Gerenciamento de Dados de Redes Sociais com Análise de Redes e Modelagem de Tópicos

Isabella Carmo[1], André L. C. Rêgo[1], Mariana Barreto[1], Marina Schuler[1], Alexandre Heine[1], Marcos V. Villas[1], Sérgio Lifschitz[1]

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

{isabellalcarmo, andreregо}@aluno.puc-rio.br,

{marianabarreto, marina_schuler}@aluno.puc-rio.br,

{aheine, villas, sergio}@inf.puc-rio.br

Resumo. *Este artigo diz respeito à gestão, manuseio e análise de dados obtidos de redes sociais digitais. Inicialmente, explica-se como as informações coletadas foram usadas para realizar análises, detalhando o uso de um software dedicado à análise de redes, para observar as possíveis comunidades, e o uso de algoritmos para modelagem de tópicos. São apresentados resultados práticos para cada um destes processos de enriquecimento dos dados para pesquisas envolvendo postagens sobre vacinação no Brasil e respostas a mulheres candidatas nas eleições de 2022 no Brasil. Dentre os desafios encontrados, destacam-se a capacidade para lidar com grandes volumes dados, a aplicação dos conceitos de análise de redes e a inferência de tópicos após a aplicação dos algoritmos. Apesar dos exemplos e experimentos serem relacionados com o Twitter, os assuntos aqui investigados e discutidos podem se aplicar a qualquer rede social.*

Abstract. *This article concerns the management, handling and analysis of data obtained from digital social networks. Initially, it is explained how the information collected was used to perform analyses, detailing the use of software dedicated to network analysis, observing possible communities, and topic modeling algorithms. Practical results are presented for each of these data enrichment processes for researches involving posts about vaccination in Brazil and responses to women candidates in 2022 elections in Brazil. Among the challenges encountered, we highlight the ability to deal with large volumes of data, the application of network analysis concepts and the inference of topics given the algorithms results. Although the examples and experiments are related to Twitter, the subjects investigated and discussed here can apply to other social networks.*

1. Introdução

A ferramenta eTC (ePOCS Twitter Crawler)[1] foi desenvolvida para coletar *tweets* relacionados com palavras de busca, com ou sem *hashtag*, de determinado período de tempo, usando uma interface amigável para usuários não técnicos e com gestão de filas de pedidos

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

de coleta. Uma vez obtidos os dados, pode-se realizar análises estatísticas sobre os *datasets* coletados ou ainda fazer o download da coleção de *tweets*. A eTC é uma aplicação que vem sendo desenvolvida por alunos do DI PUC-Rio que, inicialmente, faziam a coleta a partir de um *web crawler* mas, em 2022, passaram a utilizar a API disponibilizada pelo Twitter. Recomenda-se a leitura do artigo [Heine et al. 2021] para maiores detalhes sobre a ferramenta. Além de alunos e professores da PUC-Rio, a eTC tem usuários de várias universidades e instituições de pesquisas no país.

Naturalmente, os usuários da eTC também têm necessidade de suporte em etapas de pré-processamento e análises posteriores à coleta. Motivados por casos reais de demandas recebidas, foram aplicados novos procedimentos para enriquecer as amostras de postagens (*tweets*) (1) representando os relacionamentos entre usuários no formato de rede ao identificar comunidades de usuários que publicam e seguem outros usuários; e (2) extraindo os tópicos para identificar os assuntos mais relevantes dos *datasets*.

Neste artigo, são detalhados alguns desses enriquecimentos de *datasets* coletados pela eTC do ponto de vista computacional. Ao final, serão dados exemplos reais de casos de uso com temas de vacinação e política no Brasil, pela atenção e questionamentos dados a esses temas nos últimos anos.

2. Trabalhos Relacionados

Atualmente, há várias ferramentas para coleta e análise de dados de redes sociais. Muitas delas, principalmente algumas das comerciais como Brandwatch e Talkwalker, apresentam semelhanças significativas entre si e não são adequadas para atender a buscas e coletas de dados para fins de pesquisa em torno de temas específicos, pois se concentram principalmente na análise das próprias contas dos usuários de seus clientes nas redes sociais.

Assim, ao buscar por outras ferramentas para coleta e análise de dados de redes sociais, com objetivos propostos mais similares à ferramenta eTC, foram encontrados o SOCRATES[2] e o Netlytic[3]. No que se refere às análises, o SOCRATES, por meio dos dados importados ou coletados pela ferramenta, permite fazer algumas análises como contagem de palavras, análise de sentimentos e análises estatísticas com a combinação dos campos de postagens. Ainda assim, apresenta erros no tratamento de volumes de dados na faixa de algumas dezenas de milhares de linhas, o que limita a usabilidade da ferramenta. Quanto ao Netlytic, é possível analisar o texto por nuvens de palavras e análise categórica, além de realizar análises de redes. Para a análise categórica, é necessário montar um dicionário de palavras relacionadas a uma determinada categoria. Quanto à análise de redes, eles utilizam redes dos nome de usuários e quem os referencia e redes de referência direta (”quem responde a quem”), gerando *clusters* automaticamente. Mesmo assim, a ferramenta tem limites em torno de 10 mil registros para uso livre por *dataset*.

No eTC, é possível realizar essas tarefas para um volume maior, em torno de 100 mil registros para contas comuns. Apesar disso, com nossos parceiros realizamos também análises maiores de forma automática ou com algum apoio manual para utilizar técnicas de redução de dados em *datasets* de milhões de registros, a partir da importância deles para as análises desejadas. Assim, a ferramenta do eTC ainda é relevante na área de

[2]https://socrates.peopleanalytics.org/

[3]https://netlytic.org/

coleta e análise do conteúdo de redes sociais, buscando aderir novas tecnologias, como transferência de aprendizado, e outras formas de visualização dos dados para agregar ainda mais valor à ferramenta em relação às demais concorrentes.

3. Enriquecimento e Análise de Conjunto de Postagens

Nesta seção são explicadas algumas atividades relacionadas ao enriquecimento de *datasets* obtidos pela coleta *tweets*, visando incrementar o conjuntos de informações relacionadas aos textos por meio de identificação de tópicos tratados pelos usuários e de linguagem ofensiva. Detalha-se também o processo de construção de redes de perfis de usuários usando mecanismos tradicionais devidamente customizados e implementados.

3.1. Análise de rede

Em relação à classificação das comunidades de usuários, foram consideradas algumas metodologias existentes, decidindo-se, por fim, montar uma rede de *retweets* usando o programa de análise e visualização de redes Gephi[4] e, posteriormente, usar o método de Louvain [Blondel et al. 2008] para a detecção de comunidades, assim como em [Cherepnalkoski and Mozetic 2015], [Gargiulo et al. 2020] e [Novak et al. 2018].

Retweet, no contexto do Twitter, significa compartilhar um *tweet* postado por outra conta. Assim, terão-se arestas partindo do usuário que fez o *retweet* em direção ao usuário que postou o *tweet* original. Se um usuário *retweetou* outro múltiplas vezes, a aresta terá um peso maior, proporcional a este número. A seguir, serão detalhadas as etapas envolvidas na montagem e análise desta rede.

O primeiro passo é a elaboração do grafo usando o Gephi. Para isso, é necessário realizar mais um pré-processamento, do *dataset* de *tweets* coletados, de forma manual, a fim de se obter a informação do autor do *tweet* original, ao qual um *retweet* se refere. Para isso, foi desenvolvido um *script* que itera sobre todos os *retweets* e extrai, a partir do seu campo de texto, o nome do usuário que postou o original. Assim, a importação para o Gephi na forma de uma planilha *source-target* pode ser feita, em que a primeira coluna se refere ao nó de origem e a segunda, ao destino.

Após a importação dos dados e construção do grafo no Gephi, podem-se utilizar as funcionalidades do programa para calcular métricas para os nós da rede - usam-se, como exemplo, grau de entrada, grau de saída e *pagerank*[Page et al. 1999]. Em seguida, utiliza-se o método de Louvain, do próprio Gephi, para a detecção das comunidades. Com isso, obtem-se comunidades distintas, cada nó do grafo - isto é, cada usuário - pertencendo a uma delas. Feito isso, são exportados tais dados calculados pelo programa para uma nova tabela do banco de dados da eTC, para que possa se realizar consultas relevantes às pesquisas dos usuários do sistema.

3.2. Modelagem de Tópicos

Além da análise de redes, outro tipo de análise que pode ser feita é a modelagem de tópicos. Nela são extraídos os tópicos de um conjunto de textos, a partir da clusterização das palavras. Para isso, primeiro é realizada uma etapa de tratamento do texto, na qual se elimina caracteres especiais, números e *links* por expressão regular e as *stop words*

[4] `https://gephi.org/`

com base na biblioteca *nltk*. Com isso, utiliza-se de algum meio de extração de *features* dos textos processados como o TF-IDF (Term Frequency - Inverse Document Frequency). Por último, o algoritmo assume que cada documento é uma mistura de múltiplos tópicos e que cada tópico é a distribuição de probabilidade sobre as palavras. Logo, o resultado são vários *clusters* de palavras, em que é necessário realizar uma interpretação para inferir o que aquele tópico representa. Com base nisso, neste trabalho, foram estudados o LDA (Latent Dirichlet Allocation) [Chauhan and Shah 2021] e o BERTopic [Grootendorst 2022], ambos algoritmos para modelagem de tópicos para o enriquecimento de dados coletados de redes sociais.

A abordagem do algoritmo de LDA envolve tratar documentos como misturas de tópicos, onde cada tópico é composto de palavras-chave. Ao fornecer o número de tópicos desejado, eles são reorganizados para construir uma distribuição de tópicos com suas respectivas palavras-chave. Para utilizar o método, foram utilizados os pacotes *Gensim*, *Spacy* e *pyLDAVis* para pré-processamento dos textos, *tokenização* e visualização dos dados obtidos. Por outro lado, o método BERTopic adota uma variante de TF-IDF orientada a classes, modelando a importância das palavras em relação aos *clusters*, ao invés de documentos individuais. Também, é empregado o Sentence-BERT [Reimers and Gurevych 2019] para *embedding* de documentos, o qual se baseia na rede neural BERT (Bidirectional Encoder Representations for Transformers) [Devlin et al. 2019]. Assim, a modelagem de tópicos incorpora semântica e contexto, considerando suas relações com os *clusters*. Neste trabalho, foi utilizado o pacote BERTopic com opção multilíngue e ajustado quanto a variação no tamanho mínimo do *cluster*, para verificar sua influência na qualidade e abrangência dos tópicos extraídos.

4. Casos de Uso e Resultados

Com o intuito de avaliar as funcionalidades para enriquecimento de texto detalhadas na Seção 3, foram utilizados dois *datasets* em português como caso de uso, obtidos por meio de buscas no eTC utilizando a API Twitter, sobre os seguintes temas: vacinação durante o período de imunização infantil contra a Covid-19 no Brasil; e respostas a mulheres na política durante o primeiro turno das eleições de 2022.

4.1. Análise de Redes em Dados de Vacinação no Período de Imunização Infantil contra a Covid-19 no Brasil

Para análise de redes, foram coletados dados sobre este tema de vacinação em um intervalo de 62 dias, entre 09/12/2021 e 09/02/2022, foram coletados 5,766,083 tweets de 986,048 usuários diferentes. Dos tweets coletados na busca mencionada, 4,052,198 (70,3%) são retweets que se referem a 674,795 usuários diferentes. Incluindo também os tweets originados por esses usuários, cerca de 89,8% do total coletado pertence, portanto, a usuários que foram analisados na rede. Em relação às 5938 classes de modularidade obtidas, destacam-se os dois grupos com o maior número de tweets. O primeiro grupo possui 2,022,270 (39,0%) de tweets e retweets associados a usuários classificados com esse grupo. Já o segundo grupo possui 1,978,426 (38,2%) do total.

Os resultados apresentados na tabela 1 reafirmam o que era esperado pela detecção das comunidades da rede. Em primeiro lugar, há coesão entre os usuários apresentados de cada grupo, já que o grupo 1 possui personalidades que se colocam a favor da vacina

Nome de Usuário	Grau de Entrada
@oatila	31,081
@PedroHallal	14,079
@PedroRonchi2	12,008
@dadourado	11,589
@ThiagoResiste	11,030

Nome de Usuário	Grau de Entrada
@GFiuza_Oficial	24,405
@Rconstantino	16,998
@Biakicis	14,542
@OsmarTerra	14,294
@revistaoeste	13,493

Tabela 1. TOP 5 usuários por grau de entrada dos grupos 1 e 2, respectivamente.

enquanto o grupo 2 se refere principalmente a figuras contrárias. Em segundo lugar, nota-se que mesmo com diversas classes de modularidades, os dois grupos com mais *tweets* apresentam opiniões opostas sobre o tema, sinalizando a polarização da discussão.

4.2. Modelagem de Tópicos em Dados de Respostas Diretas a Mulheres na Política durante o 1^o Turno das Eleições de 2022

Quanto à modelagem de tópicos, foi utilizado o *dataset* de 933,080 *tweets* sobre o tema em questão com mulheres candidatas ao senado ou ex-senadoras em um intervalo de 4 meses, entre 06/04/2022 a 06/10/2022, o que abrange todo o primeiro turno de eleições. Esses *tweets* contém tanto as *replies* quanto as postagens originais dessas mulheres.

Com o propósito de obter bons resultados para ambos modelos, foi feito o *fine-tuning* de seus hiperparâmetros com semente de aleatoriedade fixada. No caso do LDA, o modelo foi treinado diversas vezes para se obter os valores de coerência e, assim, decidir qual número de tópicos teria um melhor resultado para esse modelo, por fim, concluindo-se que esse número seriam 20 tópicos. Quanto ao BERTopic, foi estudado o tamanho mínimo do *cluster*, variando o valor entre 10 e 750. A partir disso, para o BERTopic, identificou-se que o melhor tamanho mínimo era de 500 e tópicos até o 26 representavam agrupamentos de tamanho de pelo menos 0,5% do tamanho total do *dataset*. Desse modo, foram comparados os resultados alcançados por ambos os modelos. A partir das Figuras 1 e 2 pôde-se perceber pelas distâncias inter-tópico, que o BERTopic apresentou melhores resultados que o LDA, pois a distribuição dos tópicos no espaço é mais esparsa. O mesmo pôde-se concluir pela análise individual dos tópicos identificados, comprovando-se que o BERTopic era o método mais apropriado.

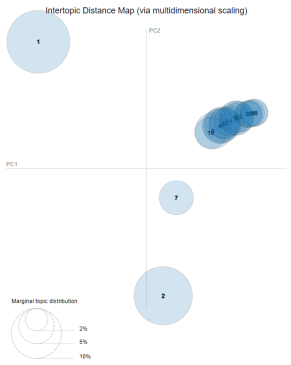

Figura 1. Mapa de Distância entre Tópicos para o LDA

Figura 2. Mapa de Distância entre Tópicos para o BERTopic

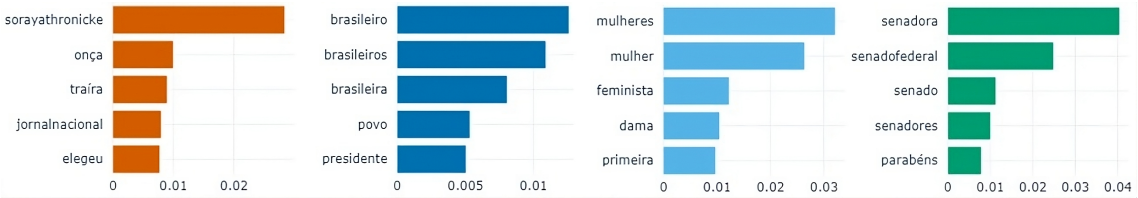

Figura 3. Alguns dos Principais Tópicos Identificados pelo BERTopic

Pela Figura 3, pode-se observar um tópico sobre a candidata à presidência Soraya Thronicke, que durante o período de eleições foi chamada de ”traíra” por apoiadores do presidente Jair Bolsonaro, o que explica a presença dessa e a outra ofensa. Além disso, também observa-se um tópico voltado mais ao povo brasileiro, inclusive a palavra presidente, o que faz menção ao contexto eleitoral. Como esperado da busca, há um tópico voltado para as mulheres em si. Por último, há um tópico para o senado, o que pode ser explicado por uma das mulheres com maior menção no *dataset* ser a ex-senadora Simone Tebet. Apesar de terem sido encontrados outros tópicos relacionados ao cristianismo, à própria senadora Simone Tebet e a candidatas apoiadoras do ex-presidente da República durante as eleições, não será possível explicá-los detalhadamente neste artigo.

5. Considerações Finais

Neste trabalho, foram apresentadas as metodologias utilizadas para enriquecer os dados retornados pela API do Twitter e para realizar análises preliminares que possam apoiar nossos parceiros de pesquisa a realizarem seus estudos. Foram realizadas operações para transformar os dados em uma rede de *retweets*, identificando, assim comunidades distintas de usuários e adicionamos um procedimento para averiguar quais os tópicos mais falados dentro de uma determinada comunidade.

Vale citar que as operações descritas neste artigo serão utilizadas em pelo menos duas linhas de pesquisa no futuro próximo. Há uma parceria em andamento com pesquisadoras da Fundação Oswaldo Cruz (Fiocruz) que visa a investigar as narrativas e os agentes envolvidos na discussão sobre vacinação no Brasil [Verjovsky et al. 2023], além de um estudo da área da comunicação política que já foi iniciado, comparando as mídias compartilhadas por apoiadores de candidatos durante as últimas eleições. Além disso, a modelagem de tópicos será utilizada para juntamente a uma análise de toxicidade em comentários feitos a candidatas femininas brasileiras durante a época das eleições de 2022, um projeto em parceria com a Universidade de Pittsburgh.

Como é possível notar, as análises realizadas com os dados coletados pela ferramenta estão em constante evolução. Uma das principais atualizações planejadas é a incorporação do BERT na análise de sentimentos, que atualmente está utilizando o modelo MLP (*Multilayer Perceptron*). Por fim, em relação à análise de redes, planeja-se passar a usar a biblioteca NetworkX [5] no lugar do programa Gephi, por ser uma ferramenta programática e com mais opções.

[5] `https://networkx.org/`

Referências

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Chauhan, U. and Shah, A. (2021). Topic modeling using latent dirichlet allocation: A survey. *ACM Comput. Surv.*, 54(7).

Cherepnalkoski, D. and Mozetic, I. (2015). A retweet network analysis of the european parliament. In *Procs. 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 350–357.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gargiulo, F., Cafiero, F., Guille-Escuret, P., Seror, V., and Ward, J. K. (2020). Asymmetric participation of defenders and critics of vaccines to debates on french-speaking twitter. *Scientific reports*, 10(1):1–12.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Heine, A., Coutinho, B., Barreto, M., Xavier, N., Villas, M., Ituassu, A., and Lifschitz, S. (2021). Análise de dados para comunicação política a partir de um sistema de coleta de tweets. In *Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 49–55, Porto Alegre, RS, Brasil. SBC.

Novak, P. K., Amicis, L. D., and Mozetič, I. (2018). Impact investing market on twitter: influential users and communities. *Applied network science*, 3(1):1–20.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

Verjovsky, M., Barreto, M. P., Carmo, I., Coutinho, B., Thomer, L., Lifschitz, S., and Jurberg, C. (2023). Political quarrel overshadows vaccination advocacy: How the vaccine debate on brazilian twitter was framed by anti-vaxxers during bolsonaro government.