O Uso das Redes Sociais Digitais e as Eleições para o Cargo de Prefeito do Rio de Janeiro

Ana Paula L. F. Vasconcelos[1], Jivago Medeiros[1], Rafael Escalfoni[2], Silas P. Lima Filho(1), and Sírius Thadeu Ferreira da Silva[1]

[1] PPGI/UFRJ
[2] CEFET/RJ N. Friburgo

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introdução

O uso das mídias sociais proporcionam a interação entre candidatos, seus partidos, e eleitores (entre si) de uma maneira direta, de modo a contornar o filtro imposto pelas mídias tradicionais (estatais e comerciais), permitindo que os eleitores tenham acesso a informações quanto ao posicionamento partidário dos candidatos e questões de interesse, levando ao uma massa de eleitores mais informados politicamente, no entanto, tais eleitores ficam vulneráveis aos vieses da ênfase seletiva dos partidos e seus apoiadores [3].

Problemas relacionados à comunicação continuam a ser comumente associados ao uso massivo das mídias sociais, por outro lado, essas mesmas mídias sociais assumiram relevante papel em processos democráticos deliberativos ao redor do mundo, como por exemplo o Brexit, as eleições de 2016 nos EUA e as eleições de 2018 no Brasil.

Nesse cenário, o Twitter continua a ser uma das principais mídias sociais da atualidade, motivo pelo qual o presente estudo analisa a relação entre o engajamento dos candidatos ao cargo de Prefeito do Rio de Janeiro e o resultado das eleições municipais, ao investigar qual o nível de protagonismo das mídias sociais nas eleições municipais do Rio de Janeiro. Adicionalmente, explorar qual a influência de tais plataformas sociais no resultado das eleições.

Dado o exposto, neste artigo, apresentado como um dos requisito para obtenção dos créditos da disciplina de Fundamentos de Data Science do Programa de Pós-Graduação em Informática (PPGI) da UFRJ, apresentaremos a análise de tweets relacionados às eleições municipais para prefeito do município do Rio de Janeiro.

2 Fundamentação Teórica

A mídia estatal e comercial atuam moldadas em modelos de comunicação que são determinados por hierarquização e manipulação da informação, com o intuito de

assegurar determinados interesses econômicos, sociais ou até mesmo políticos de quem propaga a informação ou de seus parceiros e aliados [2].

No entanto, com o advento da Internet, após um longo período de hegemonia midiática deu-se abertura de espaço para informações provenientes de outras fontes, pois a mídia tradicional passou a dividir espaço com plataformas de comunicação e redes sociais, onde as informações são "[...] produzidas e fomentadas pelos próprios usuários da rede e não por veículos de informação enviesados. [...]" (CANDIDO, 2003 apud MOURA, 2018, p. 26).

Atualmente, qualquer pessoa é capaz de produzir e propagar conteúdos online que podem vir a ter um alcance de pessoas a nível mundial com extrema rapidez e facilidade, em virtude do fluxo intenso de informações que circulam através da Internet, onde "[...] grande parte do conteúdo disponível [...] é construído de forma colaborativa entre os usuários que compõem a rede. As relações surgem do conjunto de interações estabelecido entre os usuários (sujeitos)." [2] [7].

Segundo Lupianhes (2017), o século 21 foi o marco da explosão de popularidade das redes sociais, tais como Facebook, MySpace, Twitter, LikedIn, entre outras, que transcenderam as fronteiras idiomáticas e culturais na comunicação. E ao longo dos anos, vem sendo observada uma tendência global quanto ao uso das redes sociais para interação e realização de campanhas políticas durante o período eleitoral [1].

Verifica-se que a cada dia, mais do que nunca, há mais informações relacionadas à política sendo compartilhadas, e os cidadãos comuns passaram a desempenhar um papel ativo no ciclo de disseminação de notícias. Pesquisas constataram que os usuários das mídias sociais passaram a ficar melhor informados quanto a política durante o período eleitoral, a exemplo das eleições gerais de 2015 no Reino Unido [3] [7].

Há algum tempo tal fenômeno vem despertando o interesse de pesquisadores das mais diversas áreas, em investigar possíveis impactos que o uso de redes sociais possam representar sobre o processo eleitoral, tendo em vista que há evidências de que os meios de comunicação social convencionais possuem a capacidade de influenciar não apenas a opinião pública, mas também o ideário social. Além disso, prospecta-se um crescente engajamento dos usuários convergindo ao pertencimento a uma dada identidade social que luta coletivamente por melhorias [2].

Ao analisar dados de pesquisa e relacioná-los com tweets, dos entrevistados, durante o período eleitoral (das eleições parlamentares de 2015 no Reino Unido), pesquisadores identificaram evidências de que os tweets dos candidatos e partidos mostraram-se responsáveis por conduzir a população ao conhecimento sobre fatos que eram politicamente relevantes, expor o posicionamento dos partidos frente a diferentes e importantes questões políticas, influenciando fortemente a visão dos eleitores nos mais diversos assuntos [3].

No decorrer da campanha eleitoral, a exposição dos eleitores a mensagens partidárias com relação a questões sensíveis tentando conduzir tais eleitores a direcionamentos favoráveis aos partidos, pode causar polarização do conhecimento acerca de tais questões, pois observa-se que há um direcionamento, mas

que este ocorreu com um foco impreciso, o que motiva preocupação pelos efeitos nocivos da desinformação disseminada nas redes sociais [3].

3 Metodologia

Buscando compreender o impacto do engajamento político no Twitter com o resultado das eleições, este trabalho buscou analisar dados relacionados às eleições municipais 2020 no município do Rio de Janeiro, retirados do Twitter, através da API do Twitter a partir dos tweets compartilhados.

3.1 O Dataset (DS)

O dataset possui 3.3GB, com aproximadamente 580.354 tweets coletados entre 29/09/2020 à 01/12/2020 usando a API do twitter. Tais tweets foram retornados com base em termos pré-definidos e informados à API, que combinam os nomes dos candidatos, o número de seu respectivo partido, a sigla do partido e o ano da eleição.

```
[ {nomeCandidato} AND [ {numeroPartido} OR {siglaPartido} OR $ANO] ]
```

A coleta de dados teve por objetivo criar um dataset com tweets diretamente relacionados aos candidatos à prefeitura do município do Rio de Janeiro durante o período de campanha eleitoral determinado pelo Supremo Tribunal Eleitoral.

Coleta dos Dados e Criação do Dataset Para coletar os dados via API oficial do Twitter foi utilizado um script escrito na linguagem Python com a biblioteca Tweepy[3]. A biblioteca Tweepy simplifica o uso da API do Twitter encapsulando e abstraindo alguns dos passos necessários para a coleta de dados. O script encontra-se disponível no repositório desse artigo no github[4].

A coleta de dados deu-se por meio da funcionalidade *Sampled stream*[5] pela qual a API do Twitter retorna uma amostra aleatória de 1% dos tweets públicos disponíveis em "tempo real" de acordo com os termos buscados/filtrados. Os termos utilizados na coleta de dados encontram-se descritos no Apêndice A.

O script utilizado coleta os dados exatamente da forma como são retornados pela API do Twitter e faz a persistência desses no SGBD MongoDB. Após a finalização da coleta de dados utilizou-se o comando

[3] https://www.tweepy.org/
[4] https://github.com/FundamentosDataScienceEleicoesRJ2020/Sandbox
[5] https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction

3.2 Análise e tratamento dos dados

Através da linguagem de programação python e suas principais bibliotecas para ciência de dados (como NumPy, Pandas, NLTK, Matplotlib, Seaborn, etc.) pretendemos explorar o dataset, realizando os tratamentos necessários para extrairmos as informações almejadas por esta pesquisa. Como estamos lidando com diversos tipos de dados distribuídos em colunas, cada coluna (ou conjunto de coluna com o mesmo tipo de dado) deverá passar por processos de extração e tratamento de dados distintos a fim de coletarmos as informações necessárias.

Serão utilizados métodos de tratamento de dados voltados para cada tipo de dado encontrado e técnicas de processamento de texto (Atomização, Contagem de palavras, Divisão de frases, Radicalização, Normalização e Remoção das "stop words") visando possibilitar uma análise textual automática, até mesmo realizando análise de sentimento quando possível.

No caso das colunas de tipos numéricos, verificaremos a existência de dados nulos ou NaN (Not a Number). Em caso positivo, de acordo com as regras descritivas da coluna em questão, poderemos atribuir o valor padrão 0 (para as colunas onde a ocorrência desse valor não é possível naturalmente) ou o valor padrão negativo -1 (para as colunas onde a ocorrência de valores negativos não são possíveis naturalmente).

Já no caso das colunas tipo texto (String), verificaremos a existência de dados nulos e os substituiremos pela String vazia. Além disso, utilizaremos técnicas de processamento de texto para normalizar esses dados (uniformização das letras maiúsculas e minúsculas, remoção de acentuação, pontuação, caracteres inválidos, etc.).

Também poderemos utilizar métodos de detecção de outliers baseados em estatística, distância ou modelo, cada um sendo aplicado nos casos em que forem julgados mais aptos. Para os dados textuais, o pré-processamento dos dados inclui:

- Identificação de outliers (tweets não relacionados ao contexto da Eleição, tweets inválidos/não úteis);
- Nivelamento das palavras em minúsculas;
- Remoção de stop words;
- Remoção de sufixos;
- Remoção dos atributos nulos;
- Remoção de atributos não relevantes.

3.3 Execução

Para a execução do projeto foi utilizado o Google Colab[6]. O Google Colaboratory ou Colab permite escrever código Python diretamente no navegador, com fácil compartilhamento e acesso gratuito a GPUs (e demais hardwares e softwares necessários). Nenhuma configuração extra é necessária para começar a trabalhar

[6] https://colab.research.google.com/

com essa ferramenta, exceto a instalação de algumas bibliotecas específicas do Python que porventura não estejam já instaladas no ambiente (e seja preciso utilizar durante o projeto), facilitando assim o trabalho de um cientista de dados.

O Google Colaboratory é construído sobre o Jupyter Notebook e fornece um ambiente interativo chamado notebook Colab que permite escrever e executar código (script Python). Os notebooks do Colab permitem combinar código executável e rich text em um só documento, além de imagens, HTML, LaTeX e muito mais. Os notebooks do Colab são armazenados na conta do Google Drive. É possível compartilhar os notebooks do Colab facilmente com outras pessoas e permitir que elas façam comentários ou até editem o documento. O Colab fornece completamentos automáticos para explorar atributos de objetos Python, bem como para visualizar rapidamente sua documentação.

O Colab é uma ferramenta adequada para se trabalhar com a área de Ciência de Dados, sendo possível aproveitar o potencial das bibliotecas Python para analisar e visualizar dados (numpy, matplotlib, pandas, etc.). Sendo possível importar para os notebooks do Colab os dados de uma conta do Google Drive, do GitHub e de diversas outras fontes. Além disso, o Colab é altamente integrável com o Google Drive, permitindo montar um drive virtual para leitura e armazenamento perpétuo dos dados gerados pelos scripts Python (uma vez que o ambiente de execução do Colab é temporário, sendo "zerado" após longos períodos de inatividade do usuário – ou a pedido do próprio usuário). Os notebooks do Colab executam os códigos nos servidores em nuvem do Google, significando que podemos tirar proveito da potência de hardware do Google, como GPUs e TPUs, independentemente da potência do computador local, necessitando somente de um navegador com acesso à Internet.

Foi utilizada uma biblioteca de análise exploratória de dados, chamada Sweetviz[7], que faz uso dos dataframes do pandas e cria um relatório HTML autocontido que pode ser visualizado em um navegador ou integrado em notebooks. Além de criar visualizações perspicazes e bonitas com apenas duas linhas de código, ela fornece análises que levariam muito mais tempo para serem geradas manualmente, incluindo algumas que nenhuma outra biblioteca fornece tão rapidamente, como:

- Análise de alvo: mostra como um valor alvo (por exemplo, "Sobreviveu" no dataset do Titanic) se relaciona com outros recursos;
- Comparações de dataset: entre datasets (por exemplo, "Treino x Teste") e intra-conjunto (por exemplo, "Homem x Mulher");
- Correlações / associações: integração total de correlações e associações de dados numéricos e categóricos, tudo em um único gráfico e tabela.

[7] https://pypi.org/project/sweetviz/

3.4 Proveniência

Para coletar os metadados de proveniência do dataset escolhido, e gerar o respectivo grafo de proveniência foi utilizada a biblioteca PROV do Python[8], em sua versão 2.0.0.

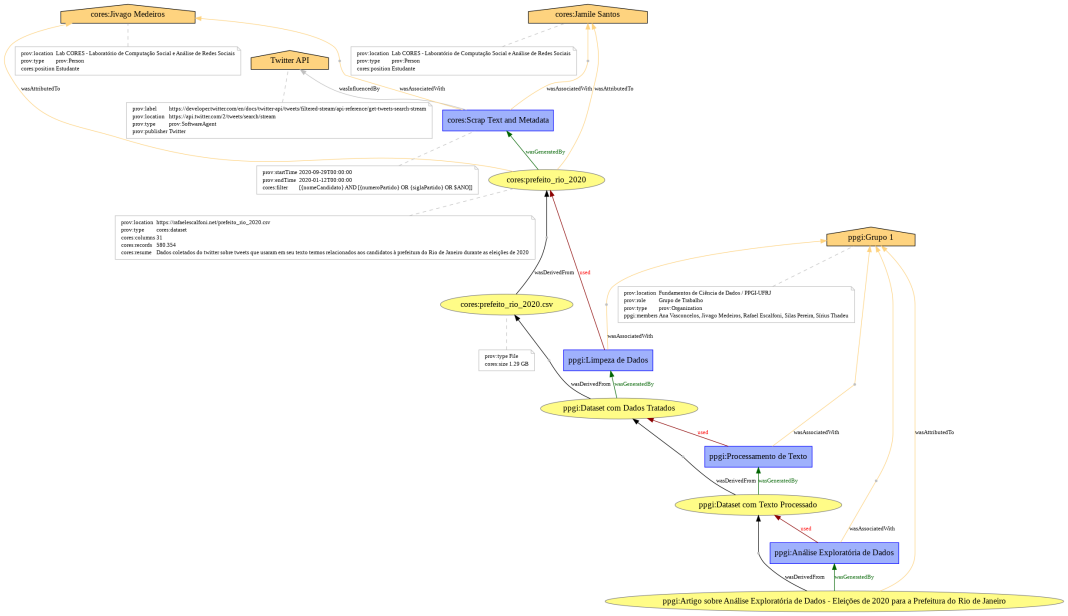

Fig. 1. Proveniência da base de dados adotada.

3.5 Reprodutibilidade

Seguindo-se os preceitos dispostos em por Benureau e Rougier [9] em "Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions", coletamos todas as informações necessárias de ambiente, tais como Sistema Operacional e sua arquitetura; Linguagem de Programação e as bibliotecas utilizadas, para tornar o experimento reprodutível. Além disso, os scripts Python estão sendo criados com o maior zelo possível, visando documentar os passos executados durante todo o experimento e seguir as boas práticas de programação para tornar esses códigos reusáveis.

No presente artigo será documentada a metodologia aplicada, assim como serão descritos os processos, algoritmos e modelos utilizados durante o projeto, visando assim permitir que este experimento seja replicado por outros cientistas.

[8] https://pypi.org/project/prov/

Para o experimento, utilizamos a ferramenta Google Colab, que possui as seguintes configurações de ambiente de execução[9]

- Sistema Operacional Ubuntu 18.04.5 LTS (Bionic Beaver), com kernel Linux versão 4.19.112+ (Chromium OS versão 10.0.0) e arquitetura x86-64 (64-bit) com ordem de byte "Little Endian" e dois processadores Intel® Xeon® de 2.20GHz cada;
- Python na versão 3.7.10 [GCC 7.5.0] e as bibliotecas Pandas versão 1.1.5, Beautiful Soup versão 4.6.3, PROV versão 2.0.0 e Sweetviz versão 2.0.9.

4 Resultados

No decorrer do trabalho uma das dificuldades encontradas foi com relação ao tamanho do dataset, pois em virtude de seus mais de 580 mil registros não foi possível abrir o DS diretamente no Pandas através do Colab, o que resultou em estouro de memória[10].

Verificou-se que a análise demanda muita memória para processamento, e apenas para a leitura do arquivo json, foi utilizada praticamente toda a memória disponibilizada pelo Colab. Em virtude desse problema, como estratégia para que sua análise torne-se possível, levantou-se possibilidades como: fragmentar o dataset, dividir em chunks de processamento, salvar o DS como jsonlines (onde cada linha vira uma string), gerar outro DS com os campos de cada processamento, fazer leitura e processamento em partes.

5 Conclusões

Em andamento.

References

1. LUPIANHES, K. A Influência das Redes Sociais na Comunicação e no Ambiente de Trabalho. REFAS - Revista Fatec Zona Sul. v. 3. n. 2. 2017. ISSN 2359-182X.
2. MOURA, R. S. Eleições 2.0: O Uso das Redes Sociais Digitais Durante as Eleições Suplementares ao Governo do Estado do Amazonas. Dissertação de Mestrado em Psicologia da Universidade Federal do Amazonas. Manaus, 2018.
3. MUNGER, K. et al. Political Knowledge and Misinformation in the Era of Social Media: Evidence from the 2015 U.K. Election. British Journal of Political Science. 2020. https://doi.org/https://doi.org/10.1017/S0007123420000198

[9] Mais informações sobre as configurações do ambiente utilizado durante o experimento podem ser visualizadas através do arquivo Environment.conf, disponível no GitHub https://github.com/FundamentosDataScienceEleicoesRJ2020/Sandbox/tree/main/artefatos

[10] A ferramenta Colab disponibiliza 13Gb de memória exclusivamente para programação

4. RECUERO, R.; SOARES, F.; ZAGO, G. Polarização, Hiperpartidarismo e Câmaras de Eco: Como Circula a Desinformação sobre Covid-19 no Twitter. Revista Contracampo. 2020.
5. SANTOS JUNIOR, M. A.; ALBUQUERQUE, A. Perda da hegemonia da imprensa: a disputa pela visibilidade na eleição de 2018. Revista do Programa de Pós-graduação em Comunicação. , v. 13, n. 3, 2019. e-ISSN: 1981-4070.
6. SOARES, F. B.; RECUERO, R.; ZAGO, G. Influencers in Polarized Political Networks on Twitter. In Proceedings of the 9th International Conference on Social Media and Society (SMSociety'18). ACM, New York, NY, USA, 168-177. DOI: https://doi.org/10.1145/3217804.3217909.
7. TUCKER, J. A. et al. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. SSRN Electronic Journal. 2018.
8. VERMELHO, S. C. et al. Refletindo sobre as redes sociais digitais. Educação & Sociedade. v. 35. n.126. 2014. ISSN: 1678-4626.
9. BENUREAU, F. C. Y.; ROUGIER, N. P. Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions. Frontiers in Neuroinformatics. V. 11. 2018. DOI 10.3389/fninf.2017.00069. ISSN 1662-5196.

A Temos de Busca Utilizados na API do Twitter

'Benedita da Silva', 'Benedita13', 'BeneditaPT', 'BeneditadaSilva2021', 'Eduardo Bandeira de Mello', 'Bandeira18', 'BandeiraRede', 'BandeiradeMello2021', 'Clarissa Garotinho', 'Clarissa90', 'ClarissaPROS', 'ClarissaGarotinho2021', 'yro Garcia', 'Cyro16', 'CyroPST', 'Cyro2021', 'CyroGarcia2021', 'Eduardo Paes', 'Eduardo25', 'EduardoPL', 'EduardoPaes2021', 'Fred Luz', 'FredLuz30', 'redLuzNOVO', 'FredLuz2021', 'Glória Heloíza', 'GlóriaHeloíza20', 'GlóriaHeloízaPSC', 'GlóriaHeloíza2021', 'Henrique Simonard', 'HenriqueSimonard29', 'HenriqueSimonardPCO', 'Simonard2021', 'Luiz Lima', 'LuizLima17', 'LuizLimaPSL', 'LuizLima2021', 'Marcelo Crivella', 'Crivella10', 'CrivellaRepublicanos', 'Crivella2021', 'Martha Rocha', 'MarthaRocha12', 'MarthaRochaPDT', 'MarthaRocha2021', 'Paulo Messina', 'PauloMessina15', 'PauloMessinaMDB', 'PauloMessina2021', 'Renata Souza', 'RenataSouza50', 'RenataSouzaPSOL', 'RenataSouza2021', 'Suêd Haidar', 'SuêdHaidar35', 'SuêdHaidarPMB', 'SuêdHaidar2021'