2024 US Presidential Election Sentiment Analysis

Team Avengers – Prachet Balaji, Dominique Eberhard, Sai Nishanth Mettu, Neha Saji Joseph, David Liu, Qile Asher Zhou

Big Data CS-GY 6513
12/10/2024

NYU

Contents

1. **Introduction**
2. **Problem Statement**
3. **Data Collection Using Big Data**
   - 3.1. Twitter Data Collection
   - 3.2. Reddit Data Collection
   - 3.3. Big Data Technologies for Efficient Data Collection
4. **Data Pre-Processing**
   - 4.1. Text Normalization
   - 4.2. Stopword Removal
   - 4.3. Lemmatization
   - 4.4. N-gram Generation
5. **Natural Language Processing**
   - 5.1. Sentiment Analysis
   - 5.2. N-gram Modeling
   - 5.3. Topic Modeling
6. **Results**
   - 6.1. Kamala Harris Sentiment
   - 6.2. Donald Trump Sentiment
7. **Big Data Patterns Used**
8. **Future Work**
9. **Conclusions**

1. Introduction

Although each US Presidential Election is touted as the “most important election” ever, the 2024 US Presidential Election certainly seemed to live up to that title. At least, that is what an outside observer consuming social media posts over the past few months would likely have begun to believe. With public debate reaching unprecedented levels, social media seemed like one of the more important aspects of the most recent election cycle. At the forefront of all of these social media posts were President Joe Biden, incumbent Vice President Kamala Harris, and former President Donald Trump. From the outside, it looked like the policy positions and the vision of the future of the United States touted by these two sides were strikingly different.

As online communities continue to gain power, Reddit and Twitter have become a crucial place for civic engagement, opinion exchange, and policy debate. These new “town squares” provide a large amount of publicly available statistics which can be analyzed for the purpose of estimating public sentiment and tracking political bias. Understanding the above sentiments and the topics emerging from the discussions in the online communities is an interesting way to take a litmus test of the overall political mood in the country.

Thus, this project is intended to examine the feelings expressed by digital media users towards Donald Trump and Kamala Harris during the 2024 election campaign. Since Joe Biden dropped out of the race in July, 2024, we will consider Harris more closely. We aim to evaluate a large number of posts using natural language processing (NLP) strategies and massive statistical systems such as Apache Spark to gain insights into how social media users (as a proxy for voters and citizens) perceived the campaigns and the major issues surrounding the election. The present analysis provides a comprehensive evaluation of the virtual civic discourse via sentiment analysis, n-gram shape, and factual image, which may perhaps shed more light on how people feel about politics than a traditional poll.

2. Problem Statement

Due to its enormous volume and the intricacy of human language, social media data analysis is difficult. The billions of posts that social media networks produce every day make it challenging to handle all of this data manually and use traditional data analysis technologies like Pandas. Given the informal language, slang, abbreviations, and distinctive communication style employed on social media platforms, it is difficult to identify significant trends in data, even when automated. Additionally, subtleties in people's opinion expression, such as irony, sarcasm, and contextual fluctuation, might make sentiment analysis difficult.

These difficulties are exacerbated in the context of political speech by the extreme polarization of modern American politics. Strong feelings, partisan rhetoric, and ideological differences are common in posts about Donald Trump and Kamala Harris. Therefore, it takes both sophisticated computational techniques and a thorough understanding of the political context to analyze sentiment in political speech.

This project tackles a number of important research issues:

How can big data technology be used to effectively gather and process massive amounts of social media data?

How may sentiment analysis be used to pinpoint the general public opinions in the discussions around Donald Trump and Kamala Harris on social media?

Which political concerns and recurring themes show up in posts about these candidates on social media?

How might recurring themes and expressions that characterize the public conversation surrounding the 2024 election be found using n-gram analysis?

By answering these questions, we hope to establish a methodology for carrying out extensive sentiment analysis research on social media platforms and offer insightful information about the political environment around the 2024 US election and US politics.

3. Data Collection using Big Data

Prior to performing sentiment analysis on social media data, it was essential to gather a sizable and representative dataset. Traditional data collection techniques are inadequate due to the volume and variety of social media posts. To manage the size and complexity of the data, big data technologies—in particular, Apache Spark—were employed.

3.1 Twitter Data Collection

Millions of people have been using X, formerly Twitter, every day to express various opinions and beliefs, making it one of the most important and influential social networking platforms in the world. Because of the magnitude of the office they were running for, Donald Trump and Kamala Harris created an enormous amount of conversation on Twitter, which makes it a perfect place to do sentiment analysis. The Tweepy library was used to gather data for this project. It connects to the Twitter API and extracts tweets in real time.

Only tweets from people based in the United States were included in the sample. This was accomplished by filtering for tweets coming from within the United States using geolocation markers. To find pertinent tweets about the candidates and election-related topics, a list of search terms was also employed, including "Kamala Harris," "Donald Trump," "2024 election," "Kamala2024," "Trump2024," and similar terms. Random sampling was used to guarantee that tweets from both the candidates' supporters and detractors were included in the dataset, making it representative and diverse.

The data collecting procedure was automated to extract tweets at regular intervals due to the high amount of tweets, which allowed for the acquisition of a sizable corpus of data over time. In our submission, we have included the scripts that served as the foundation for collecting tweets and Reddit posts, respectively. To store the data, we initially considered using Hadoop Distributed File System (HDFS), specifically with DataProc, but we found that that was a bit cumbersome. However, we ended up using Google Cloud Storage (for free for our purposes). This idea here was to make sure it was conveniently accessible for additional analysis, scalable, and fault-tolerant.

3.2 Reddit Data Collection

Compared to Twitter, Reddit offers a different kind of debate setting because it is a discussion-based site. Long, in-depth posts and follow-up comments are common in Reddit threads, which offer a forum for more nuanced and complex political debates. Since these communities are heavily interested in political disputes, subreddits like r/PoliticalDiscussion, r/KamalaHarris, r/Trump, and r/2024Election were some of the main subreddits of relevance for this investigation.

Posts and comments from these subreddits may be extracted thanks to the Reddit API and praw, the Python Reddit API Wrapper, which was used to gather data. Similar to Twitter, we tried to collect only English-language posts from users in the United States. The entire range of political discourse was captured by the data collection, which comprised both the original posts and the comments.

Conversations from various phases of the election cycle were captured by randomly sampling the posts and comments across time to guarantee that the dataset reflected a wide range of viewpoints. These posts were easily accessible for further examination because they were saved in a dispersed fashion.

3.3 Big Data Technologies for Efficient Data Collection

Big Data technologies were essential for handling and processing the gathered social media data because of the sheer amount of data involved. The data was processed in parallel across several workstations using Apache Spark. Spark was

the perfect choice for this project because of its capacity to manage complex transformations and large-scale data processing workloads.

Spark DataFrames, which offered an effective means of storing, modifying, and querying the data, was used to load the gathered data. We processed millions of social media postings in a fraction of the time it would have taken using conventional techniques thanks to Apache Spark's distributed architecture. Pre-processing and data analysis were made simple by Spark's interaction with Python-based libraries like Pandas and NLTK, as well as its compatibility with other Big Data tools like Hadoop.

4. Data Pre-Processing

By its very nature, social media data is unstructured and “dirty”, necessitating extensive pre-processing before analysis can begin. Pre-processing aims to standardize the language and eliminate unnecessary information from the data in order to tidy it up and get it ready for analysis.

4.1 Text Normalization

Normalizing the text was the initial stage of the data pre-processing pipeline. Normalization makes it possible to guarantee that the text follows a standard structure and is devoid of any letters or symbols that don't provide significant or meaningful information. All of the content was changed to lowercase in order to prevent text matching errors (for example, "Kamala" and "kamala" would be considered different words otherwise).

Following that, regular expressions were used to eliminate special symbols and non-ASCII characters. For sentiment analysis, URLs, hashtags, and mentions (like @KamalaHarris) were eliminated because they frequently offer minimal value. As a result, the model was able to concentrate on the main ideas of each post without getting sidetracked by formatting or other references.

4.2 Stopword Removal

Common words like "the," "and," "is," and "in" that are used a lot in a text but have little semantic meaning are called stopwords. The NLTK library's predefined list of English stopwords was used to eliminate stopwords from the dataset because they don't add much to sentiment analysis. By removing extraneous noise and shrinking the dataset, this procedure improved the efficiency of the analysis that followed.

4.3 Lemmatization

Cutting words down to their most basic or root form is known as lemmatization. For example, the phrases "running," "ran," and "runs" are all variants of the word "run." The consistency of the analysis is increased by lemmatizing the text to guarantee that various word forms are considered as the same word.

We performed lemmatization in this project using NLTK's WordNetLemmatizer. By reducing the complexity of the data and unifying many word grammatical forms, this technique facilitates the identification of sentiment patterns and important topics.

4.4 N-gram Generation

Word sequences known as "n-grams" can highlight significant patterns in written language. In order to identify recurring themes and subjects in social media conversations, we created bigrams (sequences with two words) and trigrams (three words) for this project. N-grams are especially helpful for recognizing significant word combinations or phrases, like "Kamala Harris" or "Donald Trump," which may not be used often as single words but which carry great weight in political contexts, at least in relation to the most recent election.

We were able to discover important concerns and subjects related to each candidate with the aid of these n-grams. Bigrams for Kamala Harris, for instance, featured terms like "Kamala Harris" and "healthcare reform," and trigrams for Donald Trump featured phrases like "Make America" and "economic growth."

5. Natural Language Processing (NLP)

Following data pre-processing, NLP methods were used to glean valuable information from the cleaned text. Topic extraction, n-gram modeling, and sentiment analysis were the main NLP tasks used in this project. These techniques were crucial for determining the main topics of discussion in regard to Donald Trump and Kamala Harris as well as for better understanding and identifying public sentiment.

5.1 Sentiment Analysis

Finding out if a text expresses a favorable, negative, or neutral sentiment is known as sentiment analysis. Sentiment analysis models based on machine learning and rules were both used for this analysis.

Initial sentiment scoring was done using a rule-based sentiment lexicon, like the VADER (Valence Aware Dictionary and sEntiment Reasoner) model. This lexicon assigns sentiment scores to words according to their polarity (positive, negative, or neutral) using predetermined rules. VADER work s well with social media text because it takes into consideration the slang, expressive language, and acronyms that are frequently used on sites like Reddit and Twitter.

We used labeled social media postings to train a machine learning model for sentiment prediction in order to achieve more complex sentiment classification. We created a model that can predict sentiment based on the presence of specific keywords, phrases, and contextual information using methods like Support Vector Machines (SVM) or Logistic Regression.

5.2 N-gram Modeling

The most common word sequences that emerged in the social media posts were identified using N-gram modeling. We were able to identify bi-grams and tri-grams—word sequences that were used to characterize Donald Trump and Kamala Harris—by using n-gram models on the cleaned text. These segments frequently relate to certain subjects or problems that were at the heart of the discussion.

Terms like "Kamala Harris" and "healthcare reform" were among the top bigrams for Kamala Harris, indicating that a large portion of the discourse surrounding her was centered on policy matters. "Make America" and "economic growth" were among Donald Trump's top slogans, suggesting that his followers were largely preoccupied with economic concerns and his pledges to revive the American economy.

5.3 Topic Modeling

To find underlying themes in the data, topic modeling was used in addition to n-grams. Based on their co-occurrence patterns, words were grouped into different categories using the well-known topic modeling technique Latent Dirichlet Allocation (LDA). This made it easier to pinpoint the main political topics and themes being brought up in respect to Donald Trump and Kamala Harris. By examining these subjects, we were able to identify the main concerns of voters and monitor their development over time.

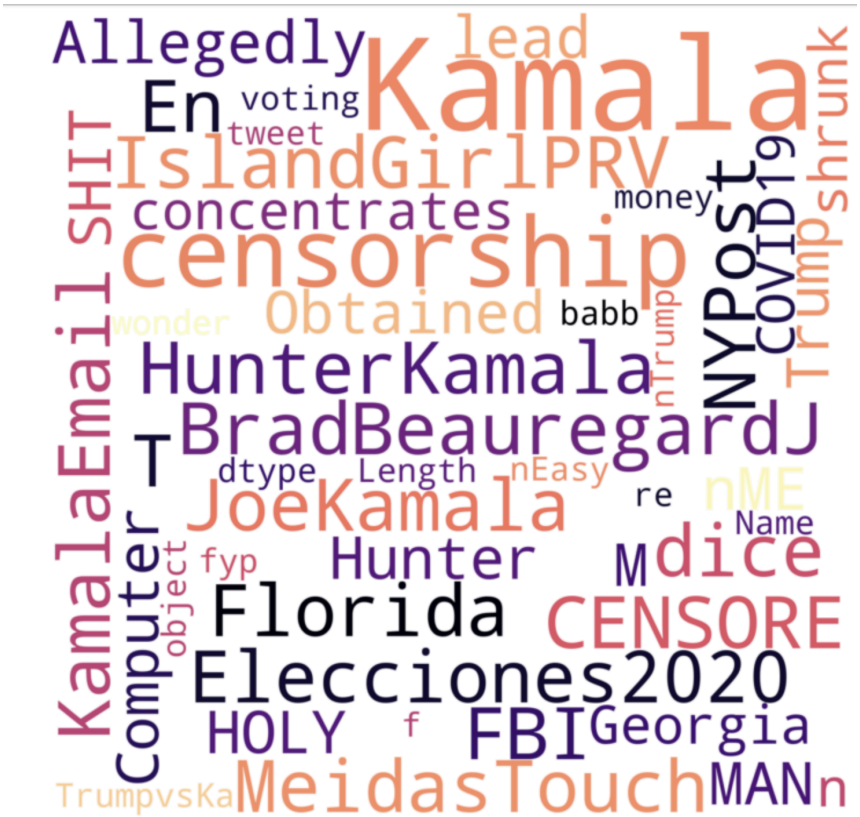
6. Results

Regarding Kamala Harris and Donald Trump in the context of the 2024 election, the sentiment analysis results provide interesting insights about popular

opinion. The polarized nature of American politics was reflected in the stark difference in opinion between the two contenders.

6.1 Kamala Harris Sentiment

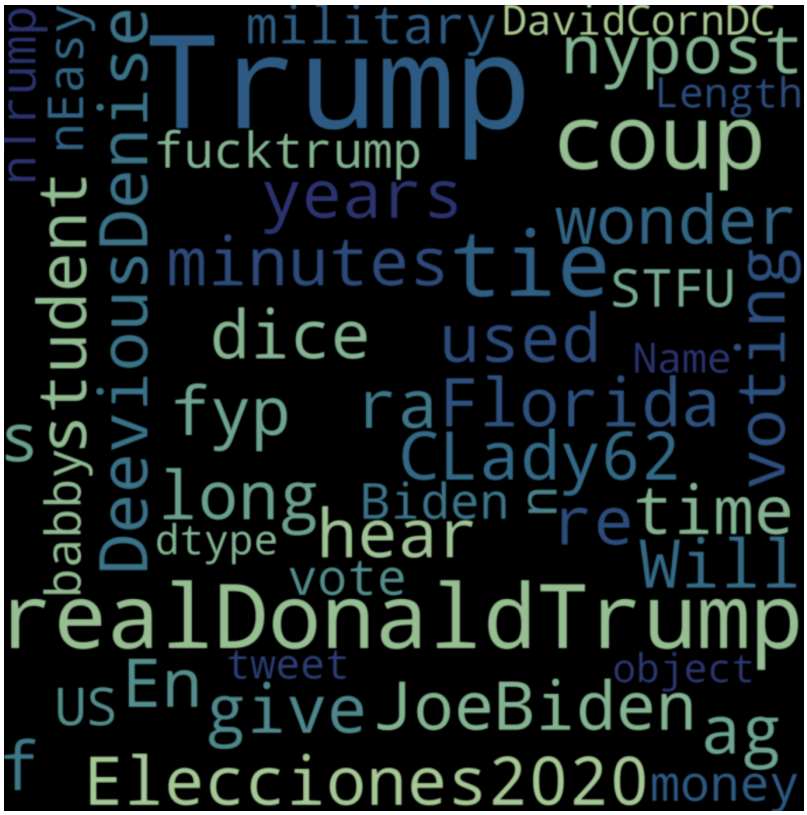
Although sentiment toward Kamala Harris was more ambiguous, it tended to be favorable. Many people expressed support for her position on progressive policy proposals, and discussions surrounding her centered on topics like healthcare reform, racial justice, and climate change. Her relationship with President Joe Biden was the subject of divisive posts; while pro-Biden fans applauded her work as Vice President, detractors voiced their displeasure with her affiliation with the Biden administration. Positive sentiment surrounding her support of social justice and climate change action, however, counterbalanced these divisive conversations.



6.2 Donald Trump Sentiment

According to Donald Trump's sentiment analysis, the political climate is extremely divisive. Trump-supporting posts were very, very positive, especially when it came to his economic ideas, his “Make America Great Again” philosophy, and his pledges to improve the economy. However, the majority of the posts that criticized Trump concentrated on how he handled the COVID Pandemic, climate change, and

racial injustice, which created a lot of negativity. As evidenced by his divisive nature as a politician, Trump's political rhetoric mirrored the ideological split in the United States.



7. Big Data Patterns Used

In order to understand the dynamics of public opinion in the 2024 election, it was essential to employ Big Data technologies such as Apache Spark and Natural Language Processing to uncover important patterns. Large quantities of data were handled effectively thanks to Apache Spark's distributed processing infrastructure, and its capacity to carry out intricate data transformations allowed for the discovery of important insights into public opinion. Sentiment analysis algorithms and n-grams were used to highlight important political concerns and show how the people viewed Donald Trump and Kamala Harris.

We were able to discover recurring themes in political discourse, follow sentiment trends over time, and extract the most popular topics addressed by social media users thanks to these Big Data patterns. Understanding how public opinion is

formed and how political campaigns might use these insights to effectively target voters requires a large-scale, detailed approach to sentiment research.

8. Future Work

Subsequent investigations may examine the introduction of machine learning algorithms to enhance sentiment prediction accuracy and the addition of other social media sites. The accuracy of sentiment categorization could be further improved by training deep learning models, such as Transformers or Recurrent Neural Networks (RNNs), particularly when we want to detect irony or sarcasm.

Future research can also look at how political speech is changing as we will enter another election cycle in 2 years and then in 4 years. This can entail monitoring changes in public opinion in reaction to significant political debates, events, or policy declarations. The applications for political campaigns of such a project are evident as they can modify their tactics to meet new challenges and strengthen their relationship with voters by keeping an eye on these trends.

9. Conclusion

The examination of social media posts about Donald Trump and Kamala Harris offers a thorough picture of popular opinion regarding the 2024 U.S. election. Millions of messages were analyzed thanks to the combination of sophisticated natural language processing (NLP) techniques and Big Data technologies, especially Apache Spark, which provided interesting insights into the topics that dominated political discourse for the past year or so. According to the findings, opinions on Donald Trump and Kamala Harris are extremely divided, which reflects the larger ideological divide in American politics.

This project emphasizes how crucial it is to use sentiment analysis and Big Data in political campaigns and voter outreach initiatives. Campaigns may more effectively adjust their messaging and tactics to appeal to voters and navigate the increasingly complicated political terrain of the political elections by understanding and analyzing the millions, maybe even billions, of opinions freely available on social media.