Uma metodologia de análise de sentimentos dos candidatos as eleições presidenciais de 2018 no Twitter

A methodology of sentiment analysis of the candidates for the 2018 presidential elections on Twitter.

Guilherme Guimarães de Queiroz [1] orcid.org/0000-0003-4038-040X
Leandro Maciel Almeida [2] orcid.org/0000-0001-8025-0517

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil.

E-mail do autor principal: Guilherme Guimarães de Queiroz ggq@ecomp.poli.br

Resumo

Este artigo apresenta uma metodologia para análise de sentimentos aplicada em *tweets* realizados pelos candidatos com maior intenção de voto no primeiro turno das eleições presidenciais brasileiras de 2018. A ideia do projeto nasceu a partir da dificuldade em avaliar o conteúdo das postagens dos candidatos, devido a escala considerável de dados gerados durante a campanha, e a possibilidade de criar uma análise das similaridades entre os comportamentos dos candidatos. Os *tweets* foram submetidos a técnicas de pré-processamento, uso de dicionários léxicos e algoritmos para agrupamento de dados. Os resultados obtidos permitiram a identificação de comportamentos como o grau de positividade ou negatividade dos candidatos, considerando fatores como a divulgação de pesquisas de intenção de votos realizadas, grau de similaridade e a frequência de termos utilizados nas postagens.

Palavras-Chave*:* Análise de Sentimento; Dicionário Léxico; *K-Means*; Twitter; Eleições.

Abstract

This paper presents a methodology for sentiment analysis applied to tweets made by the candidates with higher voting intentions in the first round of the Brazilian presidential elections 2018. The concept for the project came from the difficulty in evaluating the content of the applicant's posts due to the considerable scale of data generated during the campaign and the ability to create a similarity analysis among the conduct of candidates. The tweets were submitted to preprocessing techniques, use of lexical dictionaries and algorithms for grouping data. The results had provided the identification of conduct such as the rating of positivity or negativity of the candidates considering factors like: disclosure of intentions of votes, similarity rating and frequency of terms used in the posts.

Key-words: *Sentiment Analysis; Lexicon Dictionary; K-Means; Twitter; Elections.*

DOI: 1025286/repa.v5i1.1173

1 INTRODUÇÃO

A cada nova eleição, as plataformas *online* ganham novas possibilidades de uso, graças a atualizações realizadas na lei. O artigo 57-C da Lei nº 9.504/1997 **[1]**, por exemplo, proibia o uso de propaganda política na internet durante o período eleitoral. A partir da Lei n° 13.488/2017 **[2]**, a propaganda passou a ser permitida, inclusive por meio de impulsionamento do conteúdo publicado em redes sociais.

Em paralelo, estudos realizados pela Pew Research Center comprovam que 41% da população brasileira utilizam alguma plataforma de mídias sociais para receber notícias diariamente **[3]**. Esses fatores se transformam em oportunidade para que os candidatos consigam interagir ainda mais com uma parcela significativa do eleitorado – interação que não necessariamente precisaria se iniciar durante o período eleitoral, permitindo estreitar a relação entre candidatos e usuários das redes sociais.

Segundo Marques (2006, p. 167) **[4]**, "a internet é tomada, por diferentes autores, como uma espécie de "revigorante" da esfera pública política argumentativa". O autor considera as redes sociais como uma oportunidade para estabelecer um canal de informação e para a construção de debates, permitindo estreitar os laços a um baixo custo.

De um modo geral, graças ao crescimento acelerado de conteúdo nas redes sociais, houve incentivo ao desenvolvimento de sistemas para análise de sentimentos (AS) por meio da extração de conteúdo útil de mensagens textuais **[5]**.

Este artigo propõe uma metodologia de análise de sentimentos que realiza a extração, tratamento e classificação dos *tweets* gerados, interpretando o sentimento transmitido por cada mensagem e identificando seu grau de polaridade a partir da aplicação de dicionários para classificação das palavras em positivas ou negativas. Os dados obtidos foram analisados considerando as datas nas quais houve publicação de pesquisas de intenção de voto, com o objetivo de identificar possíveis comportamentos e similaridades entre os candidatos. Os dados gerados também foram submetidos a métodos de agrupamento de dados com o propósito de descobrir novas similaridades entre os candidatos por meio de suas postagens.

O artigo está dividido da seguinte forma: a seção 2 apresenta um embasamento teórico sobre redes sociais, análise de sentimentos e o k-means; a seção 3 tem como objetivo apresentar os materiais e métodos aplicados, descrevendo com detalhes a base de dados, o método utilizado para conexão, as técnicas de tratamento e a representação dos dados utilizados durante essa pesquisa; a seção 4 apresenta os resultados obtidos; a seção 5, por fim, apresenta as conclusões obtidas após as análises dos resultados encontrados ao final da aplicação prática dos métodos utilizados.

2 TRABALHOS RELACIONADOS

Existem outras publicações que sustentam a metodologia aplicada para este artigo. Como Exemplo, temos a publicação de Tumasjan (2010) **[6]**, que utiliza os tweets com o objetivo de prever um possível resultado nas eleições alemãs, realizadas em 2009. Foram utilizados cerca de 100 mil tweets contendo referências aos políticos, que junto a ferramenta LIWC (*Linguistic Inquiry and Word Count*), tornaram possível a identificação de sentimentos, classificando os conteúdos em emoções positivas ou negativas.

Considerando publicações realizadas com o objetivo de analisar o cenário político nacional, temos o trabalho feito por Attux (2017) **[7]**, que propõe uma análise preditiva para as eleições presidenciais de 2014, a partir de *tweets* gerados por usuários da rede social Twitter, coletados de agosto a outubro de 2014. Os dados foram submetidos a etapas de coleta e pré-processamento, para posteriormente serem aplicadas técnicas de classificação com o objetivo de identificar a polaridade dos *tweets* analisados.

Este Artigo, diferente dos citados nos parágrafos anteriores, propõe uma análise de sentimentos em *tweets* coletados diretamente nos perfis oficiais dos candidatos as eleições presidenciais de 2018, com o propósito de identificar comportamentos similares entre os usuários, utilizando técnicas de classificação de polaridade no conteúdo textual dos *tweets* e agrupamento desses dados.

3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão descritos os tópicos que constroem o referencial teórico do artigo.

3.1 Redes Sociais

Uma rede social é uma estrutura composta por indivíduos conectados socialmente ou através de dispositivos, e que tem como objetivo a construção e apoio de relações de caráter social. Segundo Aguiar (2006, p. 14) **[8]**, redes sociais são, antes de qualquer coisa, relações entre pessoas, estejam elas interagindo em causa própria, em defesa de outrem ou em nome de uma organização.

Segundo Torres (2009), redes sociais digitais caracterizam-se como "sites na internet que permitem a criação e o compartilhamento de informações e conteúdos pelas pessoas e para as pessoas [...]" **[9]**, sendo compostas de um conjunto de atores e suas relações **[10]**.

Ainda por Aguiar (2006): "Mas tanto a rede social quanto o sistema em rede podem ser mediados ou não por tecnologias de informação e comunicação (TICs); ou ainda serem híbridos – quando parte dos seus participantes não têm acesso a essas tecnologias, formando "teias invisíveis" que se comunicam com a rede através de "indivíduos-ponte"; O uso dessas ferramentas permite maior comunicação, e por consequência, a transmissão de conteúdos em uma escala considerável" **[8]**.

3.2 Análise de Sentimentos (AS)

A análise de sentimentos tem por objetivo identificar e extrair, de forma automática, as opiniões, sentimentos e emoções expressadas em um texto **[11]**.

A polarização das palavras é a representação do grau de positividade, negatividade ou neutralidade emocional de uma frase **[12]**. Para identificar essa polarização, se faz necessário o uso de métodos léxicos por meio de dicionários que classifiquem as palavras do texto de acordo com a polaridade transmitida. Sendo assim, a definição da polarização é fundamental para realizar a análise de sentimentos, pois ela irá definir o retorno que o dicionário léxico dará em cada análise.

Para realização da análise de sentimentos, é necessário efetuar, em primeiro lugar, o pré-processamento dos textos com o objetivo de eliminar acentuações, abreviações, preposições, artigos e conectivos, visto que não agregam um sentido específico a frase. Após essa etapa, se torna possível o uso de técnicas de análise de sentimentos por meio da aplicação de métodos léxicos **[13]**. A etapa de pré-processamento também se faz necessária para possibilitar a redução no volume dos dados utilizados durante a análise de sentimentos, considerando fatores como limitações referentes a memória e capacidade de processamento para bases de dados muito grandes.

3.3 TF-IDF

O TF-IDF é uma abreviação do inglês "*Term Frequency-inverse Document Frequency*". O TF indica a frequência de termos ou palavras que ocorrem em um determinado documento **[14]**. Esse cálculo é complementado pelo inverso da frequência do termo na coleção de documentos, para diminuir o peso dos termos que ocorrem com muita frequência nesses documentos **[14]**. Já o IDF representa o inverso da frequência nos documentos. Se tornou um método de interpretação estatística, tendo como fundamento base a ponderação de termos, que é especificado pela função inversa do número de documentos no qual esse termo ocorre **[14]**.

O resultado gerado a partir do TF-IDF aumenta proporcionalmente à medida que aumenta o número de ocorrências do termo no documento. Esse valor é equilibrado pelo inverso da frequência, definida para cada termo **[15]**.

3.4 Algoritmo K-Means

O algoritmo k-means é uma técnica de agrupamento de dados baseada em grupos particionados. Tem como vantagens a capacidade de escalabilidade e a eficiência em grandes conjuntos de dados e, mesmo tendo sido proposto em 1957, ainda é um dos algoritmos mais utilizados para agrupamento devido à facilidade de implementação, simplicidade e eficiência **[16]**. Caracteriza-se por não utilizar um supervisor para definir previamente os padrões que serão gerados, fundamentado em técnicas de realocação baseada em similaridade, posicionando os centróides dentro de cada grupo **[17]**.

O algoritmo k-means inicia a partir da definição do número inicial de k, que determina os grupos e centros para o agrupamento de dados. A definição do k pode ser de forma aleatória ou utilizando alguma heurística **[18]**. Com a definição dos centros de k grupos, o processo iterativo inicia com o cálculo da distância de cada dado para todos os centros. A menor distância determina a pertinência do dado a um grupo. Após a atribuição de todos os dados aos k grupos, ocorre a atualização dos centros com base nos valores médios de seus membros. Este processo se repete até que os dados se estabilizem ou por um número especificado de iterações **[19]**.

Para identificar a quantidade aproximada de k grupos necessários para a aplicação do k-means neste artigo, o uso do método de *elbow* foi o selecionado. O método de *elbow* faz crescer a quantidades de clusters a partir de 1 (um) e analisa o melhor resultado após cada incremento, com base no número da amostra. Quando o benefício não for mais relevante, encontra-se o melhor resultado de k, conforme apresentado na Figura 1.

DOI: 1025286/repa.v5i1.1173

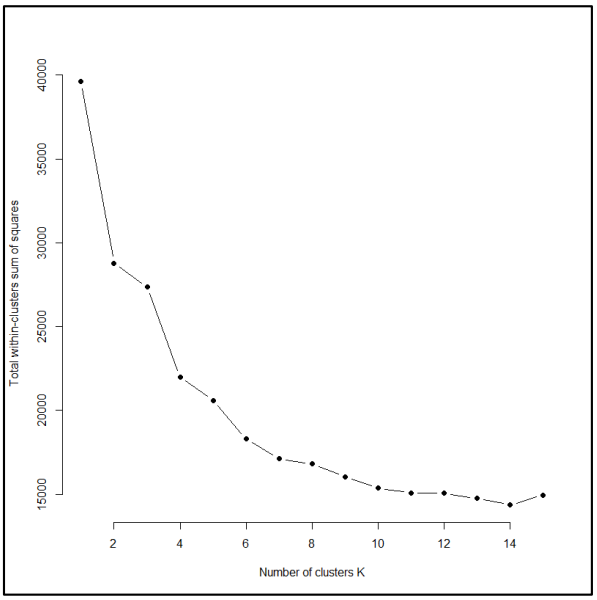

Figura 1 – Representação gráfica da técnica de *Elbow* a partir de valores da base de dados analisada neste artigo.

4 MATERIAIS E MÉTODOS

A metodologia utilizada durante o processo de desenvolvimento do artigo foi definida considerando as etapas de coleta de dados, pré-processamento, e análise de sentimentos.

Como método proposto, foram utilizadas técnicas de análise de sentimentos e agrupamento de dados por meio da linguagem R e APIs de conexão com o Twitter. A aplicação dos processos de coleta de dados, pré-processamento e análise de sentimentos é descrita a seguir.

4.1 Coleta de Dados

A etapa consiste no processo de coleta dos dados, tendo como resultado a construção de uma base de dados para análise. Para este artigo, os dados foram obtidos a partir do Twitter, por meio do uso da API disponibilizada pela própria rede social, para conexão e coleta dos *tweets*.

Foram considerados válidos para construção da base de dados os *tweets* realizados no período de 1º de agosto a 6 de outubro de 2018. A base de dados construída possui 88 atributos e 4.608 instâncias, que descrevem características associadas aos *tweets*. Para maior entendimento, esses dados podem ser descritos a partir de duas categorias: os dados referentes a informações básicas sobre o usuário (nome, descrição do perfil, total de seguidores, total de seguidos, localização, etc.) e os dados referentes aos *tweets* (conteúdo textual da mensagem publicada, data da publicação, mensagem re*tweet*ada ou gerada diretamente pelo perfil do usuário, total de caracteres gerados por *tweet*, quantidade de curtidas, etc.).

Especificamente para o objetivo deste artigo, que visa identificar a polaridade dos *tweets* realizados pelos candidatos considerando as avaliações em linha temporal, foi feito o uso dos atributos de conteúdo textual dos *tweets*, data de postagem do *tweet* e o nome do usuário na rede social. Outros atributos foram desconsiderados por não se tratarem de conteúdos fundamentais para a aplicação de análise de sentimentos descrita neste artigo.

4.2 Pré-Processamento

A etapa de pré-processamento do texto se faz necessária para as aplicações de métodos léxicos durante a análise de sentimentos **[20]**. O pré-processamento consiste em realizar procedimentos para o tratamento e limpeza nos dados que serão analisados, removendo redundâncias e dados considerados desnecessários, com o objetivo de otimizar a etapa de análise de sentimentos.

Inicialmente a base de dados construída na etapa anterior foi transformada em um subconjunto com apenas os *tweets* postados pelos usuários dos candidatos. Os dados referentes a re*tweets* foram desconsiderados da análise de sentimento, visto que não se tratam de posts realizados diretamente pelo candidato. Como processo de tratamento dos dados no conteúdo dos *tweets*, foram realizadas as seguintes atividades:

- **Substituição de caracteres maiúsculos por minúsculos:** considerando o padrão utilizado nos dicionários léxicos utilizados, o tratamento foi realizado para evitar inconsistências e permitir uma melhor visualização de indicadores que poderiam ser criados a partir dessa base de dados, como um *word cloud*.
- **Remoção de URLs:** links internos ou externos não foram considerados para o desenvolvimento deste artigo, visto que não existe classificação léxica para este tipo de dado.
- **Remoção de conteúdos visuais:** imagens, vídeos ou gifs foram removidos, visto que o estudo se trata apenas de uma avaliação de polaridade textual. Emojis também foram desconsiderados, por possuírem o mesmo sistema linguístico das palavras disponíveis.

- **Remoção de stopwords:** foram desconsideradas palavras não relevantes para a avaliação do algoritmo classificador, considerando o contexto e particularidades associadas ao processo de análise de sentimentos. Casos como preposições, artigos e conectores, que não agregam sentido específico à frase.
- **Remoção de acentos e caracteres especiais:** exclusão da acentuação e de caracteres considerados não-alfabéticos, tanto por, em alguns casos, servirem para manifestar fatores como a cultura dos usuários ao se expressar quanto para minimizar erros ortográficos ou de digitação.

4.3 Análise de Sentimentos (AS)

Nesta etapa, os dados já pré-processados foram submetidos a métodos de análise de sentimentos para extração de conhecimento. De maneira geral, o processo de AS consiste em encontrar classes (ou polaridades) para os dados, nesse caso classificando-os como positivo ou negativo **[21]**.

Na primeira parte desta seção, foram descritos os processos de análise para encontrar o sentimento dos *tweets*. Por se tratar de uma análise realizada em dados na língua portuguesa, o pacote LexiconPT **[22]**, desenvolvido e mantido por Sillas Gonzaga, foi escolhido como opção para classificar a polaridade das palavras contidas nos *tweets*.

Os dicionários léxicos selecionados no pacote foram o SentiLex-PT02 e o OpLexicon V3.0, visto que o outro dicionário disponível é o OpLexicon V2.1, sendo preterido pela sua versão mais atualizada. Em casos de palavras em que o seu conteúdo esteja em apenas um dicionário, este valor será considerado. No caso da palavra estar contida em ambos os pacotes, prevalecerá o pacote com maior presença no *tweet* em questão. Palavras não disponíveis na base de dados ou de polaridade neutra em ambos os dicionários foram desconsideradas da etapa. Cada palavra é submetida a ambos os dicionários, recebendo o valor 1 caso possua polaridade positiva e -1 caso negativa. O resultado é representado por meio de uma matriz de *tweets* e duas instâncias para os valores totais do SentiLex-PT02 e OpLexicon. Este método é descrito visualmente na Figura 2.

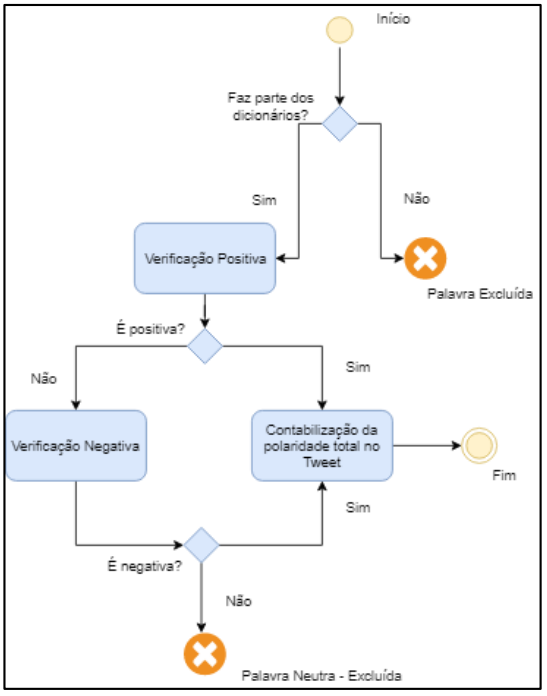

Figura 2 – Verificação da polaridade para as palavras consideradas.

Para possibilitar a classificação dos textos, cada *tweet* teve seu conteúdo fracionado, criando *n* linhas para o *tweet* baseado em n palavras disponíveis em seu conteúdo, resultando em uma tabela com uma linha por palavra e o ID do seu *tweet*. Os termos são avaliados de acordo com a sua disponibilidade nos dicionários léxicos. A Figura 3 mostra a polarização das palavras de acordo com os dicionários.

comment_id	term	polarity	lex_polarity
<int>	<chr>	<int>	<dbl>
1	uberlandia	NA	NA
1	recebeu	NA	NA
1	ciro	NA	NA
1	carinho	NA	1
1	falou	NA	NA
1	projeto	NA	NA
1	nacional	0	NA
1	publico	0	NA
1	presente	0	1
1	noite	NA	NA
1	sexta	0	NA
1	feira	NA	NA
1	memoravel	1	1

Figura 3 – Polarização das palavras de acordo com os dicionários.

Para atingir o resultado de polaridade de cada *tweet*, foi realizada a equação *X = (Y) – (Z)*, sendo X o valor total de polaridade do *tweet*; Y, o valor total de palavras positivas; e Z, o valor total de palavras negativas.

DOI: 1025286/repa.v5i1.1173

Após concluir a etapa de polarização dos *tweets*, foram realizados experimentos utilizando técnicas de agrupamento de dados, com o objetivo de identificar comportamentos ainda não apontados por meio do processo de polarização. A base de dados gerada ao final da etapa de pré-processamento foi submetida ao método de TF-IDF. Desta forma, cada *tweet* é um vetor cujo comprimento é o tamanho do vocabulário restante e os seus componentes são avaliados de acordo com a frequência da palavra na sentença e sua significância no corpus.

Os dados obtidos do TF-IDF foram submetidos a técnicas para verificação da distância euclidiana, com o objetivo de identificar o quão distante está um dado de outro. Os resultados da distância euclidiana foram aplicados ao método *elbow*, atigindo o valor de 6 grupos necessários para aplicação ao k-means, com o total de 3.467 instâncias para agrupamento.

5 RESULTADOS

Nesta seção, são expostos os resultados da análise de sentimentos de cada candidato, utilizando a base de dados tratada anteriormente. Os resultados obtidos foram descritos em duas subseções. A primeira, considerando os resultados de polaridade dos tweets a partir de um gráfico temporal, para identificação de comportamento de acordo com as pesquisas do IBOPE. Na segunda, os dados submetidos ao algoritmo K-Means foram representados através de um espaço bidimensional, considerando a distância euclidiana de cada tweet ao centro do grupo que pertence.

5.1 Polaridade dos *tweets* durante o período coletado

Após concluir a construção das polaridades na base de dados, foi realizada a análise nos dados coletados e tratados. Para a construção dos gráficos em linha temporal, foi necessário unir a base de dados tratada nas etapas anteriores a uma nova base de dados com as datas de realização das pesquisas de intenção de votos pelo IBOPE.

As Figuras representadas durante esta subseção apresentam o valor total do sentimento, considerando a polaridade positiva ou negativa. As barras em verde representam os valores totais do dia referentes aos *tweets* de polaridade positiva. As barras vermelhas são referentes aos *tweets* de polaridade negativa. Os pontos em preto são referentes as datas das pesquisas de intenção de voto. O eixo x apresenta as datas de postagem dos *tweets* do dia 1º de agosto até o dia 6 de outubro deste ano.

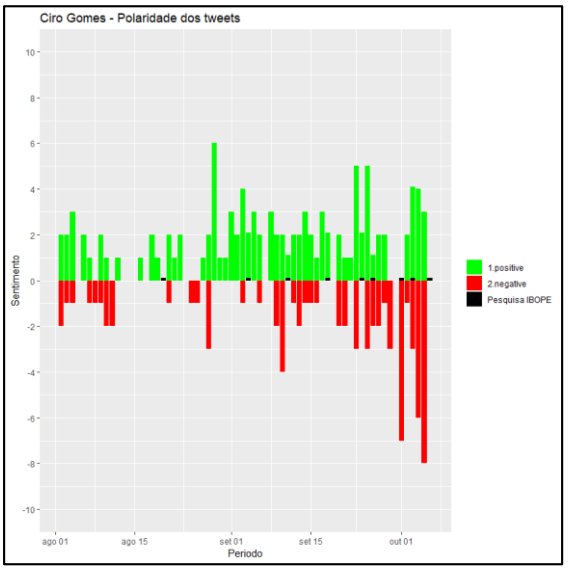

Figura 4 – Resultados de polaridade nos *tweets* realizados pelo candidato Ciro Gomes.

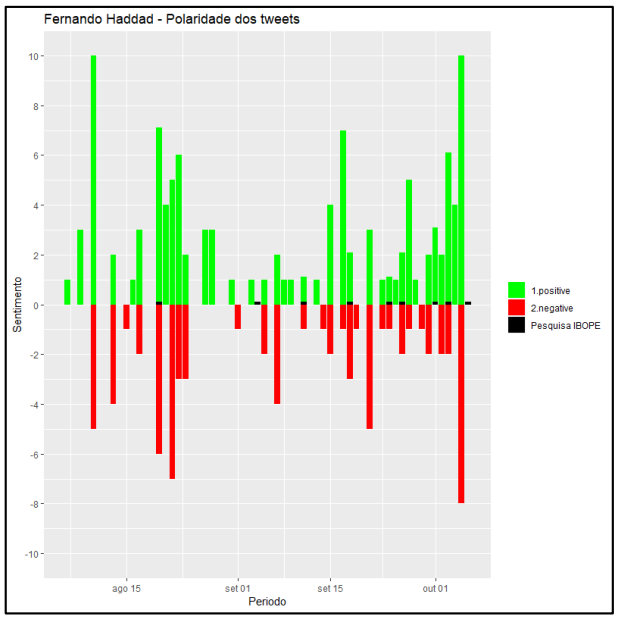

Figura 5 – Resultados de polaridade nos *tweets* realizados pelo candidato Fernando Haddad.

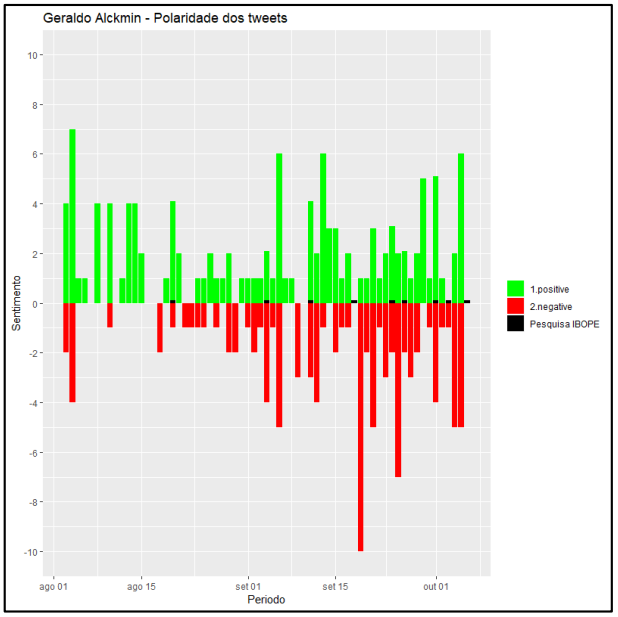

Figura 6 – Resultados de polaridade nos *tweets* realizados pelo candidato Geraldo Alckmin.

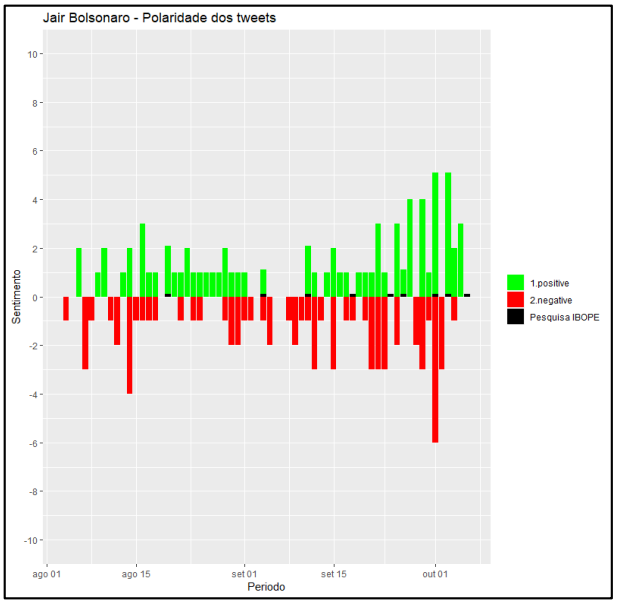

Figura 7 – Resultados de polaridade nos *tweets* realizados pelo candidato Jair Bolsonaro.

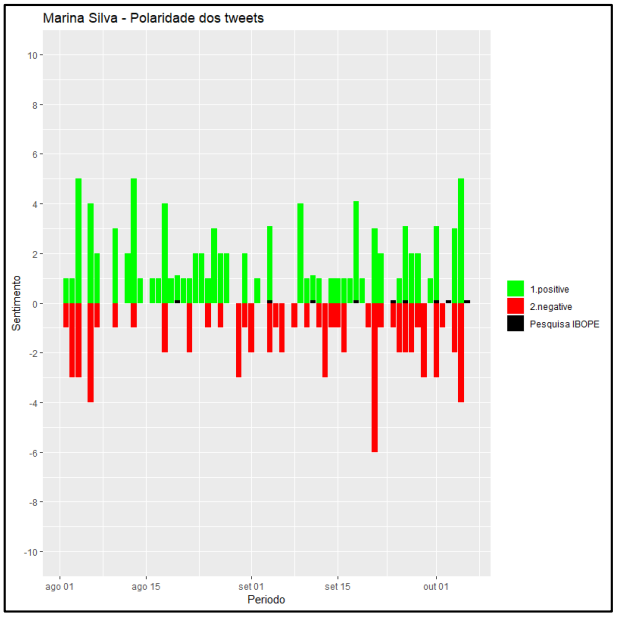

Figura 8 – Resultados de polaridade nos *tweets* realizados pela candidata Marina Silva.

Observando os resultados apresentados nas Figuras 4, 5, 6, 7 e 8, é perceptível que existe uma tendência no crescimento dos *tweets* de polaridade negativa com o passar do tempo para todos os candidatos, comportamento que pode estar atrelado a variáveis como o crescimento das campanhas eleitorais e o acirramento nas disputas por voto conforme a data para o primeiro turno se aproxima.

Também é identificada uma tendência à publicação de mensagens mais positivas no dia posterior à divulgação dos resultados das pesquisas. Essa tendência, no entanto, não é identificada nos *tweets* realizados pelo candidato Geraldo Alckmin, que apresentou *tweets* com o sentimento de polaridade negativa no dia subsequente à divulgação dos resultados em cinco das nove pesquisas avaliadas.

Também é relevante destacar o comportamento do perfil de Fernando Haddad quando houve a divulgação da primeira pesquisa de intenção de votos, em 20 de agosto: o perfil do então candidato a vice-presidente apresentou disparidade tanto positivamente quanto negativamente quando comparado aos perfis dos demais candidatos, no que diz respeito à avaliação das polaridades dos *tweets*. Tal fato pode ter relação direta com o resultado da pesquisa, uma vez que a chapa de Fernando Haddad, então candidato ao cargo de vice-presidente, liderava a disputa presidencial, com 37% das intenções de voto - 19 pontos percentuais a mais do que o segundo colocado, Jair Bolsonaro **[23]**.

DOI: 1025286/repa.v5i1.1173

5.2 Aplicação do K-Means

A Figura 9, apresentada nesta subseção, mostra o resultado obtido a partir do uso do algoritmo de agrupamento K-means. A base de dados utilizada possui 3.467 instâncias, considerando o valor da distância euclidiana de cada *tweet*, a partir dos dados de TF-IDF gerados durante a etapa de análise de sentimento. Os dados foram representados na Figura 9 considerando cinco cores, sendo distribuídas da seguinte forma: cor vermelha para o candidato Ciro Gomes; preta para o candidato Fernando Haddad; azul para o candidato Geraldo Alckmin; verde para o candidato Jair Bolsonaro; e amarelo para a candidata Marina Silva.

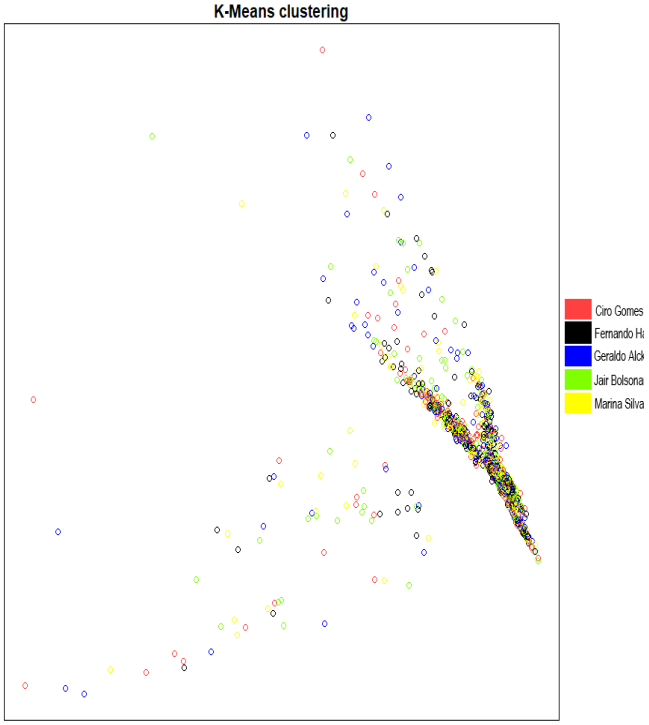

Figura 9 – Resultados do agrupamento de *tweets* através do algoritmo de K-Means.

A partir da análise do resultado gerado por meio do K-Means, fica evidente o padrão de similaridade entre os dados representados na Figura 9. Considerando a Figura 10, os valores de TF-IDF utilizados no agrupamento sugerem um modelo semelhante no que se refere ao valor médio no grau de importância das palavras utilizadas em suas mensagens.

Candidatos	Avarage of TF_IDF	Max of TF_IDF	Min of TF_IDF
Alckmin	0,8417	1,8184	0,3194
Ciro	0,8001	3,5288	0,2102
Haddad	0,8170	1,5109	0,2690
Bolsonaro	0,8422	3,4039	0,3079
Marina	0,8267	1,7272	0,3053

Figura 10 – Matriz de resultados de TF-IDF por candidato.

O motivo para o agrupamento possuir esse destaque se deve ao fato de que termos fundamentais para todos os candidatos são utilizados em boa parte dos *tweets*, conforme a Figura 11, uma vez que todos eles discutem sobre temas comuns durante o processo eleitoral, como a própria eleição, ou pautas de saúde, segurança pública, economia, social e educação. Fica evidente que apesar de diferenças ideológicas, os resultados apresentados sugerem uma maior similaridade entre os conteúdos postados pelos candidatos, no que tange a similaridade sintática.

word	freq
Brasil	718
Vamos	532
Pais	341
Governo	248
Lula	243
Hoje	225
Presidente	218
Candidato	205
Povo	181

Figura 11 – Seleção de 9 termos com maior frequência dentro da base de dados.

Ainda assim, existem *outliers*, representados em sua maioria, pelos candidatos Ciro Gomes, Geraldo Alckmin e Jair Bolsonaro. *Tweets* considerados como outliers são, em sua maioria, mensagens compostas a partir de quantidade reduzida de palavras e de termos não frequentemente utilizados. Para os dados nos quais ocorreram este comportamento, palavras como

gírias ou nomes de cidades não encontradas no corpus classificaram esses dados como diferentes do padrão identificado na Figura 9.

6 CONCLUSÕES

O artigo teve como proposta aplicar técnicas de análise de sentimentos e agrupamento de dados para a avaliação dos perfis do Twitter dos candidatos à presidência do Brasil em 2018. Para isso, foram escolhidos os cinco candidatos mais bem posicionados nas pesquisas do IBOPE, com o objetivo de identificar possíveis similaridades ou dados que caracterizassem informações relevantes a respeito do uso das redes sociais por parte destes candidatos.

Para tanto, foram utilizadas as etapas de coleta de dados para geração da base de dados e pré-processamento para realizar o tratamento de informações em linguagem natural. Após a passagem pelo pré-processamento, os dados foram submetidos a técnicas de análise de sentimentos, como a aplicação de dicionários léxicos para definição da polaridade de cada um dos *tweets* e métodos como o TF-IDF, para identificação da frequência dos termos em um determinado corpus.

Os resultados descritos no artigo permitiram identificar comportamentos referentes ao grau de positividade e negatividade transmitidos pelos *tweets*, utilizando como referência as datas de pesquisa de intenção de voto. Na segunda etapa dos resultados, os valores pré-processados foram submetidos ao algoritmo de agrupamento K-Means, representando dados de grau de similaridade e frequência dos termos através da distância euclidiana gerada com os dados do TF-IDF de cada *tweet*. Os resultados sugerem um grau de similaridade entre os candidatos, considerando dados de valor médio do TF-IDF para cada um dos avaliados.

Em trabalhos futuros, a aplicação de novos algoritmos diferentes do k-means poderia ser considerada, visto que, apesar deste ser um dos mais populares algoritmos para avaliação de conteúdos não estruturados, nem sempre ele é a melhor opção, uma vez que existem algoritmos mais específicos para a análise de sentimentos. O método da análise de sentimentos também poderia ser aplicado para a descoberta de novos sentimentos transmitidos pelos *tweets*, considerando não apenas a polaridade do seu conteúdo, mas também as demais emoções transmitidas.

REFERÊNCIAS

[1] Lei nº 9.504/1997, artigo 57-C. Disponível em:http://www2.camara.leg.br/legin/fed/lei/2017/lei-13488-6-outubro-2017-785551-publicacaooriginal-153918-pl.html

[2] Lei n° 13.488/2017, disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9504.htm

[3] Pew Research Group, 2018. Disponível em: http://assets.pewresearch.org/wp-content/uploads/sites/2/2018/01/09131309/Publics-Globally-Want-Unbiased-News-Coverage-but-Are-Divided-on-Whether-Their-News-Media-Deliver_Full-Report-and-Topline-UPDATED.pdf

[4] MARQUES, Francisco Paulo Jamil Almeida. **Debates políticos na internet: a perspectiva da conversação civil**. Opinião Pública, Campinas, vol. 12, nº 1, Abril/Maio, pp, 164-187, 2006.

[5] HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Minning: Concepts and Techniques**. 2011.

[6] TUMASJAN, Andranik. **Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment**. 2011.

[7] ATTUX, Rafael. **Predição dos resultados das eleições 2014 para presidente do Brasil usando dados do Twitter**, Universidade Federal de Uberlândia – UFU, 2017.

[8] AGUIAR, Sonia. **Redes sociais e tecnologias digitais de informação e comunicação: relatório final de pesquisa**. Rio de Janeiro: Nupef/RITs, 2006.

[9] TORRES, C. **A bíblia do marketing digital**. São Paulo: Editora Novatec, 2009.

[10] RECUERO, R. **Redes Sociais na Internet**. Porto Alegre: Sulina, 2009. (Coleção Cibercultura).

[11] NARAYANAN, Ramanathan; LIU, Bing; CHOUDHARY, Alok. **Sentiment analysis of conditional sentences**. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of

DOI: 1025286/repa.v5i1.1173

the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. [S.l.], 2009. p. 180–189.

[12] GONÇALVES, Pollyanna de Oliveira. **Um benchmark para comparação de métodos para análise de sentimentos**. UFMG, 2015.

[13] REZENDE, Solange Oliveira. **Sistemas Inteligentes - Fundamentos e Aplicações**. Barueri, SP, Brasil: Manole, 2003.

[14] DIXON, Mark. **Na Overview of Document Mining Technology**. [S.l.: s.n]. 1997.

[15] RICARDO, Baeza-Yates *et al*. **Modern information retrieval**. [S.l.]: Pearson Education India, 1999.

[16] JAIN, Anil K. **Data clustering: 50 years beyond K-means**, In Pattern Recognition Letters, Volume 31, Issue 8, 1 June 2010, Pages 651-666.

[17] PIMENTEL, E. P., FRANÇA, V. F. De, Omar, N. **A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização**, ITA, São José dos Campos, SP, Brasil, 2003.

[18] HENRIQUES, N. D., N. **Um breve estudo sobre o algoritmo K-means**. Departamento de Matemática, Universidade de Coimbra, Portugal, 2016.

[19] CARLANTONIO, L. M. Di. **Novas metodologias para clusterização de dados**. Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia, UFRJ, RJ, Brasil, 2001.

[20] REZENDE, Solange Oliveira. **Sistemas Inteligentes-Fundamentos e Aplicações**. Barueri, SP, Brasil: Manole, 2003. ISBN 85-204-1683-7.

[21] DUARTE, E. S. **Sentiment analysis on Twitter for the Portuguese language**. (Doutorado) — Universidade Nova de Lisboa, 2013.

[22] **Biblioteca LexiconPT**. Disponível em: https://cran.r.project.org/web/packages/lexiconPT/index.html

[23] IBOPE - **Resultados de pesquisa após o registro dos candidatos a eleição presidencial.** Disponível em http://agenciabrasil.ebc.com.br/politica/noticia/2018-08/ibope-divulga-1a-pesquisa-apos-registro-de-candidatos-presidente.