SOCIOECONOMIC ANALYTICS

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.eupb.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.

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. Introduction

The growth of social networks as communication platforms has generated a vast amount of textual data that offers valuable information about users emotions and preferences (Kouloumpis *et al.*, 2011; Giachanou & Crestani, 2016). Among these platforms, Reddit stands out for being a diverse discussion space, where users share opinions and experiences on a variety of topics, from news and entertainment to personal support and advice (Medvedev *et al.*, 2019). In this context, sentiment analysis has become an essential tool for deciphering the emotional particularities of posts and understanding the atmosphere of interactions on the platform (Hutto & Gilbert, 2014).

Reddit was created in 2005 by Steve Huffman and Alexis Ohanian as a news aggregator site and space for community discussion. Since then, the platform has evolved to become one of the leading forums on the internet, organizing content into specific communities known as subreddits. These communities cover a wide range of topics and are maintained by volunteer moderators, which contributes to the rich diversity of opinions and information available (Massanari, 2015). The interactive and open nature of Reddit makes it an ideal environment for studies that seek to understand social trends, emotions, and opinions expressed by its users.

Two widely used tools were used for textual data analysis in social networks are VADER (Valence Aware Dictionary and Entiment Reasoner) and BERTopic. VADER was developed by Hutto and Gilbert (2014) and is a sentiment analysis algorithm that combines lexicon-based rules with machine learning, being highly effective for short and informal texts, such as social media posts. On the other hand, BERTopic, created by Grootendorst (2020), is a machine learning-based tool that uses language embeddings and clustering techniques to extract topics from large volumes of text. Both tools have been widely adopted in academic studies and commercial applications to understand and interpret large textual datasets efficiently.

2. Objective

This work aimed to explore the use of data analysis techniques applied to Reddit, through an exploratory data analysis (EDA), in order to identify patterns of interaction between users, most recurrent themes, in addition to analyzing the distribution of popularity of topics and the frequency of comments. In addition, it was sought to explore the modeling of topics addressed in the comments of the social network, using BERTopic, its use allowed the extraction of latent topics and facilitated the comparison of different topic models, contributing to a more accurate and dynamic analysis of the textual content generated by users on Reddit.

3. Methodology

3.1. Material

The database used in this study was obtained from the Kaggle platform, containing 1 million comments collected from 40 different subreddits. Each subreddit has 25 comments, totaling the division of data into 40 distinct groups. The database consists of four columns, whose information includes the comment itself and other variables associated with the content. The analysis was conducted with a focus on word processing and the identification of sentiment patterns present in the comments.

3.2. Methods

The current study explores how data analysis techniques can reveal patterns and sentiments in Reddit users interactions. The methodology was divided into several stages that involved data collection, exploratory analysis and application of sentiment analysis models and topic modeling. Initially, we conducted an Exploratory Data Analysis (EDA) to gain an initial understanding of the data. During this step, we examined the volume of comments on different subreddits, identifying the most frequent patterns of user participation. In parallel, we did an analysis of the most common keywords to determine the predominant topics of interest. This gave us an overview of how users behave on the platform, which topics generate the most engagement, and what the main topics of discussion are.

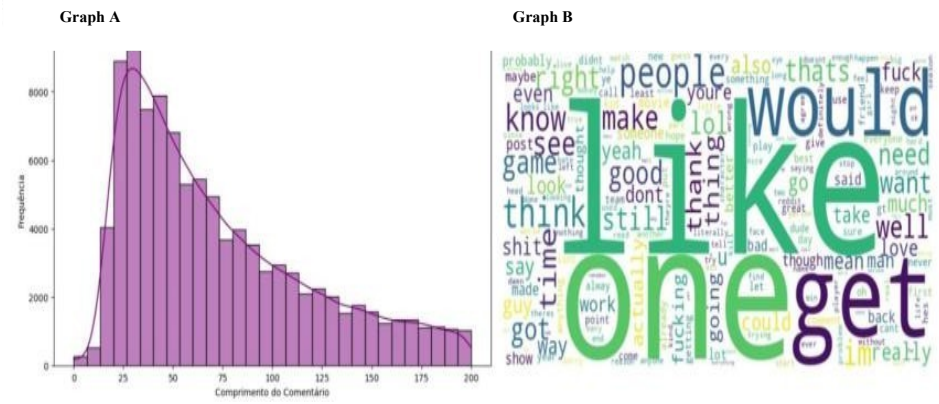
As part of this analysis, we used a word cloud to quickly identify the most frequent words in a set of Reddit comments. The cloud visually highlighted the most commonly used terms, allowing them to perceive the prevailing themes and providing clues about the overall sentiment or tone of interactions, such as positive or negative emotions. This tool has been shown to be useful in guiding more elaborate analyses, such as sentiment. Then, in order to capture the emotions expressed in the users' texts, we applied the “VADER” sentiment analysis model. This type of analysis was chosen to be an easy-to-implement model, designed specifically for short, informative texts, such as those found on social media, and to be able to capture nuances such as sarcasm and slang.

BERTopic, on the other hand, based on the BERT model (Devlin *et al.*, 2018), was used to perform topic modeling, grouping posts based on their latent themes and allowing the identification of the main categories of discussion within Reddit. Finally, the results obtained from exploratory data analysis, topic modeling, and sentiment analysis were integrated to generate a more complete picture of interaction dynamics on Reddit. This approach allowed us to identify not only the most planned themes, but also the emotional tone of the conversations, providing details about the preferences and behaviors of the platform's users.

4. Results and discussions

This section details the main results found, offering an in-depth look at the patterns identified, the prevailing sentiments, and the most discussed topics in Reddit users interactions.

Figure 1: (A) Frequency Distribution of Words; Graph (B) Word Cloud.



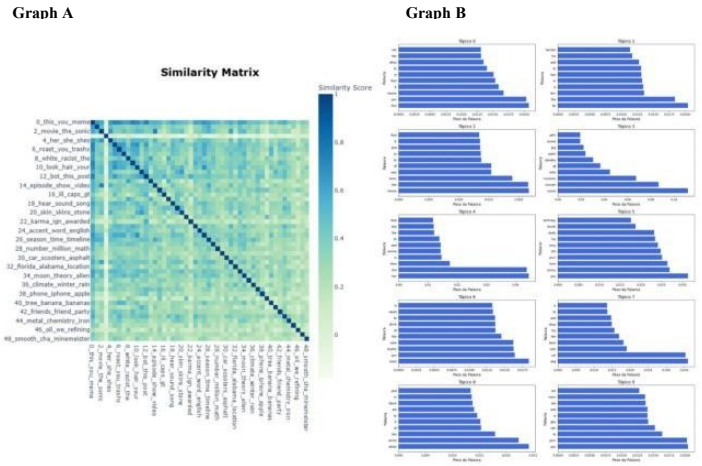
Source: Prepared by the authors (2024).

We observed that Graph A shows the distribution of the length of the comments, considering only those with up to 200 characters. On the X-axis, we have the length of comments, which ranges from 0 to 200 characters, while the Y-axis shows the frequency of comments for each length range. The highest frequency of comments is in the range between 25 and 30 characters, with more than 8,000 occurrences, which characterizes an evident peak. The distribution is asymmetrical, with a high concentration of short comments, while the frequency gradually decreases as the length increases, suggesting that most users opt for more objective and direct messages.

Graph B presents a word cloud in which the larger terms represent the most frequent ones in the comment base. Words such as “like”, “get”, “would” and “one”, indicating that they have a central role in the construction of discourse. This suggests that many messages revolve around expressions of preference, desire, actions, or references to something specific.

The presence of these words, together with the absence of other more technical or formal ones, reinforces that the analyzed base may be linked to everyday debates, outbursts or casual conversations, possibly on online platforms aimed at social interaction or communities. The centrality of these words demonstrates not only the most addressed topics, but also the predominant style and tone of communication.

Figure 2: Topic modeling by BertTopics.

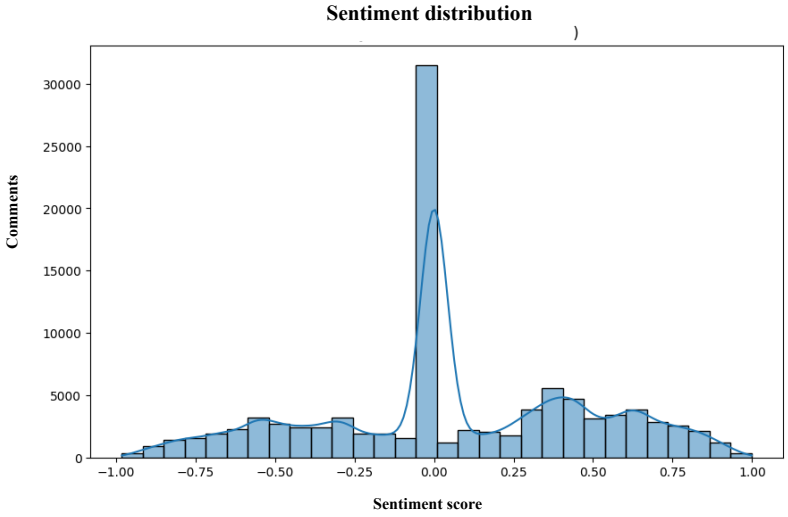


Source: Prepared by the authors (2024).

Graph A shows the similarity matrix between the texts, represented by a heat map. Each cell of the matrix indicates the degree of similarity between two texts, with values ranging from 0 (no similarity) to 1 (maximum similarity). Darker colors represent greater similarity, while lighter colors indicate less relationship between texts. This allows you to identify which texts address similar themes or have similar structures.

In Graph B, each bar represents the most common words on different topics, such as sports, movies, or games. The bars are divided according to the feeling of the words, showing whether they have a positive, negative, or neutral tone. If the bar is more to the positive side, it means that the words in that thread have an optimistic or happy tone. If it's more on the negative side, the words have a critical or sad tone. When the bar is in the middle, the words are more neutral, without a strong feeling. This graph helps you understand, for each topic, whether people are speaking in a more positive, negative, or neutral way.

Figure 3: Analysis of the distribution of feelings with Vader.



Source: Prepared by the authors (2024).

We observed that most of the comments analyzed have a neutral tone, that is, they do not express strong emotions, neither positive nor negative. This means that users used more balanced words, without showing intense feelings. There are some comments that are more emotional, with a positive or negative tone, but they appear less frequently. In general, the comments are more objective and direct, without conveying opinions that are too emotionally charged. Neutrality is what stands out the most in this analysis.

5. Conclusion

In this Natural Language Processing analysis, applied to Reddit comments, we were able to extract valuable insights into users behavior. From Data Exploration, we identified important trends in the distribution of subreddits and the polarization of comments, highlighting relevant variations in score and controversy. The sentiment analysis, carried out with VADER, indicated that most comments express a neutral feeling, while polarized opinions (positive and negative) appear less frequently, suggesting a balanced posture of users in interactions.

The topic modeling, done through BERTopic, allowed the identification of the main topics discussed, evidencing the diversity of topics in the comments. The combination of these techniques provided deep insight into the patterns of discussion. In this way, the analysis offers a broad and detailed understanding of the discursive and emotional behavior of the community on Reddit, contributing to future studies on interaction dynamics on social media platforms.

These future studies can explore topics such as politics, sports, and day-to-day outreach on Reddit. In politics, it is possible to analyze ideological polarization, the spread of fake news, and the impact of events such as elections on interactions. In sports, crowd dynamics, engagement in events such as the World Cup, and discussions about athletes can reveal emotional and social patterns. In daily discussions, topics such as work, mental health, and consumption can be investigated, focusing on emotional tone, behavioral trends, and mutual support in online communities.

Acknowledgements

Our sincere thanks go, first, to the “Wanda 2024” event and its organizers, whose commitments and attributions were fundamental to the realization of this work.

Conflict of Interest Declaration

The authors declare no conflicts of interest. All co-authors fully agree with the content of the manuscript, having no financial interests to be disclosed.

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.