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

Giovanni Vurro

Introduction

The problem the project is trying to solve is a lack of timely, scalable, nuanced insight into public sentiment at times of elections that are needed to make sense of the behavior of voters and the shaping of political strategy. Although valued, traditional polling methods are usually time-consuming, expensive, and limited in scope, capturing just a snapshot of opinions sans real-time shifts or the dynamic nature of public discourse. With the advent of the digital era, Twitter has become one of the major platforms where political conversation occurs; this also reflects a huge amount of publicly available data that is diverse and spontaneous in nature. However, the volume and unstructured nature of this data make it difficult to analyze either manually or conventionally. This project addresses these challenges by leveraging advanced NLP techniques, specifically sentiment analysis, for efficient processing and analysis of large-scale Twitter data, thus providing timely, detailed insights into voter sentiment and trends, shifts, and emotional responses in real time. This will, in turn, enable a range of stakeholders, political campaigns, policymakers, and researchers to make better, evidence-based decisions that would enhance their capacity for effective engagement with the electorate and respond to public concerns.

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 that described sentiment analysis that was conducted for the Coronavirus Outbreak and did a comparative experiment of VADER vs. BERT. 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.

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, date, 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.

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. 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 - having the neutral label was important as it could capture tweets that had a more ambiguous sentiment. 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, date, 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 removing 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) as well as a confidence score associated with the predicted label from 0 to 1. Although at this point the model has only been evaluated using one "chunk" of data (due to time) the confidence scores were then examined to assess the distribution of sentiments across the selected dataset. Based on the results

I plan on creating visualizations that better illustrate what the overall sentiment was regarding the election. 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. In today's social media anyone's voice can be amplified and it offers us an even more direct insight than a simple survey or basic election polling. I am positive that the results can be used to identify patterns, trends, and insights regarding public opinion.