USO DE DADOS DE REDES SOCIAIS COMO INDICADOR DO COMPORTAMENTO SOCIAL DURANTE A EPIDEMIA COVID-19: ANÁLISE PRELIMINAR DO COMPORTAMENTO POLÍTICO SOBRE AS ELEIÇÕES NO BRASIL

Palavras-Chave:BIG-DA, INDICADORES SOCIAIS, ANÁLISE DE SENTIMENTOS

Autores(as):
DÉCIO MIRANDA FILHO – IMECC - UNICAMP
Prof. Dr. ALEXANDRE GORI MAIA (orientador), IE - UNICAMP

INTRODUÇÃO:

A Pandemia de Covid-19 representou um momento crucial na história da humanidade e, à medida que se espalhou globalmente, ganhou destaque nos noticiários, reportagens e, especialmente, nas redes sociais. Diante desse cenário, este projeto tem como objetivo utilizar uma das maiores plataformas de divulgação social, o Twitter, que hoje é uma das principais fontes de dados para extrair informações precisas e relevantes sobre diversos temas, a fim de compreender as opiniões e comportamentos dos seus usuários.

A análise preliminar do uso de dados do Twitter para construção de um classificador sobre as eleições no Brasil agiu conforme a diretriz do orientador(até onde a pesquisa continuou). Essa classificação dos tweets relacionados ao comportamento dos usuários do Twitter que comentaram sobre o sentimento político dos dois principais candidatos na primeira etapa da eleição presidencial da República Federativa do Brasil tratou em utilizar dados sobre os dois candidatos à presidência com maiores intenção de votos: Luiz Inácio Lula da Silva e Jair Messias Bolsonaro. O objetivo é testar a eficácia do algoritmo de classificação e compreender a relação entre o sentimento político expresso nos tweets e os votos por regiões onde foram publicados. A ideia seria que a partir do resultado desses dados testar sobre dados da COVID, mas que não foi possível devido à interrupção da pesquisa conforme resultados de Gori Maia *et. al* (2023).

METODOLOGIA:

O projeto teve como regulamentos indicados pelo orientador que a priori construir-se-ia um classificador de sentimento político baseado em palavras contextuais e relacioná-lo com os votos por

região. Nessa etapa inicial, o objetivo proposto foi a estruturação desse classificador de sentimento político objetivando à validação e ajuste.

Os dados do Twitter foram fornecidos pelo orientador e, junto com os conjuntos de dados eleitorais e de mesorregiões, foram tratados para reduzir a quantidade de observações. O classificador utilizou termos positivos e negativos associados a cada candidato (Lula e Bolsonaro) para categorizar os tweets. Após o processamento dos dados, o classificador conseguiu classificar 52.994 tweets relacionados a Lula e 30.534 tweets relacionados a Bolsonaro (excluindo tweets neutros). Os tweets foram submetidos a tratamento para remover palavras irrelevantes e acentos.Além disso, foram utilizados mais dois conjuntos de dados auxiliares para obter informações sobre as mesorregiões e dados eleitorais, a fim de calcular a proporção de votos por mesorregião.

Em suma, o desenvolvimento do e teste do classificador de sentimento político dos tweets, e correlação com os votos por região, culminando na elaboração de uma tabela com a média agregada do classificador e a proporção de votos por mesorregião.

Tabela 1. – Número de Observações de Tweets após cada Etapa

Etapas	Número de Observações	
	Lula	Bolsonaro
Dados Brutos Raspados do Tweeter	1190013	
Remoção de Duplicados	847156	
Tratamento Strings e captura dos dados com localização adequada*	220033	
Passagem pelo Classificador e que contivessem o respectivo candidato**	52994	30534
Agregação por Usuário e Local	2657	2221
Agregação por Mesorregião***	137	137

**Em tratamento por strings foi aplicada transformações de strings passadas para minúscula, retirada de acentos e outros caracteres.*

*** Foi aplicado o classificador de palavras contextuais, filtrando os tweets não classificados e feita a captura dos locais dos tweets.*

*** Agregado pelas 137 mesorregiões do Brasil*

Fonte: Elaborada pelo autor

Com isso, o classificador operou com base em palavras contextualmente associadas a conotações positivas ou negativas. Algumas dessas palavras foram extraídas do próprio conjunto de dados de tweets para melhorar seu desempenho. O classificador consiste em itens que geralmente remetem a um contexto de sentimento específico para cada candidato político. Para o candidato Lula, foram utilizadas palavras positivas como "desigual", "fome" e "amo", e negativas como "corrupção", "ladrão" e "quadrilha". Já para o candidato Bolsonaro, foram utilizados termos como "mito", "deus" e "família", e negativos como "gado", "vacina" e "milícia". A Tabela 2 em anexo contém os termos utilizados para cada um dos candidatos.

Além disso, o classificador funciona com um algoritmo desenvolvido em linguagem Python. Ele verifica se alguma das palavras positivas está presente no tweet e atribui +1 para positivo e -1 para negativo caso algum dos termos negativos seja identificado no tweet.

Após a breve explicação do algoritmo, a base de dados tratada com 220.033 observações foi submetida ao classificador, permitindo a classificação de 52.994 tweets relacionados ao candidato Lula e 30.534 tweets relacionados ao candidato Bolsonaro (considerando apenas tweets não neutros, ou seja, diferentes de 0). É importante mencionar que, durante esse processo, todas as palavras de um tweet foram processadas por stop-words, ou seja, foram excluídos pronomes e artigos que não têm relevância para o processamento da linguagem natural, e também foram retirados os acentos.

Tabela 2 – Palavras e Termos Contextuais utilizadas para Classificação do Sentimento Político

Termos	Lula	Bolsonaro
Positivos	inocente, bom, pobre, fome, miser, parab, grand, excel, desigual, bem, ilustre, carid, hero, trab, amo, desigu, maior, crescim, melhor, burgues, trab, lider, amo, votam lula, desempre, fome, inflaç, lindo, milic, antipet, :), (:, :D, :-), ;), (;, 😉, 😄, 😀, 😭, 😍, 😘, ♥, 🟢, ▶, 😚, campeao, educ, lula presidente, 👏, votar lula, ▧, reeleito, corint, australia, nova zelandia, vence, ♥, lula primeiro, ganhar, lula importante	molusco, inocente, bom, parab, grand, excel, bem, ilustre, carid, hero, trab, amo, maior, crescim, melhor, trab, lider, votam bolso, lindo, antipet, :), (:, :D, :-), ;), (;, 😉, 😄, 😀, 😭, 😍, 😘, ♥, 😚, campeao, mito, imbrox, imbroch, arma, imunidade de rebanho, meu presidente, capitao, deus, patria, familia, acima de, naçao, china, cloroquina, ivermec, cuba, venez, confirma bolsonaro, 👏, reeleito, voto impresso, vence, bolsonaro presidente, ganhar, deus acima
Negativos	mesmo, roub, corrup, mensa, petro, perdeu, perd, ladr, ladrao, comuni, esquerd, cuba, venez, idio, imbec, mald, band, imprest, covard, pior, mariel, molusco, chefe, quadrilha, empr, empobr, fraude, endiv, milit, destr, mamata, mortadela, regul, :(,):, >_<, > _<*, (>_<), ;(, 💔, 😂, 🤣, 😈, ☹, 😒, 😑, 😭, 😕, 😩, 💩, 🤢, doente, imundo, demo, cachac, repugnante, maligno, doente, desastr, condenado, pcc, cabo eleitoral, crime organizado, matou, pinguco, burro, mentir, marcola, cego, ladroagem, mensalao, canalha, maldito, 👌, cadeia, 💩, fraude, vigarista, ptralha, lixo	queimadas, bozo, gado, mesmo, inflaç, fome, genoc, vacina, roub, corrup, mensa, petro, perdeu, perd, ladr, ladrao, idio, imbec, mald, band, imprest, covard, pior, chefe, quadrilha, empr, empobr, fraude, endiv, milit, destr, mamata, mortadela, regul, :(,):, >_<, > _<*, (>_<), ;(, 💔, 😂, 🤣, 😈, ☹, 😒, 😑, 😭, 😕, 😩, 💩, 🤢, doente, imundo, demo, diab, repugnante, maligno, nojento, doente, desastr, gado, bozo, fora, cabo eleitorial, odeia pobre, crime organizado, matou, rachad, odeio, sufocando, burro, bater mulher, mentir, milicia, micheque, sociopata, cego, falec, covid, orcamento secreto, canalha, maldito, 👌, odeia pobre, cadeia, 💩, fraude, vigarista, lixo

Além disso, foram utilizados mais dois conjuntos de dados auxiliares. Um deles continha informações sobre as mesorregiões que abrangem todos os municípios do Brasil, e os dados foram obtidos a partir de <https://dadosabertos.tse.jus.br/dataset/resultados-2022>. O outro conjunto continha dados eleitorais, e foi feita uma filtragem dos votos por candidato para criar a proporção percentual de votos absolutos por município, considerando apenas os dois principais candidatos.

Após essa etapa, os dois conjuntos de dados, um contendo os votos e o outro com informações sobre as mesorregiões, foram combinados em um único conjunto, com o objetivo de obter a proporção de votos por mesorregião. A união de todos esses dados em uma única tabela, agrupada por usuários e por cada mesorregião, resultou em uma tabela com 137 observações. Essa tabela continha o

resultado da média agregada do classificador, bem como a proporção média agregada de votos para ambos os candidatos à presidência da república.

RESULTADOS E DISCUSSÃO:

Assim, após realizar todas as etapas de obtenção, correção, análise e classificação dos dados, foi possível identificar uma correlação entre a média da proporção percentual de votos de cada mesorregião e a média do sentimento associado a cada candidato: um r-Pearson de 27,93% para o candidato Lula e um r-Pearson de 16,09% para o candidato Bolsonaro.Outra métrica utilizada em vez da média da proporção percentual de votos foi a classificação em resultados binários, onde o candidato que recebeu a maior proporção de votos percentuais em um município recebeu o valor 1 e, caso contrário, recebeu o valor 0. Com essa métrica, obteve-se um r-Pearson de 29,44% para Lula e um r-Pearson de 13,92% para Bolsonaro. Uma tentativa adicional de aprimorar o classificador foi adicionar ao conjunto de palavras contextuais o dicionário lexicon da biblioteca nltk. Esse dicionário consiste em uma lista de palavras genéricas, desprovidas de um sentido político específico. No entanto, a inclusão do dicionário na análise piorou os resultados, evidenciando que um classificador com poucas palavras contextuais específicas do cenário brasileiro pode obter resultados bastante significativos.

CONCLUSÕES:

Considerando que o projeto tinha como meta inicial validar um classificador sobre dados políticos sobre as eleições no Brasil, as conclusões até o momento da interrupção da pesquisa foi que o classificador apresentou bons indicadores de correlação, sendo necessário aprimorar os termos de palavras contextuais para um melhor resultado. Isso demonstra que numa hipotética fase de testes com termos contextuais ligados à COVID-19 haverá uma melhor identificação dos termos estratégicos para estudo de correlação entre as variáveis. Mesmo assim, em termos de utilidade, o classificador e a análise de correlação realizada demonstrou resultados promissores e parece ser capaz de ser aprimorada e adaptada para diferentes contextos de análise de sentimentos. Ratificando, o classificador somente com as palavras contextuais apresentou melhores indicativos que o uso de um pacote com termos genéricos.

BIBLIOGRAFIA

Gori Maia, A., Martinez, J.D.M., Marteleto, L.J. et al. **Can the Content of Social Networks Explain Epidemic Outbreaks?**. Popul Res Policy Rev 42, 9 (2023). Disponível em . Acesso em 4 març. 2023.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56). SciPy. Disponível em Acessado em 25 fev. 2023

NumPy 1.20.1. (2021). NumPy: A fundamental package for scientific computing with Python. Disponível em Acessado em 25 fev. 2023

PYTHON SOFTWARE FOUNDATION. Welcome to Python.org. Site de Internet. Disponível em: . Acesso em: 24 set. 2022.

Unidecode 1.2.0. (2018). Unidecode: ASCII transliterations of Unicode text. Acessado em https://pypi.org/project/Unidecode/1.2.0/ Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc. Disponível em https://www.nltk.org/book/. Acessado em 25 fev. 2023