Análise de Dados para Comunicação Política a partir de um Sistema de Coleta de *Tweets*

Alexandre A. P. Heine[1], Bruno Coutinho[1], Mariana Barreto[1], Nicholas Xavier[2], Marcos V. Villas[1], Arthur Ituassu[3], Sérgio Lifschitz[1]

[1]Departamento de Informática – (PUC-Rio) - Rio de Janeiro - RJ

[2]Faculty of Science – The University of British Columbia - Vancouver, Canada

[3]Departamento de Comunicação – (PUC-Rio) - Rio de Janeiro - RJ

{xandeaph,brunocoutinho64,marianaportobarreto}@gmail.com,

nichogx@student.ubc.ca, {villas, ituassu, slifchitz}@puc-rio.br,

Resumo. *Este artigo apresenta o projeto, a pesquisa e o desenvolvimento de uma ferramenta de coleta e análise de dados do Twitter, que tem por objetivo avaliar os dados publicados nesta rede social, em particular voltadas para a área de comunicação política. Além de explicar brevemente a arquitetura do sistema, descrevemos algumas funcionalidades importantes, a saber: a coleta de dados por streaming; as análises relativas aos links compartilhados; a identificação de usuários que realizaram retweets; e o estudo de polaridade dos sentimentos expressos no corpo dos tweets. Dentre os desafios encontrados destacamos o pré-processamento dos dados coletados, as limitações no uso da API do Twitter e a obtenção e preparação da bases de dados para a análise de sentimentos.*

1. Introdução

A análise de dados de redes sociais se apresenta como um dos principais instrumentos para a realização de estudos na área da comunicação política. Com o aumento da presença de candidatos a cargos públicos no meio digital e a expansão global do número de pessoas com acesso à internet, as mídias sociais se tornaram um dos principais espaços de debate político [Rabelo 2010]. Nesse contexto, as informações públicas contidas na plataforma Twitter[1] são de extremo interesse para os acadêmicos interessados em comunicação política. Dentre os possíveis usos desses dados, podemos destacar as análises que se propõem a caracterizar e interpretar o debate público no contexto digital, fundamentais para a compreensão das dinâmicas de interações políticas na contemporaneidade.

Nesse cenário, a ferramenta eTC[2] (ePOCS *Twitter Crawler*) é uma aplicação de agendamento e coleta de dados do Twitter, desenvolvida por alunos da PUC-Rio, a partir do uso de um *web crawler*. Por meio de uma interface *web*, pesquisadores interessados em análise de redes sociais podem usar a ferramenta para obter dados de *tweets* antigos e, com esses, fazer suas análises [Rodriguez et al. 2019]. A coleta de *tweets* é feita mediante agendamento, etapa na qual o usuário deve informar o termo a ser buscado e o período de

[1]https://twitter.com/twitter

[2]https://etc.biobd.inf.puc-rio.br/

tempo desejado. Após a realização desse sistema centrado na coleta de dados, percebeu-se a necessidade de ampliá-lo, visando a apoiar, de forma integral, as pesquisas do grupo ePOCS[3], centradas na área da comunicação política, e demais acadêmicos interessados nesse campo.

No contexto dessa área de interesse, informações relativas ao posicionamento dos usuários em relação a um candidato, às fontes de notícia compartilhadas e à polaridade do sentimento em resposta a eles servem de base para diversas linhas de produção acadêmica [Ituassu et al. 2018]. Além disso, no período eleitoral, dados em tempo real são importantes para determinar mudanças de opinião que ocorram devido a atitudes tomadas pelos candidatos e para monitorar o debate público. Assim, a coleta de dados por *streaming* também é necessária para a realização de análises na área de comunicação política. Este trabalho apresenta, portanto, o projeto, a pesquisa e o desenvolvimento de uma ferramenta de coleta e análise de dados do Twitter direcionadas para essa área.

2. Soluções Propostas

Com o intuito de ampliar o sistema em funcionamento e dar maior apoio a pesquisas no âmbito da comunicação política, foram feitas quatro novas funcionalidades, a saber: a análise de mídia, a análise de *retweeters*, a análise de sentimentos e o *streaming* de *tweets*. Planejou-se implementá-las de maneira a aproveitar a estrutura existente da ferramenta do *crawler*, incluindo o sistema *web* e o banco de dados relacional. A Figura 1 exibe a estrutura dos componentes e suas tecnologias, destacando o desacoplamento entre as partes, ligadas somente pelo uso do mesmo banco de dados.

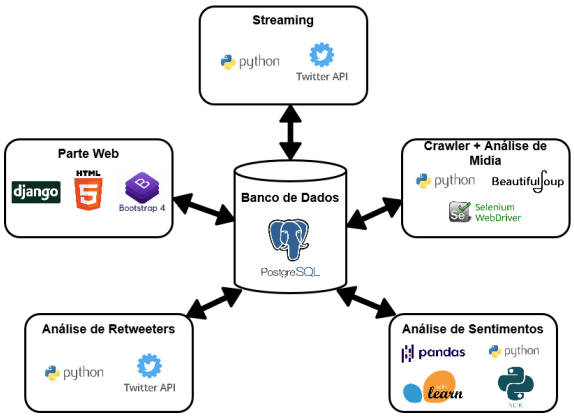


Figura 1. Diagrama com a estrutura interna do sistema e seus componentes

Cada uma das novas funcionalidades, contudo, foi projetada para operar em momentos distintos e de formas distintas. Como ilustrado na Figura 2, a análise de mídia ocorre durante a coleta de *tweets* antigos, enquanto a análise de *retweeters* e a de sentimentos precisam ser agendadas, via interface *web*, pelo usuário do sistema após o término de uma coleta. Já o *streaming* ocorre por meio de agendamento próprio, sem relação direta com a coleta de dados do *crawler*. Diferentemente das três primeiras funcionalidades, que são análises feitas utilizando-se de dados históricos obtidos pelo *crawler*, o *streaming* de *tweets* trata-se de uma coleta separada de dados em tempo real. Cada funcionalidade

[3]https://www.comprio.com.br/epocs

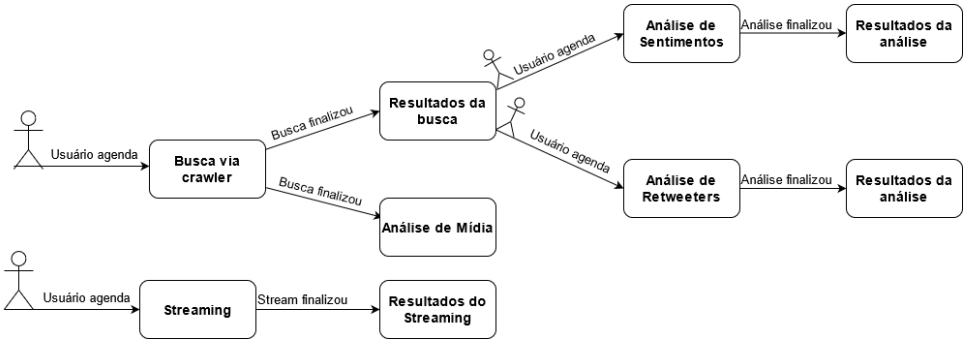


Figura 2. Sequência temporal do uso das novas funcionalidades do eTC

tem a sua própria fila de execução, à exceção da análise de mídia. No instante em que ocorre um agendamento no site, a análise agendada é inserida no final de sua respectiva fila e, de modo assíncrono, cada algoritmo executa o próximo pedido.

2.1. Análise de Mídia

Os *links* compartilhados pelos usuários nos *tweets* são informações importantes para acompanhar quais mídias são mais citadas em determinada busca. Além disso, é possível observar a evolução da *mainstream media versus* as "mídias complementares" [Ituassu et al. 2018]. A ferramenta eTC já apresentava, em sua última versão, uma análise que disponibilizava as mídias mais citadas em uma determinada busca [Rodriguez et al. 2019]. No entanto, tal análise estava limitada aos *links* que se encontravam no texto do *tweet*. Após atualizações da própria plataforma, as URLs deixaram de compor o texto, agora aparecendo apenas em sua versão encurtada (com domínio t.co). Dessa maneira, a análise não podia ser obtida devido à falta de informação.

Para conseguir saber, portanto, quais mídias eram mais compartilhadas, foi preciso encontrar uma forma de obter pelo menos os domínios das URLs. A princípio, seria possível obtê-los fazendo *requests* a cada *link* coletado. Contudo, o volume de *requests* feitos a diferentes *sites*, em um curto período de tempo, seria grande o bastante para que isso fosse visto como ataques de DoS (*Denial of Service*). Com o objetivo de encontrar uma outra alternativa, foi observado que o *Twitter* disponibiliza, atualmente, o acesso aos *links* compartilhados de duas formas: a primeira por um *card* com imagem e resumo do *link*; e a outra por meio da URL incompleta no texto do *tweet*. Em ambas as opções, as URLs completas não podem ser obtidas, porém o domínio da URL pode ser coletado na própria página HTML, em que é feita a coleta do *crawler*. Como apenas o domínio é necessário para atender a essa análise, a solução encontrada é capaz de substituir a anterior sem a perda de informações.

2.2. Análise de *Retweeters*

Além dessa funcionalidade, outra demanda que se colocou foi a de obter a informação de quais usuários deram *retweet* em determinados *tweets*, ou seja, quais foram os *retweeters* de um conjunto de *tweets*. *Retweet*, no contexto do *Twitter*, significa republicar um *tweet* postado por outra conta. Tal necessidade partiu de uma pesquisa interdisciplinar [dos Santos et al. 2021], a qual almejava analisar os perfis que realizaram *retweets* de

candidatos políticos e, em seguida, avaliar, por meio da ferramenta Pegabot[4], a probabilidade desses perfis serem automatizados. Como não é possível extrair dados acerca dos *retweeters* de um *tweet* via *crawling*, fez-se necessária a coleta dessas informações pela API[5] (*Application Programming Interface*) do *Twitter*.

Por meio do uso do ID de um *tweet* como parâmetro, pode-se utilizar um *endpoint*[6] da API, que fornece uma amostra de até 100 IDs de usuários que fizeram seu *retweet*. No contexto dessa API, um *endpoint* representa uma requisição, em que são fornecidos parâmetros e são retornados dados estruturados. Após a coleta dos IDs de usuário dos *retweeters*, foi utilizado outro *endpoint*[7], para obter *usernames* dos mesmos perfis, pois o Pegabot trabalha somente com esse tipo de identificação de usuário. Assim, a análise se dividiu em duas etapas lógicas: a coleta de IDs de *retweeters* e a obtenção dos *usernames* a partir dos IDs. Em termos computacionais, as etapas foram implementadas como processos distintos, executados em paralelo e conectados indiretamente pela utilização do mesmo banco de dados relacional.

2.3. Análise de Sentimentos

Como forma de identificar a polaridade dos sentimentos dos usuários do *Twitter*, que interagiram com candidatos de eleições analisadas, foram estudados métodos de análise de sentimentos como: o uso da polaridade de sentimentos de palavras individuais, a partir de dicionários de sentimentos de léxicos [Silva and de Oliveira 2018]; o uso de ferramentas de tradução em apoio a aplicações de aprendizado de máquina já existentes em outras línguas [Araújo et al. 2020], por não haver, até o momento desta pesquisa, um modelo treinado para a língua portuguesa; e o uso de bases de dados em português classificadas em conjunto com técnicas de aprendizado de máquina ou métodos probabilísticos, desejando-se obter uma nova ferramenta, para realizar análises na linguagem original.

Com isso, optou-se pelo uso de bases de dados em português, já classificadas com métodos probabilísticos e aprendizado de máquina, pois, nas demais opções, percebeu-se a necessidade de parcerias, conhecimento especializado na área de Linguística ou do uso de serviços pagos. Para isso, foi realizada uma pesquisa por bases, que oferecessem informação sobre a polaridade dos sentimentos das palavras e frases. Desse modo, foram encontrados o *SentiLex-PT* [Carvalho and Silva 2015] e o *OntoPT* [Gonçalo Oliveira et al. 2014], dicionários léxicos de sentimentos em português, que mapeiam a polaridade de palavras e expressões; e outras duas bases com o sentimento de frases de *tweets*: o TAS-PT[8] e o "*Tweets from MG/BR*"[9].

Com essas bases de dados, foi montado um *script* de pré-processamento de dados, a fim de remover ou substituir trechos das frases analisadas, como algumas palavras escritas de forma errada, caracteres especiais, dentre outros. Dada essa etapa de limpeza,

[4] https://pegabot.com.br/

[5] https://developer.twitter.com/en/docs/twitter-api/v1

[6] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-retweeters-ids

[7] https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/api-reference/get-users-lookup

[8] https://github.com/pauloemmilio/dataset

[9] https://www.kaggle.com/leandrodoze/tweets-from-mgbr

as bases de *tweets* foram separadas em dois conjuntos, um para treino (65%) e outro para teste (35%). Depois dessa separação, foram removidas *stop words* e foram convertidos os textos, do conjunto de treino, em uma matriz de contagem de *tokens*. Um *token* é a representação de um agrupamento de uma ou mais palavras.

A partir disso, foram utilizados os conjuntos de treino e teste para treinar modelos probabilísticos e de aprendizado de máquina. Neste trabalho, como foram utilizadas bases com dados rotulados, optou-se por métodos supervisionados. Além disso, foram utilizadas três classificações para os dados: positivo, neutro e negativo. Logo, foram treinados e comparados os seguintes modelos: *multinomial* Naive Bayes[10], regressão logística[11], máquina de vetores de suporte (SVM[12]) e *multi-layer perceptron*[13].

2.4. Coleta de Dados por *Streaming*

No método de obtenção de *tweets* passados através de *crawling*, dois problemas principais surgiram: o tempo necessário para capturar os *tweets* e a quantidade de *tweets* que o Twitter disponibilizava para esse tipo de coleta. A alternativa explorada, para resolver esses pontos, foi a API de *Streaming*[14], disponibilizada pela própria rede social.

Essa ferramenta possibilitou a captura em tempo real de *tweets* de acordo com termos procurados em uma busca. Dessa forma, foi possível capturar todos os *tweets*, inclusive aqueles que poderiam ser apagados posteriormente, os quais não seriam encontrados pelo *crawler*. Assim, foi desenvolvido um *software* para acessar a API, armazenando o resultado em banco de dados para uso posterior. O *software* tem a capacidade de fazer apenas uma busca simultânea e, para controle de acesso, foi feita uma interface em que os usuários podem agendar uma busca e ver os horários disponíveis para o agendamento.

3. Resultados

Os resultados das análises são persistidos em um banco de dados e se encontram disponíveis para visualização no *site* do eTC. No caso das análises de *retweeters* e de sentimentos, além de poderem consultar estatísticas e gráficos, há a opção de baixar planilhas com os resultados. Após a conclusão do *streaming*, a consulta aos resultados é feita exclusivamente pela opção de *download* dos dados presente no *site*. Em relação ao volume de dados suportado pela ferramenta, na análise de retweeters temos um limite imposto pela API do Twitter de 300 requisições a cada 15 minutos, o que resultou em um tempo de execução de 39 horas para a coleta de 10857 retweeters provenientes de 3539 tweets, por exemplo. No caso do *streaming*, a principal restrição trata-se do limite de 10 milhões de tweets coletados por mês. Mesmo assim, em um de nossos testes, nossa ferramenta coletou 30.629 tweets no espaço de uma hora. Já a análise de sentimentos é limitado apenas pelo tempo de execução dos algoritmos, analisando, por exemplo, 200.120 tweets em 1h25.

[10] https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes

[11] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

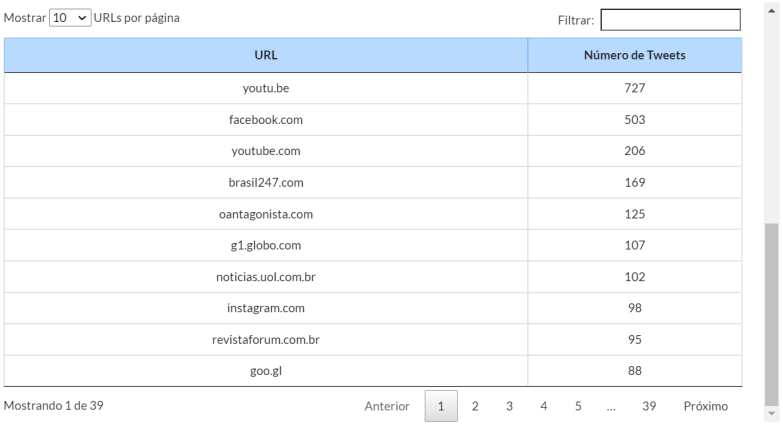
[12] https://scikit-learn.org/stable/modules/svm.html

[13] https://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron

[14] https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time

Em relação aos modelos treinados na análise de sentimentos, foi escolhido o *multilayer perceptron* como padrão, por ter apresentado os melhores resultados: 81,3% de acurácia e 72,2% de precisão. Foi feita a variação também do tamanho dos conjuntos de treino e teste e a retirada ou não de *stop words*. Mesmo assim, há um painel na aplicação *web* em que é possível selecionar os demais modelos.

O modelo escolhido permitiu verificar o quanto a polaridade de sentimentos influencia na probabilidade de voto de um determinado candidato. Usando três classificações (positiva, neutra e negativa), foi realizada uma classificação manual, obtendo-se inicialmente um acerto de 40% para esse caso. Vários dos casos que geraram erros eram referentes a sentimentos em relação a candidatos de oposição ou menção de múltiplos candidatos, casos os quais não foram contemplados no tratamento dos dados analisados.


Mostrar 10 URLs por página

Filtrar:

URL	Número de Tweets
youtu.be	727
facebook.com	503
youtube.com	206
brasil247.com	169
oantagonista.com	125
g1.globo.com	107
noticias.uol.com.br	102
instagram.com	98
revistaforum.com.br	95
goo.gl	88

Mostrando 1 de 39 Anterior 1 2 3 4 5 ... 39 Próximo

Figura 3. Exemplo da tabela da página de Análise de Mídia

Para análise de mídia, anteriormente, os domínios das URLs compartilhadas eram extraídos do texto do *tweet* por meio de uma consulta ao banco de dados que listava as URLs e uma expressão regular que filtrava os domínios. Agora, todas as etapas de limpeza e filtragem dos dados são feitas durante a coleta dos *tweets* e as informações são persistidas no banco de dados. A página de visualização da análise de mídia, conforme exibido na Figura 3, apresenta uma tabela com os todos domínios e a quantidade de vezes que eles foram mencionados durante a busca.

4. Considerações Finais

Neste projeto, a partir do *streaming* e das análises feitas, foram obtidos grandes volumes de dados estatísticos, importantes para o conhecimento do meio político: quais mídias determinados grupos utilizam, os sentimentos associados ao voto em um determinado candidato, além dos perfis que mais interagem com determinado político através de *retweets*.

Como é possível notar, a ferramenta está em constante evolução. Uma das principais atualizações previstas é a migração da coleta do *crawler* para a API do Twitter, tendo em vista a nova versão gratuita voltada para a academia, que possui um *endpoint* de busca histórica de *tweets*. Para a análise de sentimentos, um caminho a ser seguido é tentar reconhecer as preferências de voto dos usuários, não mais verificando a polaridade. Já na parte de análise de mídia, os próximos passos envolvem aprimorar a visualização dos resultados e fornecer o *download* dos dados. Por fim, um dos maiores interesses do eTC é ampliar sua coleta de dados, incluindo outras redes sociais, como *Facebook* e *Instagram*.

Referências

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

Carvalho, P. and Silva, M. J. (2015). Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*, 7(1).

dos Santos, J. G. B., Ituassu, A., Lifschitz, S., Guimarães, T., Cerqueira, D., Albu, D., Fernando, R., Ferreira, J. H., and Mondelli, M. L. (2021). Das milícias digitais ao comportamento coordenado: métodos interdisciplinares de análise e identificação de bots nas eleições brasileiras. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 187–192. SBC.

Gonçalo Oliveira, H., Paulo-Santos, A., and Gomes, P. (2014). Assigning polarity automatically to the synsets of a wordnet-like resource. In *Symposium on Languages, Applications and Technologies (SLATE)*, OASICS, pages 169–184.

Ituassu, A., Lifschitz, S., Capone, L., Vaz, M. B., and Mannheimer, V. (2018). Compartilhamento de mídia e preferência eleitoral no twitter: uma análise de opinião pública durante as eleições de 2014 no brasil. *Palabra Clave [online]*, 21(3):860–884.

Rabelo, L. (2010). As mídias sociais e a esfera pública: mudanças de paradigma na comunicação contemporânea. In *Anais do XII Congresso de Ciências da Comunicação na Região Centro-Oeste*, pages 27–29.

Rodriguez, A. M., Sava, P. S., Ituassu, A., and Lifschitz, S. (2019). Sistema web crawler para coleta automática de tweets, persistência em bancos de dados e análises estatísticas. In *Companion Proceedings SBBD*, pages 325–332.

Silva, E. A. and de Oliveira, L. F. R. (2018). Análise de sentimentos: Identificando sentimentos em comentários da rede humaniza sus. In *II Workshop de Informação, Dados e Tecnologia, UFPB*.