



Reddit Comment Analysis: Sentiment Prediction and Topic Modeling using VADER and BERTopic

Denilson de Oliveira Silva ^{1*},
Richard Matheus Avelino da Silva ²,
Patrícia Virgínia de Santana Lima ³,
Jéssica Cristina Pereira Batista ⁴,
Sílvio Fernando Alves Xavier ⁵,



^{1 2 3 4 5} Departamento de Estatística, State University of Paraíba, Campina Grande, Brazil

Emails: ¹ denilson.oliveira.silva@aluno.uepb.edu.br; ² richard.silva@aluno.uepb.edu.br;
³ patricia.lima@aluno.uepb.edu.br; ⁴ jessica.batista@aluno.uepb.edu.br; ⁵ silvio@servidor.uepb.edu.br.

* Corresponding author

How to cite this paper: Silva, D. de Oliveira, da Silva, R. M. A., Lima, P. V. de Santana, Batista, J. C. P., Xavier, S. F. A. (2024). Reddit Comment Analysis: Sentiment Prediction and Topic Modeling using VADER and BERTopic. *Socioeconomic Analytics*, 2(1), 130-136. <https://doi.org/10.51359/2965-4661.2024.265074>

RESEARCH ARTICLE

Socioeconomic Analytics
<https://periodicos.ufpe.br/revistas/SECAN/>
ISSN Online: 2965-4661

Submitted on: 11.11.2024
Accepted on: 15.12.2024.
Published on: 27.12.2024.

Copyright © 2024 by author(s).

This work is licensed under the Creative Commons Attribution International License CC BY-NC-ND 4.0
<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>



Abstract

This work aims at exploring data analysis techniques applied to the social media platform Reddit, highlighting the execution of an Exploratory Data Analysis (EDA) to identify trends and patterns of interaction among users. For sentiment analysis of the comments, the VADER model ("Valence Aware Dictionary and Sentiment Reasoner") is used, and topic modeling is performed with BERTopic ("Bidirectional Encoder Representations from Transformers for Topic Modeling"). The goal is to compare the accuracy and effectiveness of the models in classifying emotions and themes expressed in the comments. The comparison of the models allows identifying which approach yields the most accurate results, which is aligned with the context of discussions on Reddit, providing valuable insights into user behavior and preferences.

Keywords

Sentiment Analysis, Text Mining, Topic Modelling, Exploratory Data Analysis, Reddit, VADER, BERTopic.

1. Introdução

O crescimento das redes sociais como plataformas de comunicação tem gerado uma vasta quantidade de dados textuais que oferecem informações valiosas sobre as emoções e preferências dos usuários (Kouloumpis et al., 2011; Giachanou & Crestani, 2016). Dentre essas plataformas, o Reddit se destaca por ser um espaço de discussão diversificado, onde os usuários compartilham opiniões e experiências em uma variedade de tópicos, desde notícias e entretenimento até suporte e conselhos pessoais (Medvedev et al., 2019). Nesse contexto, a análise de sentimentos tornou-se uma ferramenta essencial para decifrar as particularidades emocionais das postagens e compreender a atmosfera das interações na plataforma (Hutto & Gilbert, 2014).

O Reddit foi criado em 2005 por Steve Huffman e Alexis Ohanian como um site agregador de notícias e espaço para discussão comunitária. Desde então, a plataforma evoluiu para se tornar um dos principais fóruns da internet, organizando conteúdo em comunidades específicas conhecidas como subreddits. Essas comunidades abrangem uma ampla gama de tópicos e são mantidas por moderadores voluntários, o que contribui para a rica diversidade de opiniões e informações disponíveis (Massanari, 2015). A natureza interativa e aberta do Reddit o torna um ambiente ideal para estudos que buscam compreender tendências sociais, emoções e opiniões expressas por seus usuários.

Foram utilizadas duas ferramentas amplamente usadas para análise de dados textuais em redes sociais são o VADER (Valence Aware Dictionary and sEntiment Reasoner) e o BERTopic. O VADER foi desenvolvido por Hutto e Gilbert (2014) e é um algoritmo de análise de sentimentos que combina regras baseadas em lexicon com aprendizado de máquina, sendo altamente eficaz para textos curtos e informais, como postagens em redes sociais. Por outro lado, o BERTopic, criado por Grootendorst (2020), é uma ferramenta baseada em aprendizado de máquina que utiliza embeddings de linguagem e técnicas de clustering para extrair tópicos de grandes volumes de texto. Ambas as ferramentas têm sido amplamente adotadas em estudos acadêmicos e aplicações comerciais para entender e interpretar grandes conjuntos de dados textuais de forma eficiente.

2. Objetivo

Este trabalho teve como objetivos explorar o uso de técnicas de análise de dados aplicadas ao Reddit, por meio de uma análise exploratória de dados (EDA), a fim de identificar padrões de interação entre os usuários, temas mais recorrentes, além de analisar a distribuição de popularidade dos tópicos e a frequência de comentários. Além disso, buscou-se explorar a modelagem de tópicos abordados nos comentários da rede social, utilizando o BERTopic, sua utilização permitiu a extração de tópicos latentes e facilitou a comparação de diferentes modelos de tópicos, contribuindo para uma análise mais precisa e dinâmica do conteúdo textual gerado pelos usuários no Reddit.

3. Metodologia

3.1. Material

A base de dados utilizada neste estudo foi obtida da plataforma Kaggle, contendo 1 milhão de comentários coletados de 40 subreddits diferentes. Cada subreddit possui 25 mil comentários,

totalizando a divisão dos dados em 40 grupos distintos. A base de dados é composta por quatro colunas, cujas informações incluem o comentário em si e outras variáveis associadas ao conteúdo. A análise foi conduzida com foco no processamento de texto e na identificação de padrões de sentimentos presentes nos comentários

3.2. Métodos

O presente estudo explora como técnicas de análise de dados podem revelar padrões e sentimentos nas interações dos usuários do Reddit. A metodologia foi dividida em diversas etapas que envolveram a coleta de dados, análise exploratória e aplicação de modelos de análise de sentimentos e modelagem de tópicos. Inicialmente, realizamos uma Análise Exploratória de Dados (EDA) para obter uma compreensão inicial dos dados. Durante esta etapa, examinamos o volume de comentários em diferentes subreddits, identificando os padrões mais frequentes de participação dos usuários. Em paralelo, fizemos uma análise das palavras-chave mais comuns para determinar os tópicos predominantes de interesse. Isso nos trouxe uma visão geral de como os usuários se comportam na plataforma, quais assuntos geram mais engajamento e quais são os principais temas de discussão.

Como parte desta análise, utilizamos uma nuvem de palavras para identificar rapidamente as palavras mais frequentes em um conjunto de comentários do Reddit. A nuvem destacou visualmente os termos mais utilizados, permitindo perceber os temas predominantes e fornecendo pistas sobre o sentimento geral ou o tom das interações, como emoções positivas ou negativas. Essa ferramenta foi mostrada útil para orientar análises mais elaboradas, como a de sentimentos. Em seguida, a fim de capturar as emoções expressas nos textos dos usuários, aplicamos o modelo de análise de sentimentos “VADER”. Esse tipo de análise foi escolhido para ser um modelo de fácil implementação, projetado especificamente para textos curtos e informativos, como os encontrados nas redes sociais, e para ser capaz de capturar nuances como sarcasmo e gírias.

Já o BERTopic, baseado no modelo BERT (Devlin et al., 2018), foi utilizado para realizar a modelagem de tópicos, agrupando postagens com base em seus temas latentes e permitindo identificar as categorias principais de discussão dentro do Reddit. Finalmente, os resultados obtidos na análise exploratória de dados, na modelagem de tópicos e na análise de sentimentos foram integrados para gerar uma visão mais completa das dinâmicas de interação no Reddit. Essa abordagem nos permitiu identificar não apenas os temas mais planejados, mas também o tom emocional das conversas, fornecendo detalhes sobre as preferências e comportamentos dos usuários da plataforma.

4. Resultados e discussões

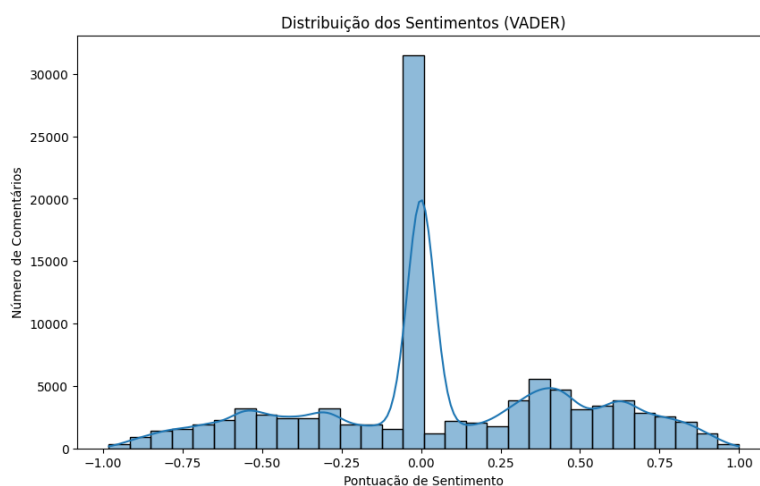
Esta seção detalha os principais resultados encontrados, oferecendo uma visão aprofundada sobre os padrões identificados, os sentimentos predominantes e os tópicos mais discutidos nas interações dos usuários do Reddit.

Figura 1: (A) Distribuição de Frequência de palavras; Gráfico (B) Nuvem de palavras.

No gráfico A Mostra a matriz de similaridade entre os textos, representada por um mapa de calor. Cada célula da matriz indica o grau de similaridade entre dois textos, com valores que vão de 0 (nenhuma similaridade) a 1 (máxima similaridade). As cores mais escuras representam maior similaridade, enquanto as mais claras indicam menor relação entre os textos. Isso permite identificar quais textos abordam temas semelhantes ou possuem estruturas parecidas.

Já no gráfico B, cada barra representa as palavras mais comuns em diferentes tópicos, como esportes, filmes ou jogos. As barras são divididas de acordo com o sentimento das palavras, mostrando se elas têm um tom positivo, negativo ou neutro. Se a barra estiver mais para o lado positivo, significa que as palavras naquele tópico têm um tom otimista ou feliz. Se estiver mais para o lado negativo, as palavras têm um tom crítico ou triste. Quando a barra fica no meio, as palavras são mais neutras, sem um sentimento forte. Esse gráfico ajuda a entender, para cada tópico, se as pessoas estão falando de maneira mais positiva, negativa ou neutra.

Figura 3: Análise da distribuição dos sentimentos com Vader.



Fonte elaborado pelo autores (2024)

Observamos que a maior parte dos comentários analisados tem um tom neutro, ou seja, não expressam emoções fortes, nem positivas nem negativas. Isso significa que os usuários usaram palavras mais equilibradas, sem demonstrar sentimentos intensos. Há alguns comentários que são mais emocionais, com tom positivo ou negativo, mas eles aparecem com menos frequência. No geral, os comentários são mais objetivos e diretos, sem transmitir opiniões muito carregadas de emoção. A neutralidade é o que mais se destaca nessa análise.

5. Conclusão

Nesta análise de Processamento de Linguagem Natural, aplicada aos comentários do Reddit, conseguimos extrair informações valiosas sobre o comportamento dos usuários. A partir da Exploração de Dados, identificamos tendências importantes na distribuição de *subreddits* e na polarização dos comentários, destacando variações relevantes em relação ao score e à controvérsia. A análise de sentimentos, realizada com o VADER, indicou que a maioria dos comentários expressa um sentimento neutro, enquanto as opiniões polarizadas (positivas e negativas) aparecem com menor frequência, sugerindo uma postura equilibrada dos usuários nas interações.

A modelagem de tópicos, feita por meio do *BERTopic*, permitiu identificar os principais temas discutidos, evidenciando a diversidade de tópicos nos comentários. A combinação dessas técnicas forneceu uma visão profunda dos padrões de discussão. Dessa forma, a análise oferece uma compreensão ampla e detalhada do comportamento discursivo e emocional da comunidade no *Reddit*, contribuindo para estudos futuros sobre dinâmica de interação em plataformas de mídia social. Esses futuros estudos podem explorar temas como política, esportes e divulgação do dia a dia no *Reddit*. Na política, é possível analisar a polarização ideológica, a propagação de notícias falsas e o impacto de eventos como eleições nas interações. Nos esportes, a dinâmica das torcidas, o engajamento em eventos como a Copa do Mundo e as discussões sobre atletas podem revelar padrões emocionais e sociais. Já nas discussões diárias, tópicos como trabalho, saúde mental e consumo podem ser investigados, focando no tom emocional, nas tendências de comportamento e no apoio mútuo mútuo em comunidades online.

Acknowledgements

Nossos sinceros agradecimentos vão, primeiramente, ao evento 'Wanda 2024' e aos seus organizadores, cujos compromissos e atribuições foram fundamentais para a realização deste trabalho.

Conflict of Interest Declaration

Os autores declaram não possuir conflitos de interesse. Todos os coautores concordam integralmente com o conteúdo do manuscrito, não tendo interesses financeiros a serem divulgados.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
2. Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2), 1-41.
3. Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1).
4. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.
5. Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis (Vol. 5)*. Cham: Springer International Publishing.
6. Reddit. (n.d.). Reddit. <https://www.reddit.com>
7. Magnan, S. (2024). 1 million Reddit comments from 40 subreddits. Kaggle. <https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits>, (acesso em 11 novembro de 2024).
8. Danushi, D. (2021). Topic modelling with BERTopic. Medium. <https://medium.com/@danushidk507/topic-modelling-with-bertopic-249095144555> (acesso em 11 de novembro de 2024)

9. Medvedev, A., Silva, B., & Pereira, C. (2019). Análise de sentimentos em plataformas digitais: Estudo no Reddit. *Revista Brasileira de Tecnologia*, 15(3), 25-40.
10. Massanari, A. L. (2015). Networks of race and gender: Black women's online presence and the role of social media in the fight for racial justice. [PhD dissertation, University of Minnesota]. ProQuest Dissertations & Theses Global.