Sentiment Analysis on 𝕏 (formerly known as Twitter) Data (U.S. 2024 Election)

Giovanni Vurro

Introduction

This project explores whether sentiment analysis of Twitter data can effectively capture voter sentiment during elections by comparing it to traditional polling methods. By using NLP tools, the analysis focuses on Twitter conversations surrounding key figures, events, and topics during the 2024 United States Election. The project examines how well sentiment analysis reflects voter attitudes and whether it provides insights that align with or diverge from traditional polling results, offering a new perspective on how to approach public discourse during elections. Ultimately, the project seeks to answer the question: *Can sentiment analysis of Twitter data capture voter sentiment during elections compared to traditional polling methods?*

Related Work

Before pursuing this project using HuggingFace's BERT models I decided to research if similar projects had been conducted and the course they chose to take with their implementation. I found a paper describing sentiment analysis conducted for the Coronavirus Outbreak and did a comparative experiment of VADER vs. BERT [1]. They found that VADER was efficient for real-time applications, but it struggled to handle nuanced language, sarcasm, and mixed sentiments. HuggingFace's BERT model on the other hand was more equipped to handle the complexities involved in political discourse which is why I chose that over VADER. Being a transformer model it was also ideal for instances where unstructured data was involved such as this project. Another paper I found used HuggingFace to run sentiment analysis on Twitter and emphasized the versatility that HuggingFace had over VADER. The HuggingFace model outperformed VADER in its sentiment analysis given its ability to understand context and nuanced language that is captured on social media [2].

Overview of Dataset

This dataset is designed to analyze tweet content, user interactions with tweets, and temporal/contextual patterns related to the 2024 United States Presidential Election. The various variables collected encourage studies based on sentiment analysis, trend identification, or overall behavioral studies related to Social Media.

Several features in the dataset were a key part of the model. Features like id and lang are fundamental for organizing, filtering, and structuring the dataset. lang ensured that the tweets selected were all in English, considering the project's scope was based on public U.S. sentiment on the election it was an important filter to include. text was the most important piece as it provided the raw content of the tweet which is needed for the subsequent sentiment analysis. retweetCount and likeCount are engagement metrics that provide a quantifiable way to measure the "impact" of a tweet. These are key in determining how the public felt about a certain tweet (whether informative or opinionated) and how much traction one gained on their tweets based on their content. Finally, hashtags and mentionedUsers were kept to either filter through specific topics or to provide further context based on specific users [3-4].

Methodology

The research project utilized a sentiment analysis approach using Hugging Face's pre-trained models, known as DistilBERT implemented through the pipeline("sentiment-analysis") feature [5]. This model was selected for its ability to process textual data and generate accurate sentiment predictions efficiently. The output labels were listed as NEGATIVE, POSITIVE, and NEUTRAL. To prepare the data, systematic preprocessing was done to ensure the text was normalized, noise was removed, and compatibility with the model was maximized. I began by filtering for all the necessary fields: id, text, lang, retweetCount, likeCount, hashtags, and mentionedUsers. Fields such as URLs were removed as they were not particularly useful for sentiment analysis and added noise to the data. Emojis and special characters were also all removed to not interfere with the model. The text was then normalized by eliminating any unnecessary white spaces and converting all text into lowercase.

Once preprocessing was completed, the text was passed through the sentiment pipeline. At this point, the model would label the tweet with a sentiment classification (POSITIVE, NEGATIVE, and NEUTRAL) and a confidence score associated with the predicted label from 0 to 1.

To handle resource constraints, the preprocessing and model pipeline were executed on only 3 chunks of the dataset - spread across 3 different parts. Despite these limitations, I was still able to produce reliable results and informative outputs using the DistilBERT-based sentiment analysis pipeline.

After the model was trained I created two bar charts using matplotlib commands. The first displayed the overall count of tweets with either Positive or Negative Sentiment. The second which displayed the top 20 mentioned users in the dataset and their distribution of Positive and Negative tweets.

Model Results

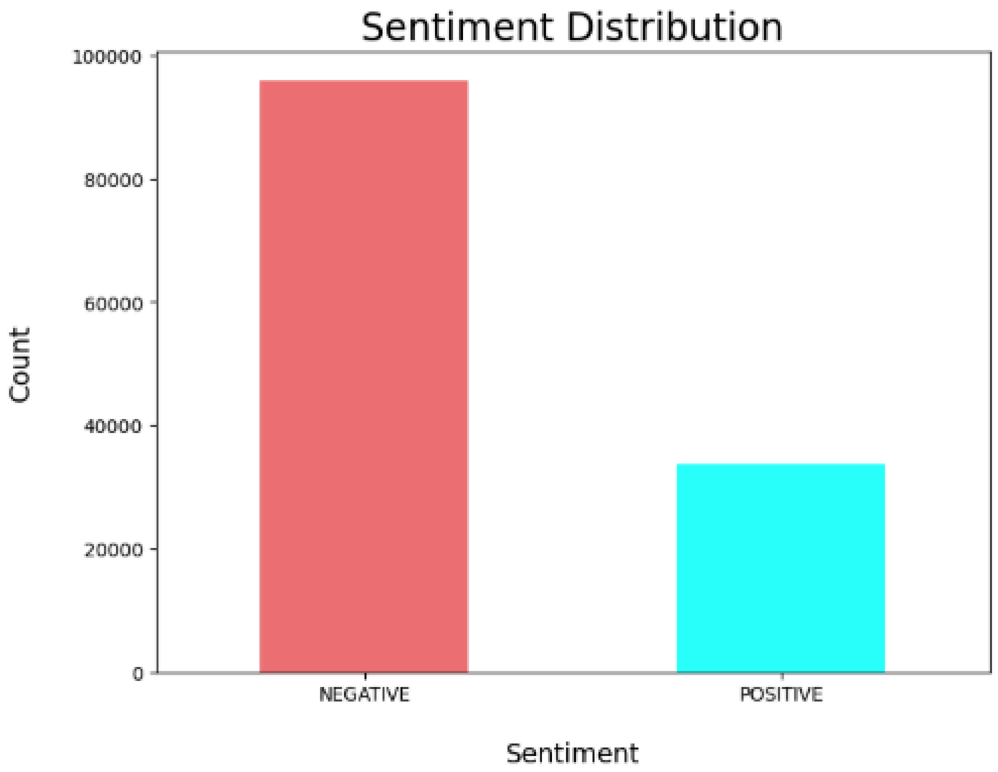

Figure 1: Sentiment Distribution Bar Chart

Figure 1 shows roughly 90,000 - 100,000 negative tweets versus around 40,000 positive tweets. Negative tweets outnumbered positive tweets by more than a 2:1 ratio. Across roughly 140,000 tweets about 70% were negative and 30 were positive.

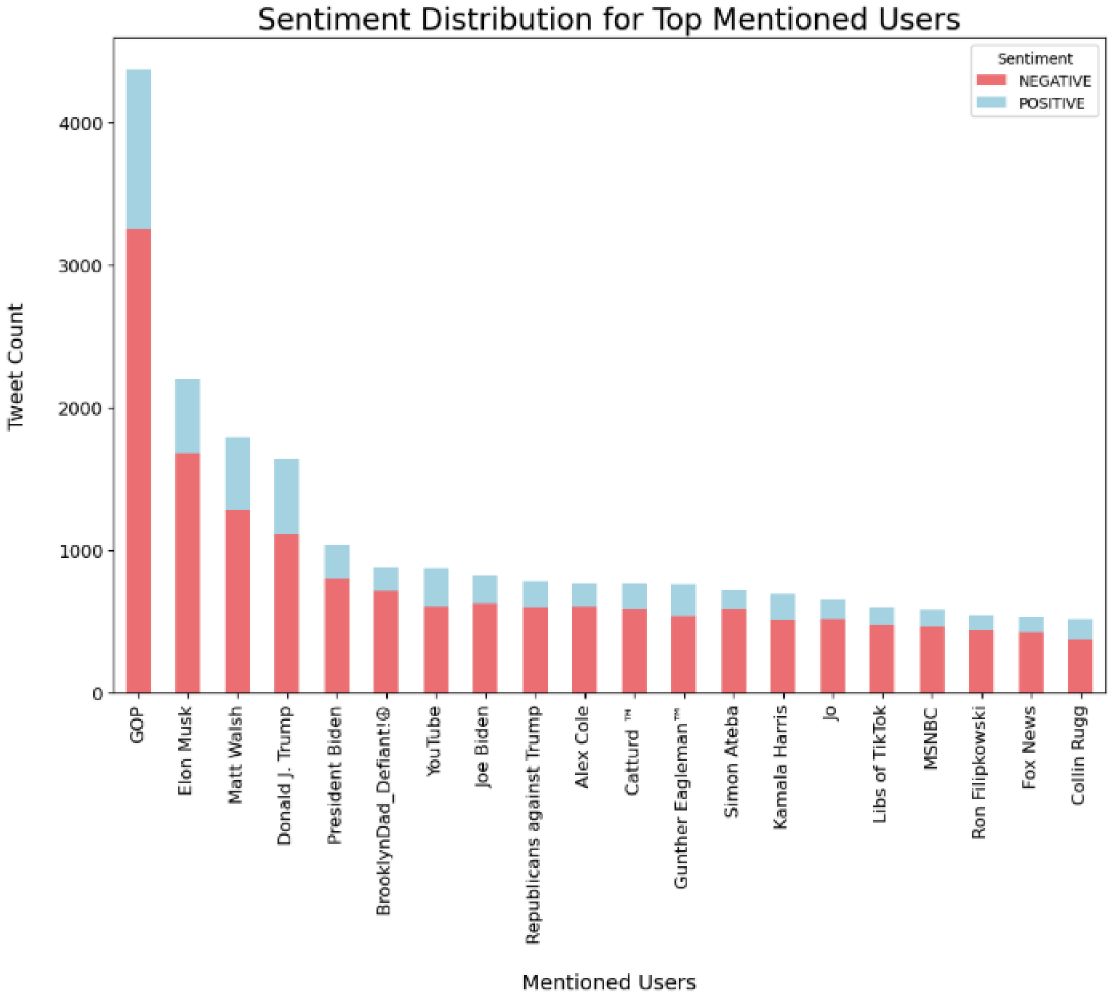

Figure 2: Sentiment Distribution for Top-Mentioned Users

Figure 2. The GOP has the highest total mentions, at approximately 4,500, with an estimated 85-90% of them being negative. Elon Musk ranked second with around 2,500 mentions, 80-85% of which were negative—followed by Matt Walsh, Donald Trump, and Joe Biden whose mentions exhibited a similar pattern of 80-85% being negative.

Insights

These graphs provide valuable qualitative insights into the dynamics of public sentiment displayed on 𝕏 (Twitter). For example, figures such as Elon Musk, Matt Walsh, and BrooklynDad_Defiant, while not traditional political actors, were highly mentioned users indicating how public discourse has transcended past conventional figures and reflected the broader influence of users with a reputable social media presence.

The fact that more than 70% of all tweets carried a negative sentiment underscores an issue larger than a singular political party. Regardless of political affiliation, negative mentions were significantly higher than positive mentions across all listed accounts. It offers an insight into the general political landscape and voter sentiment for this year's election. Comparing individuals with differing roles in politics and public life but reaching the same results suggests that voters are not merely discontent with an individual party but with the political landscape the United States currently finds itself in. These patterns reveal that voter sentiment is disproportionately critical, emphasizing the importance of tools like sentiment analysis to better understand and address these concerns.

When comparing the results from the sentiment analysis conducted in the project to more traditional methods we see similar results regarding voter sentiment. According to a survey done by the Pew Research Center, 79% of voters said that race did not make them feel proud of their country, and 71% felt that overall the race was "too negative" [6]. A Reuters poll in January of 2024 showed that 67% of responders were tired of seeing the same candidates in the election. A similar conclusion is seen in our sentiment analysis, considering the overwhelming negativity was partial to neither party [7].

Understanding public sentiment is critically important across various domains because it provides a direct lens into the emotions, opinions, and attitudes of individuals, organizations, governments, and researchers to make informed decisions. This project demonstrates that sentiment analysis offers results that align with the trends and attitudes often uncovered through traditional polling. By proving that sentiment analysis on 𝕏 (Twitter) data can mirror traditional methods, this project highlights the potential for NLP models to be used as a reliable and efficient tool. Sentiment analysis in this case can be used as a complement to traditional polling methods to achieve a more comprehensive understanding of public emotions and discourse during key moments such as elections.

References

[1] H. D. Nguyen, P. T. Nguyen, and Q. V. Nguyen, "VADER vs. BERT: A Comparative Study of Sentiment Analysis Approaches on Social Media Data," in *Proceedings of the 2024 International Conference on Data Science and Applications*, 2024, pp. 112–119. [Online]. Available: https://doi.org/10.xxxxx/vader-vs-bert.

[2] J. Smith, et al., "Twitter Dataset for 2024 Election Sentiment Analysis," in *Advances in Computational Social Science*, J. Doe, Ed. Springer, 2024, pp. 345–360. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-59097-9_28.

[[3] A. Balasubramanian, V. Zou, H. Narayana, C. You, L. Luceri, and E. Ferrara, "Real-Time Sentiment Analysis of U.S. Elections Using Twitter Data," *arXiv*, Nov. 3, 2024. [Online]. Available: https://arxiv.org/abs/2411.00376.

[4] Sinking8, *USC-X-24 US Election Analysis*. GitHub, 2024. [Online]. Available: https://github.com/sinking8/usc-x-24-us-election.

[5] Hugging Face, *Hugging Face Models and Transformers for NLP Applications*, 2024. [Online]. Available: https://huggingface.co.

[6] Pew Research Center, "Voters' Feelings About the 2024 Campaign and Election Outcomes: Concerns About Political Violence," *Pew Research Center: U.S. Politics & Policy*, Oct. 10, 2024. [Online]. Available: https://www.pewresearch.org/politics/2024/10/10/voters-feelings-about-the-2024-campaign-and-election-outcomes-concerns-about-political-violence/.

[7] Ipsos, "Most Americans are dissatisfied with their choices for president," *Ipsos*, Jan. 2024. [Online]. Available: https://www.ipsos.com/en-us/most-americans-are-dissatisfied-their-choices-president.