10ª Jornada Científica e Tecnológica
7º Simpósio da Pós-Graduação
INSTITUTOS FEDERAIS 10 ANOS
INSTITUTO FEDERAL Sul de Minas Gerais
ISSN 2319-0124

EXTRAÇÃO E ANÁLISE DE SENTIMENTOS DE INFORMAÇÕES RELACIONADAS À ELEIÇÃO PRESIDENCIAL DO BRASIL EM 2018

Bruce D. A. ALMEIDA[1]; **Hugo RESENDE**[2]

RESUMO

A rede social Twitter possibilita a criação e o compartilhamento de conteúdos, os quais podem compreender opiniões e informações relevantes. Particularmente, em anos de eleição, partidos políticos tendem a se beneficiar dessa rede social, pois a utilizam como recurso de divulgação de políticas de candidatos e, além disso, como mecanismo investigativo de opiniões e estatísticas referentes aos partidos. A partir da coleta de dados do Twitter, torna-se passível a análise de objetos desses dados por meio de diversos algoritmos da literatura computacional, dentre os quais se tem o classificador Naive Bayes e a função conhecida por *polarity*, utilizados na análise de sentimentos de conteúdos textuais. Deste modo, este trabalho visa realizar uma análise de sentimentos expressos por usuários do Twitter em relação ao primeiro turno da eleição presidencial brasileira de 2018. Serão coletados os dados objetos de análise e os algoritmos supracitados serão utilizados para classificá-los. A partir dos resultados da classificação, os dados obtidos serão empiricamente analisados, de modo a verificar se o que foi reportado pelos algoritmos é próximo do resultado real do primeiro turno das eleições.

Palavras-chave:
Mineração de Texto em Tweets; Algoritmo de Naive Bayes; Twitter.

1. INTRODUÇÃO

As redes sociais tornaram-se nos últimos anos ferramentas de comunicação usadas por milhões de pessoas, as quais produzem informações todos os dias. Uma dessas redes sociais é o Twitter, que possui cerca de 330 milhões de usuários ativos mensalmente (INVESTOR RELATIONS, 2017), a qual fornece serviços para que todos os usuários sejam capazes de criar e compartilhar ideias e informações instantaneamente por meio de tweets (TWITTER, 2018).

Com a popularização das redes sociais, diversos estudos e ferramentas foram desenvolvidos para recuperação, reconhecimento de padrões e análise de sentimentos das informações nelas produzidas (PANG; LEE, 2008). Nesse contexto, a Mineração de Textos, pode ser definida como um conjunto de técnicas que buscam obter informações textuais, com o objetivo de extrair padrões e informações não triviais em dados não-estruturados (TAN, 1999). Tais dados são entendidos pelos computadores como uma sequência de caracteres não passíveis de extração do conhecimento.

[1] Bruce D. A. Almeida, IFSULDEMINAS – *Campus* Passos. E-mail: bruce.almeida@alunos.ifsuldeminas.edu.br.
[2] Hugo Resende, IFSULDEMINAS – Campus Passos. E-mail: hugo.resende@ifsuldeminas.edu.br.

Assim, utilizando-se o Processamento de Linguagem Natural (PLN), é possível estruturar os elementos textuais, visando facilitar a extração de conhecimento dos respectivos dados.

O PLN, uma das técnicas de Mineração de Textos, pode ser definido como um conjunto de técnicas computacionais que utiliza de análises linguísticas para representar dados textuais (LIDDY, 2001). A Análise de Sentimentos, uma área de pesquisa do PLN, tem como finalidade entender os sentimentos que os usuários apresentam a respeito de alguma entidade de interesse e tem sido bastante utilizada neste aspecto.

Diferentes abordagens para a classificação de textos são estudadas e aplicadas em diversos contextos. O algoritmo supervisionado Naive Bayes, devido à sua eficiência em alguns tipos de cenários, é amplamente utilizado na literatura para classificação de textos e análise de sentimentos (WANG; MANNING, 2012). Por essa razão, ele será utilizado neste trabalho em conjunto com a função *polarity*, que classifica tweets de acordo com sua polaridade. Essa função, assim como diversas funções aplicadas na classificação de textos, pode ser utilizada por meio da biblioteca TextBlob, implementada em linguagem de programação Python, (TEXTBLOB, 2018).

No ano de 2018 haverá a eleição presidencial brasileira e, devido à repercussão nacional desse evento, é esperado que o Twitter contenha uma ampla variedade de informações a respeito dos candidatos presidenciáveis. Por meio da filtragem dessas informações, as quais podem incluir sentimentos e/ou opiniões, torna-se viável indicar futuros cenários para as eleições.

Nesse sentido, este trabalho tem como objetivo principal obter um modelo capaz de minerar tweets com conteúdo relacionado ao primeiro turno da eleição presidencial brasileira em 2018 e, a partir da análise de sentimentos desses tweets, analisar potentes padrões comportamentais. Por meio dessa análise, espera-se identificar o nível de aceitação de cada candidato presidenciável e seus respectivos indicadores de influência.

2. MATERIAL E MÉTODOS

Na etapa de coleta de tweets referentes à eleição presidencial brasileira de 2018, optou-se por desenvolver um *Web Crawler Agent* (WCA), um tipo de programa capaz de buscar por informações de interesse na Internet e coletar textos que serão utilizados para a extração de conhecimento. Para a extração de tweets, o WCA utilizará de uma das bibliotecas disponibilizadas pelo Twitter, conhecida como REST API.

Uma vez que o WCA esteja concluído, espera-se executá-lo durante o período de uma semana, em horários coincidentes e após o término dos debates eleitorais. O principal objetivo da

extração nesse período é obter tweets com conteúdos relacionados às eleições como um todo e aos candidatos de forma individualizada.

Uma vez que a extração de dados seja realizada, aplicar-se-á um pré-processamento nos tweets com o propósito de descartar informações que são irrelevantes para a etapa de classificação. As etapas de pré-processamento a serem realizadas são (i) remoção de links, uma vez que eles não possuem conteúdo semântico; (ii) conversão de todas as letras para minúsculas, com o objetivo de padronizar o texto; (iii) remoção de caracteres especiais e acentuação, devido ao fato de não possuírem relevância para a classificação; e (iv) remoção de stopwords, como artigos, preposições e palavras sem muito valor semântico, a fim de otimizar o tempo de processamento.

Uma vez que os dados estiverem pré-processados, será implementado o algoritmo de classificação Naive Bayes e far-se-á o uso da função *polarity* para análise de sentimentos dos tweets, a fim de classificá-los como sentimentos *positivos*, *negativos* ou *neutros*. Como o algoritmo Naive Bayes é do tipo supervisionado, será realizado um treinamento desse algoritmo com um conjunto de tweets coletados, os quais terão os seus sentimentos rotulados previamente e de forma manual, com o objetivo de se aumentar a taxa de acerto do algoritmo. Uma vez implementado, o modelo com o classificador e com a função de polaridade será utilizado para classificar a base de tweets coletados pelo WCA.

Na análise, os dados extraídos e classificados serão comparados com os resultados do primeiro turno obtidos nas urnas, de modo a ratificar as informações coletadas a partir da rede social. Essa comparação será norteada por meio da criação de gráficos que indicarão os candidatos mais populares e as suas respectivas influências sobre a população no Twitter, bem como possibilitará uma análise sobre os resultados obtidos se comparados ao resultado real do primeiro turno da eleição.

3. RESULTADOS ESPERADOS E CONSIDERAÇÕES FINAIS

Como resultados esperados, estima-se que as tecnologias escolhidas sejam capazes de suprir as necessidades para um adequado desenvolvimento desta pesquisa. Almeja-se que as tecnologias utilizadas para a extração e a análise de tweets mostrem-se adequadas para as suas finalidades pré-estabelecidas. Particularmente, é esperado que o algoritmo Naive Bayes e a função *polarity* sejam capazes de classificar os tweets com uma boa taxa de acerto, se comparado aos demais contextos nos quais eles são utilizados como classificadores.

Após o desenvolvimento do modelo responsável por coletar os tweets e classificá-los de

acordo com os seus sentimentos, espera-se obter informações que possam indicar futuros cenários para a eleição presidencial referente ao primeiro turno. Dentre os dados coletados, uma porção será utilizada para criação da base de dados rotulada que será responsável pelo treinamento do algoritmo Naive Bayes.

Com os tweets classificados, almeja-se apresentar informações com a porcentagem de opiniões relacionadas a cada candidato e suas respectivas aceitações por regiões geográficas. Por fim, serão realizadas análises nas informações obtidas, com intuito de verificar se essas informações condizem com o resultado real do primeiro turno das eleições.

Como contribuição principal desta pesquisa, espera-se que seja possível realizar um paralelo comportamental entre as informações que são geradas na rede social Twitter e o resultado do primeiro turno da eleição presidencial. Com isso, poderá ser viável verificar se as informações sobre os candidatos que são geradas realmente condizem com a real situação do primeiro turno, ou se o que é reproduzido no Twitter é de certa maneira tendencioso e manipulado.

REFERÊNCIAS

INVESTOR RELATIONS. **Selected Company Metrics and Financials**. Disponível em: <https://investor.twitterinc.com>. Acesso em: 10 abr. 2018.

LIDDY, Elizabeth D. Natural Language Processing. In Encyclopedia of Library and Information Science, 2001, New York.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, v. 2, n. 1–2, p. 1-135, 2008.

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

TEXTBLOB. **TextBlob**: Simplified Text Processing. Disponível em: <https://textblob.readthedocs.io>. Acesso em: 20 mai. 2018.

TWITTER. **Rules and Policies**. Disponível em: <https://help.twitter.com/pt/rules-and-policies>. Acesso em: 11 abr. 2018.

WANG, Sida; MANNING, Christopher D. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012, Stanford, p. 90-94.