", "psol", "50"
Geraldo Alckmin	"alckmin", "xuxu", "chuchu", "45"
Henrique Meirelles	"meirelles", "meireles", "15"
Jair Bolsonaro	"bolsonaro", "coiso", "bolso", "mito", "bonoro","bozonaro", "bolzonaro", "elenão", "elesim","bolsonaro17", "17neles", "b17", "jair", "obrasilvota17", "17"
João Amoêdo	"amoedo", "amoêdo", "partidonovo", "30"
Marina Silva	"marina", "18"

3.4 Pré-processamento dos tweets

Após a coleta e anotação dos dados, deu-se início a etapa de pré-processamento dos tweets. Esta etapa tem como finalidade padronizar e preparar os dados para serem utilizados nas próximas etapas. As técnicas utilizadas neste processo foram desenvolvidas utilizando as bibliotecas NLTK[1] e Scikit-Learn[2] do python, que estão descritas abaixo.

[1]Disponível em: https://www.nltk.org/
[2]Disponível em: https://scikit-learn.org/stable/

3.4.1 *Preparação e padronização.* Primeiramente, todo o *corpus* foi padronizado em letra minúscula. Após a padronização, algumas informações utilizadas pelo Twitter, que não definem a opinião dos usuários, foram removidas, como: *links* para *websites*, citações de outros usuários e *hashtags*.

3.4.2 *Tokenização.* A primeira etapa no pré-processamento de textos é a tokenização, que consiste em quebrar o fluxo de caracteres em palavras [Weiss et al. 2010]. Nesta etapa, cada *tweet* foi quebrado em um vetor de palavras, utilizando como delimitadores os espaços em brancos entre as palavras.

3.4.3 *Remoção de stopwords.* Uma das tarefas mais utilizadas em pré-processamento de textos é a remoção de *stopwords*. Esse método consiste em remover palavras muito frequentes que não agregam conteúdo semântico ao texto, sendo informações irrerelevantes para o modelo [Manning et al. 2010]. Para isso, foi utilizado a lista de *stopwords* em português, disponível pela biblioteca NLTK em python.

3.4.4 *Lematização.* A lematização é uma técnica muito utilizada em buscadores de palavras em *websites*, por possibilitar que a busca incorpore o maior número de palavras relacionadas à busca. Esta técnica consiste em reduzir uma palavra ao seu lema, que é a sua forma no masculino e singular, diminuindo o número de palavras no vocabulário [De Lucca and Nunes 2002]. Para isso, foi utilizado o pacote *Wordnet Lemmatizer* da biblioteca NLTK do python.

3.4.5 *TF-IDF.* O TF-IDF é uma técnica muito utilizada em mineração de textos para descobrir a importância de palavras em um texto semi ou não estruturado [Lima and de Castro 2012]. Essa técnica consiste em percorrer todo o texto e buscar palavras que possuam maior importância no texto. Os termos mais comuns em um único ou pequeno grupo de documentos tende a ter maior TF-IDF em relação as palavras que aparecem várias vezes, como preposições, pronomes e artigos.

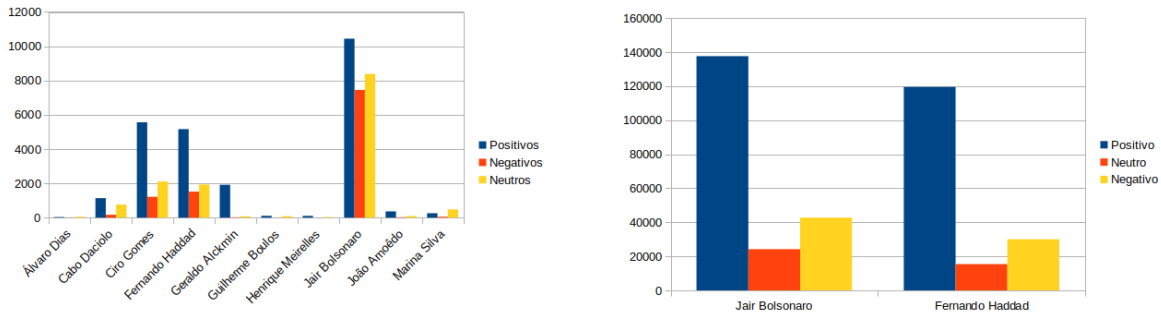
O desempenho deste método depende fortemente da quantidade de itens no vocabulário, não sendo indicado para documentos de tamanhos pequenos, pois a probabilidade de repetição de termos é mínima, gerando valores próximos para os termos [Choi et al. 2014]. Para utilizar o método TF-IDF bem como calcular seu valor para cada termo, foi utilizado sua implementação disponibilizada pela biblioteca Scikit-Learn em python.

3.5 Algoritmos de aprendizado supervisionado

Para a classificação dos *tweets*, dois algoritmos de aprendizado de máquina supervisionados foram utilizados, sendo o Naive Bayes e SVM. Para isso a implementação destes algoritmos, foi utilizado a biblioteca Scikit-Learn.

O Naive Bayes é um simples classificador probabilístico, baseado no teorema de Bayes. Este classificador descreve a probabilidade de um evento, de acordo com a base de conhecimento prévio das condições relacionadas ao evento. Ele utiliza-se dos dados de treinamento para criar um modelo probabilístico baseado nas características dos dados, supondo que há uma independência nas características do modelo, fazendo com que uma característica seja independente das outras para a classificação [Lacerda and Braga 2004]. Este algoritmo é um dos mais utilizados para a tarefa de classificação de textos, por essa razão fez-se a escolha do mesmo. Para este trabalho, a versão Multinomial do Naive Bayes foi utilizada.

O SVM é um classificador binário muito utilizado na tarefa de classificação de textos, principalmente, pelo seu bom desempenho [Firmino Alves et al. 2014]. Este classificador busca traçar um hiperplano entre dados de duas classes, buscando maximizar a distância entre os pontos mais próximos em relação a cada uma das classes. O SVM é uma boa opção para classificação de textos, principalmente quando a base de dados de treinamento contém poucos exemplos, sendo um dos motivos principais para a seleção desse algoritmo. Para este trabalho, foi considerado kernel linear com parâmetro $C = 1.0$.



(a) Coleta do primeiro domingo eleitoral.

(b) Coleta do segundo domingo eleitoral.

Fig. 1: Resultados obtidos para as coletas referentes ao primeiro e segundo domingo eleitoral.

4. RESULTADOS

Esta seção tem como objetivo analisar os resultados obtidos por meio do *corpus* coletado e pré-processado e buscar uma relação entre a opinião dos usuários do Twitter e o resultado final das eleições de 2018.

4.1 Classificação dos *tweets*

Primeiramente, foi realizado um experimento para avaliar o desempenho dos dois classificadores e selecionar o melhor deles para classificar o conjunto de dados completo. As medidas de avaliação adotadas foram: acurácia, precisão, recall e f-measure. Para isso, 70% dos dados de treinamento foram utilizados para treinar os classificadores e os 30% restantes foram utilizados para avaliar seus desempenhos. A Tabela III mostra os resultados obtidos pelos dois algoritmos. É possível notar que o classificador SVM obteve melhor desempenho em todas as medidas de avaliações analisadas.

Table III: Desempenho obtidos pelos classificadores.

Classificador	**Acurácia**	**Precision**	**Recall**	**F-Measure**
Naive Bayes	56,11%	54,01%	63,23%	54,21%
SVM	66,66%	65,22%	71,05%	66,16%

Com base nisso, o modelo de classificação foi gerado a partir do classificador SVM e todo o conjunto de dados foi rotulado. Como os dados estavam divididos por candidatos, os tweets coletados no primeiro domingo eleitoral foram divididos em dez candidatos; e os coletados no segundo domingo eleitoral foram divididos em dois candidatos. No total, doze conjuntos de dados diferentes foram classificados por meio do modelo de classificação gerado.

4.2 Análise de sentimento dos *tweets*

Nas Figuras 1a e 1b é possível visualizar a distribuição dos *tweets* e os sentimentos expressos para cada um dos candidatos, considerando as três polaridades analisadas. A Figura 1a representa os sentimentos expressos para os dados coletados durante o domingo eleitoral do primeiro turno das eleições. Neste evento, foram analisados 48.985 *tweets* referentes aos dez candidatos que concorreram à presidência do Brasil. É possível notar que cinco candidatos sobressaíram-se em quantidade de mensagens positivas, que são: Jair Bolsonaro, Ciro Gomes, Fernando Haddad, Geraldo Alckmin e Cabo Daciolo. O mesmo acontece para esses candidatos em quantidades de mensagens negativas, com exceção de Geraldo Alckmin, que quase não obteve mensagens negativas.

A Figura 1b representa os sentimentos expressos para os dados coletados durante o domingo eleitoral do segundo turno das eleições. Neste evento, foram analisados 369.800 *tweets* referentes aos dois

candidatos que foram para o segundo turno das eleições e concorreram à presidência do Brasil. É possível notar que o candidato Jair Bolsonaro teve uma quantidade maior de citações positivas à seu adversário, porém, também obteve um maior número de citações negativas.

4.3 Comparação com o resultados finais das eleições de 2018

Com o intuito de analisar os dados obtidos por este trabalho, os resultados foram comparados com pesquisas feitas durante as eleições presidenciais, realizadas por dois órgãos de pesquisas, que são: Instituto de Pesquisas Datafolha e Ibope (Instituto Brasileiro de Opinião Pública e Estatística). Os dados foram obtidos do site UOL e estão separados em primeiro[4] e segundo[5] turnos.

A Tabela IV apresenta os quatro candidatos com maior número de eleitores, de acordo com os resultados obtidos pelas pesquisas do Datafolha e Ibope e uma comparação com os dados obtidos da rede social Twitter. Analisando a Figura 1a, é possível notar que os mesmos candidatos apontados pelos orgãos de pesquisa, como os que possuem maior parte dos eleitores a seu favor, foram também os que mais tiveram *tweets* positivos publicados, o que mostra que o resultado final pode ter uma correlação com o sentimento dos usuários da rede social.

Table IV: Comparação com pesquisas realizadas no primeiro turno.

Candidato	**Datafolha**	**Ibope**	**Twitter**		
			Positivo	**Neutro**	**Negativo**
Jair Bolsonaro	40%	41%	10.437 (21.31%)	7.451 (15.21%)	8.383 (17.11%)
Fernando Haddad	25%	25%	5.163 (10.54%)	1.521 (3.11%)	1.933 (3.95%)
Ciro Gomes	15%	13%	5.564 (11.36%)	1.220 (2.49%)	2.108 (4.30%)
Geraldo Alckmin	8%	8%	1.290 (2.63%)	22 (0.04%)	78 (0.16%)

A Tabela V apresenta os dois candidatos que concorreram no segundo turno das eleições de 2018. É possível notar que, assim como na pesquisa realizada pelo Ibope e Datafolha, o candidato Jair Bolsonaro teve um número maior de mensagens positivas, quase 5% a mais que seu adversário, a mesma diferença apontada pelos orgãos de pesquisa nacionais.

Table V: Comparação com pesquisas realizadas no segundo turno.

Candidato	**Datafolha**	**Ibope**	**Twitter**		
			Positivo	**Neutro**	**Negativo**
Jair Bolsonaro	55%	54%	137.691 (37.23%)	24.219 (6.54%)	42.781 (11.57%)
Fernando Haddad	45%	45%	119.602 (32.34%)	15.448 (4.18%)	30.059 (8.13%)

Os resultados mostram que o Twitter é uma ótima fonte de informação, principalmente, como fonte de pesquisas sobre a opinião de seus usuários. Os números mostram que a quantidade de mensagens positivas possuem uma forte relação com as pesquisas divulgadas e com o resultado final das eleições de 2018. Embora os estudos não mostraram a relação entre os *tweets* classificados como neutros e negativos com as pesquisas e o resultado final das eleições, estas informações podem ser úteis para auxiliar em diversos tipos de tomadas de decisão.

5. CONCLUSÃO

Esse trabalho apresentou um processo de mineração de textos e análise de sentimentos para estudar mensagens coletadas da rede social Twitter, referentes às eleições presidenciais de 2018, buscando uma relação entre o resultado final das eleições e a opinião dos usuários da rede social. Para isso, diversas etapas de análise de sentimentos foram executadas, como: coleta de tweets, anotação dos *tweets*, pré-processamento dos *tweets*, classificação dos *tweets* e análise dos resultados.

[4]Disponível em: https://noticias.uol.com.br/politica/eleicoes/2018/pesquisas-eleitorais/brasil/1-turno/
[5]Disponível em: https://noticias.uol.com.br/politica/eleicoes/2018/pesquisas-eleitorais/brasil/2-turno/

Após testar os classificadores com o conjunto de dados anotados e analisar os resultados com diversas métricas de avaliação, o algoritmo SVM foi eleito como o melhor para a classificação de todos os *tweets* do conjunto de dados. Com isso, foi possível notar que esse classificador é uma ótima escolha para tarefas de classificação de textos quando se trata de abordagens utilizando o aprendizado de máquina supervisionado, visto que ele obteve uma alta acurácia, precisão e recall.

Com base nos resultados obtidos, foi possível concluir que essa abordagem é útil para conduzir pesquisas e estudos sobre a opinião dos usuários da rede social referentes às eleições do Brasil. Os resultados apontaram que a análise dos *tweets* positivos para os candidatos possuem uma forte relação com as pesquisas oficiais realizadas por grandes empresas.

É importante ressaltar que a metodologia utilizada neste trabalho pode ser utilizada em outros tipos de estudos e pesquisas que envolvam conteúdos textuais. Isso vale também para o conjunto de dados, que se torna público e pode ser utilizado em outros estudos. Como trabalhos futuros, pretende-se melhorar o conjunto de dados de treinamento, aumentando a quantidade de exemplos rotulados e o número de anotadores. Além disso, utilizar algoritmos mais recentes, como redes neurais ou técnicas que utilizam *word embedding*, buscando aumentar a eficiência da classificação.

6. AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001.

REFERENCES

Budiharto, W. and Meiliana, M. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big Data* 5 (1): 51, Dec, 2018.

Choi, D., Ko, B., Kim, H., and Kim, P. Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications* vol. 38, pp. 16–21, 2014.

Correa, I. T., Abdala, D. D., Miani, R. S., and Faria, E. R. Sentiment analysis of twitter posts about the 2017 academy awards. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, Porto Alegre, RS, Brasil, pp. 320–331, 2018.

De Lucca, J. and Nunes, M. d. G. V. Lematização versus stemming. *USP, UFSCar, UNESP, São Carlos, São Paulo*, 2002.

Faria Silva, C., Caseli, H., and Teixeira, C. Classificação de tweets por relevância para concepção de um modelo de aprendizado de máquina para uso em aplicações de tv social. In *Anais do XIV Encontro Nacional de Inteligência Artificial e Computacional*, 2017.

Firmino Alves, A. L., Baptista, C. d. S., Firmino, A. A., Oliveira, M. G. a. d., and Paiva, A. C. d. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*. WebMedia '14. ACM, New York, NY, USA, pp. 123–130, 2014.

França, T. C., Faria, F., Miceli, C., Rangel, F., and Oliveira, J. Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. *Anais do SBBD (Porto Alegre). SBC*, 2014.

Jianqiang, Z., Xiaolin, G., and Xuejun, Z. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access* vol. 6, pp. 23253–23260, 2018.

Lacerda, W. and Braga, A. Experimento de um classificador de padrões baseado na regra naive de bayes. *INFOCOMP Journal of Computer Science* 3 (1): 30–35, 2004.

Lima, A. C. and de Castro, L. N. Automatic sentiment analysis of twitter messages. In *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*. IEEE, pp. 52–57, 2012.

Manning, C., Raghavan, P., and Schütze, H. Introduction to information retrieval. *Natural Language Engineering* 16 (1): 100–103, 2010.

Pak, A. and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. Vol. 10. pp. 1320–1326, 2010.

Teixeira, D. and Azevedo, I. Análise de opinião expressas nas redes sociais. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, 12, 2011.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media, 2010.

Öztürk, N. and Ayvaz, S. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics* 35 (1): 136 – 147, 2018.