Sistema para Coleta e Tratamento Textos Brasileiros sobre Polarização Política

Paulo Costa[1], **Luciano Gallegos**[1]

[1]Universidade de Fortaleza – UNIFOR
Fortaleza – CE – Brasil

{paulo7676, luciano.gallegos}@gmail.br

Abstract. *Popular political polarization stands out when there is a clear party lines division regarding political, public and public subjects, where one side sees the other as a threat. This type of polarization can be studied through texts shared on social networks, such as on Facebook and Twitter. In many occasions, researchers and interested people collect texts from these social networks, but the process of collecting and processing these texts are not shown. In this article, we show the development of a system aiming at collecting and treating texts involving the popular political polarization subject from Twitter, capable of including tweets' localization and the sentiment analysis of their texts.*

Resumo. *A polarização política popular caracteriza-se quando há uma clara divisão de linhas partidárias em relação a questões políticas, políticas governamentais e figuras públicas, onde um dos lados observa o outro como ameaça. Este tipo de polarização pode ser estudado por meio de textos compartilhados em redes sociais, como no Facebook e no Twitter. Muitas vezes, pesquisadores e interessados coletam textos destas redes sociais, mas não mostram como desenvolver o processo de coleta e tratamento destes textos. Neste artigo, mostramos o desenvolvimento de um sistema com o objetivo de coletar e tratar textos envolvendo o tema da polarização política popular do Twitter, capaz de incluir a localização dos tweets e a análise de sentimento destes textos.*

1. Introdução

A polarização política é um termo que vem sendo bastante abordado em meios de comunicação e entre as pessoas, e pode ser abordada de diferentes forma. Uma destas formas, estudada por sociólogos e cientistas políticos, chama-se polarização política ”popular”e caracteriza-se por haver uma clara divisão de linhas partidárias em relação a questões políticas, políticas governamentais e figuras públicas, onde um dos lados observa o outro como ameaça [Baldassarri and Gelman 2008].

Hoje, a polarização política popular pode ser estudada a partir de textos compartilhados em redes sociais, como o Facebook e o Twitter, onde estados afetivos e reflexões cognitivas, e tendências comportamentais podem ser verificados e correlacionados à polarização política em indivíduos [Moslehm et al. 2021]. Exemplo deste tipo de abordagem pode ser encontrado em [Jiang et al. 2020], onde foi desenvolvida uma análise da polarização política nos Estados Unidos entre janeiro a março de 2020, ano de eleições presidenciais e início da pandemia do coronavírus naquele país. Nessas análises, ficou

evidente a existência de polarizações maiores em tweets pertencentes a categorias conspiracionistas, inclinações políticas, e saúde pública. Já em [Kamienski et al. 2021], foi analisanda a polarização política no Brasil entre Março e Junho de 2020. Os autores identificaram categorias de tweets envolvendo maiores polarizações em cada um desses dias, assim como a quantidade de usuários, retweets, e a inclinação e posicionamento políticos. Obtivemos, por meio destes artigos, informações sobre a API utilizada para a extração de tweets e acesso a base de dados mas, por outro lado, não é mostrado como desenvolver computacionalmente o processo de coleta e tratamento destes tweets.

Neste artigo, mostraremos o desenvolvimento de um sistema para coleta e tratamento de textos envolvendo polarização política popular por meio de tweets do Twitter, assim como o acoplamento de um analisador léxico para determinar o estado afetivo nestes textos. Neste mesmo sistema, a origem geográfica destes tweets são agregadas. Este artigo está organizado da seguinte forma: na Seção 2, fazemos a fundamentação teórica e descrevemos os passos necessários para o desenvolvimento do sistema proposto, e na Seção 3 são descritos os resultados obtidos da coleta de tweets com localização agregada. Por último, na Seção 4, fazemos nossa reflexões sobre o trabalho descrito neste artigo.

2. Fundamentação Teórica e Metodologia do Trabalho

O sistema desenvolvido e descrito neste artigo é composto por 4 componentes internos: coleta de tweets, hidratação, limpeza e tradução, e localização e análise de sentimentos. Existem mais 3 componentes externos: Twitter (rede social de onde extraímos os tweets), e uma Banco de Dados (armazenamento dos tweets coletados e processados). Este sistema foi programado em linguagem Python 3.8.3, e seus componentes podem ser visualizados na Figura 1.

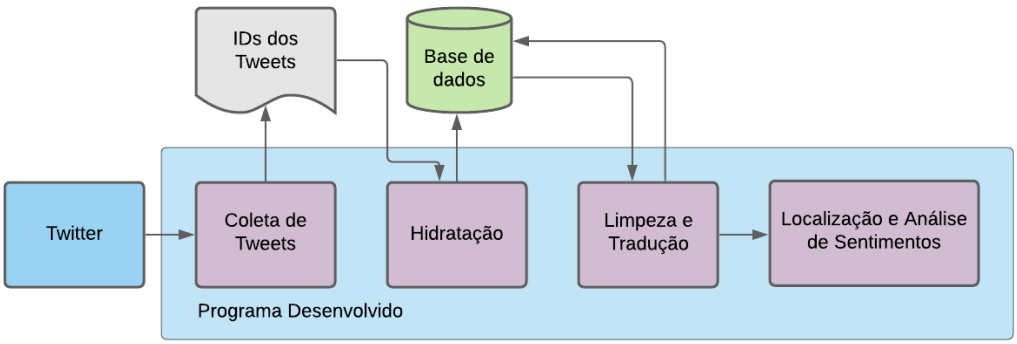

Figura 1. Sistema para Coleta e Tratamento de Tweets Brasileiros sobre Polarização Política.

A **coleta de tweets** é iniciada pela seleção de tópicos de interesse, por meio do Twitter Trending Archive[1]. Ele lista os tópicos mais frequentes do Twitter para dias e horários específicos, por país, e são atualizados a cada 30 minutos [Aiello et al. 2013]. Selecionamos empiricamente os tópicos relacionados à polarização política popular no Brasil entre 1^o de fevereiro e 30 de julho de 2020, o que coincide com o início da pandemia do Novo Coronavírus (Covid-19) no Brasil[2].

[1] https://archive.twitter-trending.com/
[2] https://coronavirus.saude.gov.br/linha-do-tempo/

Os tópicos de interesse viabilizam a coleta de tweets, por meio do Twitter Developer[3], onde existem 2 opções para a realização da coleta de tweets gratuitamente: o *Search Tweets: 30-days* onde é possível coletar até 25 mil tweets por mês que tenham sido escritos nos últimos 30 dias, e o *Search Tweets: Full Archive* que possibilita coletar até 5 mil tweets por mês, independente da data de publicação. O *Search Tweets: Full Archive* mostra-se mais adequado ao nosso caso, mas é bastante limitado pela quantidade de tweets que podem ser coletados.

A limitação existente no *Search Tweets: Full Archive* foi contornada por meio do desenvolvimento de um algoritmo de coleta manual, capaz de utilizar a função "busca avançada" da própria plataforma que permite buscar os tweets desejados selecionando automaticamente as características desejadas (período desejado, idioma, etc). Neste algoritmo, utilizamos as seguinte bibliotecas do python: *pyautogui*, para controlar o mouse e teclado e selecionar tweets para coletar o Id (dígitos existentes no final de cada tweet) por meio de sua Url, e o *win32clipboard*, onde o Id é copiado e registrado em um documento no formato txt, para nosso armazenamento. Para evitar bloqueios quanto ao tempo utilizado para coleta de dados no Twitter, utilizamos a biblioteca *time* para parar o código por poucos segundos, de forma intermitente. Ao todo, coletamos 104.678 tweets.

O Twitter não permite o compartilhamento de bases de dados contendo atributos como localização, usuário e textos, mas apenas os Ids. Para se ter acesso aos textos dos tweets, é necessário **hidratar tweets**, por meio do Twitter Developer, onde as chaves de acesso são fornecidas após cadastro. Com estas chaves, é possível utilizar as funções da biblioteca Twarc[4] do Python para manipular a API de hidratação do Twitter e captar os dados em formato JSON. O armazenamento dos tweets devidamente hidratados é feito em um banco de dados MongoDB.

A maior parte dos tweets coletados não possuem localização como atributo. Para contornar esta limitação, optamos por utilizar a localização existente nos perfis dos usuários. Utilizamos a biblioteca *geopy.geocoders* do Python, capaz de identificar uma localização escrita em um texto, retornar esse atributo de forma padronizada e as suas respectivas coordenadas geográficas. Obtivemos 53.9% dos tweets com localização dentro do território brasileiro, 6.2% dos tweets fora do Brasil, e 40% sem localização definida. Os tweets com localização em território brasileiro (56.421 tweets) passaram por um **processo de limpeza** (ex: pontuações, quebra de linhas, stop words, links), e de **tradução** para a língua inglesa por meio da biblioteca $google\_trans\_new$ da linguagem de programação Python.

Os tweets, limpos e traduzidos, passam a ser classificados por meio do analisador de sentimentos VADER (Valence Aware Dictionary for Sentiment Reasoning). O VADER é um modelo para análise de sentimentos desenvolvido especialmente para o contexto de redes sociais, sem requerer treinamento, capaz de avaliar padrões como excesso de pontuações, utilização de letras maiúsculas, *emojis*, *emoticons* e conjunções que podem inverter a polaridade do sentimento da mensagem [Hutto and Gilbert 2015]. Ao analisar entidades em um texto, o VADER retorna resultados em quatro categorias: *negative*, *neutral*, *positive* e *compound*, sendo este último uma normalização dos valores anteriores entre -1 (extremamente negativo) a 1 (extremamente positivo), 0 como neutro.

[3]`https://developer.twitter.com/en`

[4]`https://github.com/DocNow/twarc`

3. Resultados

Obtivemos tweets com localização de todas as regiões e estados brasileiros. 58% dos tweets originários da região Sudeste, 16% da região Sul, 15% da região Nordeste, 6% da região Norte e 5% da região Centro-Oeste. Os estados com mais tweets compartilhados são Rio de Janeiro, São Paulo, e Minas Gerais.

A análise de sentimentos dos tweets utilizando a opção *compound* do VADER demonstra que as regiões e os estados presentes nas mesmas possuem estados afetivos proporcionais em relação aos textos analisados: 32% possuem textos com sentimentos positivos, 31% possuem textos com sentimentos negativos e 36% possuem textos com sentimentos neutros, considerando como intervalos de classificação entre $-1,0$ e $-0,2$ para sentimentos negativos, $-0,19$ e $+0,19$ para sentimentos neutros, e $+0,2$ e $+1,0$ para sentimentos positivos A exceção desta proporção ocorre em períodos específicos (tópicos como $ForaBolsonaro$ ou $Morte\_Covid19$), onde encontramos maior polarização a determinados políticos, onde os sentimentos analisadas são predominantemente negativos (menores que -0,5).

4. Conclusões

Neste artigo é mostrado o desenvolvimento de um sistema para coleta e tratamento de textos envolvendo polarização política popular por meio do Twitter, e sua localizações obtidas a partir de palavras do próprio texto. Estes tweets foram hidratados, tratados e traduzidos, e tiveram seus textos analisados pelo VADER, um dos analisador léxicos mais utilizados hoje para determinação de estados afetivos. O sistema aqui proposto pode ser automatizado, o que e permitirá,o desenvolvimento de um *embedding* de palavras sobre polarização políticas no Brasil, assim como a montagem de um visualizador para análise gráfica de resultados e tendência dos tweets coletados.

Referências

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Baldassarri, D. and Gelman, A. (2008). Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446.

Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Jiang, J., Chen, E., Yan, S., Lerman, K., and Ferrara, E. (2020). Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211.

Kamienski, C., Mazim, L., Penteado, C., Goya, D., Genova, D. D., Franca, F. D., Ramos, D., and F. Horita, F. (2021). A polarization approach for understanding online conflicts in times of pandemic: A brazilian case study. In *Hawaii International Conference on System Sciences*.

Moslehm, M., Pennycook, G., Arechar, A., and Rand, D. (2021). Cognitive reflection correlates with behavior on twitter. *Nature Communication*, 921(12).