Analyzing Public Sentiment on Ongoing Political Campaigns in the US using Reddit Data

Raga Naga Sowmya Eemani, Vineeth Reddy Beeram, Vinithra Sadras, Dileep Kumar Perala

Syracuse University, reemani, vbeeram, vsadras, dperala, as@syr.edu

*Abstract* –This research project focuses on analyzing public sentiment regarding ongoing political campaigns conducted by the Republican and Democratic parties in the United States using Reddit data. With the increasing significance of social media platforms like Reddit in shaping political discourse, understanding user sentiments becomes imperative. Leveraging advanced data mining algorithms and sentiment analysis techniques, this study aims to provide nuanced insights into public perceptions and reactions. The project employs the Reddit API to collect data, emphasizing the extraction of information from hashtags and discussions directly associated with political debates. Essential preprocessing steps, including tokenization and lemmatization, are implemented to enhance sentiment analysis accuracy. Various data mining algorithms, such as Decision Trees and Natural Language Processing, are explored to predict sentiment and classify posts into positive, negative, or neutral categories. Domain-specific hashtags and embeddings are leveraged to capture nuanced sentiments unique to political discussions. The project evaluates chosen algorithms using key metrics like accuracy and recall unveiling trends and patterns in public sentiment, ultimately providing actionable insights for political strategists, policymakers, and researchers. This research contributes to a deeper understanding of contemporary political landscapes and offers valuable guidance for stakeholders in navigating the complexities of public opinion.

*Index Terms* – Data Collection, Data Mining, Political Sentiment Analysis, Public Opinion, Reddit API, Reddit Data Analysis, Sentiment Analysis, Social Media Data Mining.

introduction

In this digital era, the platforms of social media are becoming key battlegrounds for political discussions. In this regard, people use these platforms to voice their opinions, debate ideas, and share information about political events and candidates. If politicians, analysts, and policymakers wish to understand the mood of the public, they must first understand the mood that people express on these social media platforms.

Our research project aims to dive deep into these sentiments, specifically focusing on the ongoing political campaigns of both the Republican and Democratic parties in the United States. We are honing in on Reddit, a massive online community containing many varied users and interests, to see how people perceive and react to these political campaigns.

We have planned to apply data mining algorithms and sentiment analysis techniques to extract detailed insights from the vast array of discussions and hashtags related to political factions on the platform. To begin, we have already gathered and pre-processed Reddit data that is relevant to political discussions and campaign-related hashtags to ensure that the data we are working with is authentic and directly applicable to our analysis. We then go on to apply advanced data mining algorithms to sift through the data we have collected. We will analyze the data to identify any patterns, trends, or correlations that might shed light on public opinion regarding political campaigns. In this way, we will use sentiment analysis techniques to put these feelings into positive, negative, or neutral, giving us a well-rounded understanding of public opinion.

The importance of our research lies in the potential these findings have to offer timely insights into how public opinion changes with time in the digital age. By offering a nuanced understanding of how Reddit users perceive and interact with political campaigns, we hope to provide practical guidance to those involved in the political sphere. These findings are especially significant because of the ever-changing digital landscape, where social media holds an increasingly powerful sway over political discourse and decision-making.

Literature Survey

Using the data generated by social media to track the political sentiment analysis has gained significant attention in the recent times. The research study conducted by [1] Elvyna Tunggawan and Yustinus et al. titled , "Bayesian Twitter-based prediction on 2016 U.S. presidential election", the authors have simplified the preprocessing methodology while maintaining predictive accuracy . The researchers have managed to simplify the preprocessing by focusing on a Naive Bayesian approach and utilizing a more streamlined preprocessing pipeline. They believe that this methodology will reduce the complexities behind the preprocessing that is required before implementing any model to predict the sentiment analysis based on twitter data. Since we are not using twitter data, our data is not as complicated as of twitter data, but we have taken reference from their method to preprocess our data. A numerous number of studies have been conducted on the same topic, implementing different kinds of methodologies, and further comparing which methodology gives the best results. These studies have delved into various aspects of Twitter-based analysis to gauge public opinion and potentially predict electoral outcomes. In the paper presented by [2] Ibrahim et al. (2015) the researchers have used statistical approaches, leveraging sentiment analysis techniques to extract insights from Twitter data and forecast election results with varying degrees of accuracy.

Opposing to them, [3] Oyebode & Orji (2019) have proposed a lexicon-based approach. This method utilizes sentiment lexicons to classify tweets and predict electoral outcomes in different contexts. Additionally, supervised learning approaches, demonstrated by [4] Mahmood et al. (2013) and Razzaq, Qamar & Bilal (2014), employ machine learning algorithms such as Naive Bayes and Support Vector Machines to train predictive models based on labeled tweet data. Addressing all these diffferent types of methodologies used by different researchers and organization to predict the sentiment analysis, [5] Bansal & Srivastava have proposed a research paper studying more about these individual methods and identifying the underlying challenges to all of these methods. These issues include but not limited to data sparsity, noise, and the need for robust evaluation metrics. The researchers have further proposed methods to overcome these challenges, such as the development of hybrid topic-based sentiment analysis approaches which highlight the ongoing evolution of methodologies in this field.

reddit connection

The data collection process involved utilizing Reddit's Application Programming Interface (API) in conjunction with the Python Reddit API Wrapper (PRAW) library. Through this interface, we systematically gathered relevant information pertaining to ongoing political campaigns within the United States. Specifically, we employed a custom function named 'headlines' integrated into our Python script, which facilitated the extraction of submission titles containing specific hashtags corresponding to individual political candidates.

The Reddit connection process involves setting up a Reddit application and using the Python Reddit API Wrapper (PRAW) to interact with Reddit's API. Initially, a Reddit application is created through the user's account preferences, where a unique `client\_id` and `client\_secret` are obtained.

A screenshot of a computer

Description automatically generated

Figure-1 Reddit App Connection

These credentials are essential for authenticating the PRAW connection. In the Python environment, PRAW is installed using the package manager `pip`, and the library is imported into the script. Additional libraries such as `pandas`, `numpy`, and `re` for regular expressions may also be imported as required for data handling and pattern matching.  
  
A screenshot of a computer code

Description automatically generated

Figure-2 Importing required modules

To initiate the connection, a `user\_agent` string is defined, which describes the purpose of the script to Reddit, in this case, for fetching and analysis. The PRAW instance is then created by passing the `client\_id`, `client\_secret`, and the `user\_agent` as parameters. This instance acts as a gateway to Reddit, enabling the script to access, fetch, and perform analysis on the data from specified subreddits or posts. This process sets the stage for any subsequent data mining and sentiment analysis tasks aimed at understanding public sentiment on various topics discussed on Reddit.

data collection

In this project, the data collection process is a critical initial step that lays the groundwork for the entire analysis. Utilizing the Python Reddit API Wrapper, a specialized script is developed to interact with the Reddit API, targeting the 'politics' subreddit—a hub for political discourse. The script functions by dynamically inserting specific hashtags related to the names of the election contestants into the search query to filter the data precisely.

For the first experiment, it fetches the titles of the top 200 trending posts regardless of their submission date, aiming to capture the most immediate and influential discussions. This approach allows the team to gauge which topics are gaining traction among Reddit users in real-time and could influence public opinion about the contestants.

A computer screen shot of a black screen

Description automatically generated

Figure-3 Fetching top 200 posts

In contrast, the second experiment extends the data collection to encompass all posts from the last month, providing a broader and more historical view of public sentiment. By setting the time filter to 'month', the script selectively gathers post titles that reflect the ongoing narrative and shifts in perspective regarding the election contestants over a substantial period. The dataset is meticulously compiled to include only unique entries, ensuring the quality and integrity of the data.

A computer screen with text on it

Description automatically generated

Figure-4 Fetching posts in last 30 days

The resulting collection of Reddit post titles forms a comprehensive dataset that serves as the foundation for subsequent analytical steps. This dataset is then prepped for a thorough sentiment analysis, which will seek to unearth the underlying tones and attitudes present in public discussions surrounding the 2024 US election contestants. To ensure a comprehensive and up-to-date dataset, we focused on collecting data spanning a period of one month. During this timeframe, submissions related to five prominent political contestants were targeted for analysis. These contestants included Joe Biden, Donald Trump, Robert F. Kennedy Jr., Kanye West, and Jill Stein.

The 'headlines' function was instrumental in retrieving submission titles from Reddit, leveraging various parameters and filters to tailor the data collection process to our specific requirements. By specifying the subreddit to search within—whether it be the entirety of Reddit ('all') for a broad perspective or specific subreddits for more targeted discussions—we were able to hone in on relevant conversations surrounding the aforementioned political figures.

Furthermore, sorting parameters such as 'new' and time filters such as 'month' were applied to ensure that the collected data remained current and aligned with our designated timeframe. Each retrieved submission title was meticulously stored in a set data structure to eliminate duplicates and streamline the dataset for subsequent analysis.

Overall, this approach allowed us to compile a rich and diverse dataset encompassing a wide array of perspectives and opinions expressed across various Reddit communities. By leveraging the power of Reddit's API and PRAW library, we were able to gain valuable insights into public sentiment surrounding the ongoing political campaigns, laying the foundation for our subsequent sentiment analysis and interpretation.

Data Pre-Processing

In the "Analysing Public Sentiment on Ongoing Political Campaigns in the US using Reddit Data" project, data pre-processing is a critical step that significantly impacts the accuracy and quality of the sentiment analysis. The raw data collected from Reddit encompasses a wide variety of content, including links, emoticons, and special characters that are not pertinent to understanding user sentiment and can potentially mislead the analytical process.

To address this, a specialized text pre-processing function has been implemented. This function systematically cleans the data by applying a series of regular expressions (regex) designed to filter out irrelevant or disruptive elements. The first step in this cleaning process is the removal of URLs, HTML tags, emojis, and emoticons, which, while common in online communication, do not contribute to sentiment and can introduce noise into the analysis.

A screenshot of a computer code

Description automatically generatedFigure-5 Data Preprocessing

Following this, the function removes punctuation and special characters. Such characters can lead to incorrect interpretations of words and phrases, as they may cause similar strings to be mistaken for unique tokens during analysis. The function also targets words that are unlikely to carry sentiment value—specifically, very short words (less than three characters) and unusually long strings (greater than fifteen characters), which often represent noise rather than meaningful content.

Lastly, the preprocessing includes the elimination of excessive whitespace. This not only contributes to a more uniform dataset but also ensures that the sentiment analysis algorithm operates on neatly formatted and consistent input, thereby improving the reliability of the sentiment measures extracted. Through this rigorous preprocessing routine, the dataset is refined to a form that is optimally structured for the subsequent sentiment analysis, ensuring that the focus is directed solely on the meaningful textual content that reflects user sentiment regarding the election contestants.

Data Frequency visualization

Following the data pre-processing phase in our project, the next step involves analyzing the frequency of words within the processed text. To accomplish this, we have developed a user-defined function that utilizes the WordCloud library in Python. The purpose of this function is to visually represent the most frequent terms that appear in our corpus of Reddit post. titles.

A screen shot of a computer code

Description automatically generatedFigure-6 Code for plotting a word cloud

The function, `generate\_word\_cloud`, first concatenates all preprocessed text from a specified column of our dataset into a single string. It then generates a word cloud from this text, which is configured to display the most prominent words in a specified number of maximum words; in our case, the limit is set to 100. The WordCloud library is instructed to ignore common stopwords, ensuring that the visual representation focuses on the most meaningful words. The word cloud is generated with specified dimensions and a black background for optimal contrast and visual appeal.

A close up of words

Description automatically generated

Figure-7 Sample Word cloud output

Once the word cloud is created, it is displayed using the matplotlib library. The size of the figure is adjusted to ensure clarity and detail, and axes are turned off for a cleaner image presentation. This graphical output helps us to quickly identify which words are most dominant in discussions related to U.S. election contestants, thereby providing a snapshot of the key themes and topics that are resonating with the public. Through this analysis, we aim to gain deeper insights into the collective mindset of the electorate, as reflected in online conversations on Reddit.

Sentiment Scores

As part of the sentiment analysis in our project, we have integrated functions that utilize the TextBlob library to determine the subjectivity and polarity of the text. The `getSubjectivity` function calculates the amount of personal opinion versus factual information within the text. Higher subjectivity indicates a more opinionated text, while lower subjectivity suggests the text contains more factual information.

Conversely, the `getPolarity` function assesses the overall sentiment of the text. Polarity scores range from -1 to 1, where a higher polarity value indicates a more positive sentiment and a lower polarity value denotes a more negative sentiment. This method allows us to quantitatively gauge the sentiments expressed in Reddit posts, providing a nuanced understanding of the public's opinion toward the election contestants.

A screen shot of a computer code

Description automatically generated  
Figure-8 Sentiment score calclulation

These functions are pivotal in analyzing and interpreting the tone and subjectivity of social media discourse. By systematically applying these measures to the collected data, we can ascertain the general sentiment trends, whether they lean towards positive, negative, or neutral, and how much of the conversation is driven by personal beliefs as opposed to objective statements. This level of analysis is instrumental in decoding the complex dynamics of public sentiment in online political discussions.

Results

Here are the results of Subjectivity for 5 politicians in both the experiments for top 200 and last month experiments:

|  |  |  |
| --- | --- | --- |
| **TOP 200  POSTS** | **NEUTRAL (GENERAL STATEMENTS)** | **POSITIVE (OPINION)** |
| **BIDEN** | 56.77% | 43.22% |
| **TRUMP** | 56.37% | 43.62% |
| **KENNEDY** | 50.21% | 49.78% |
| **WEST** | 58.33% | 41.66% |
| **STEIN** | 43.96% | 56.03% |

Table-1 Subjectivity for top 200 posts

|  |  |  |
| --- | --- | --- |
| **LAST MONTH** | **NEUTRAL (GENERAL STATEMENTS)** | **POSITIVE (OPINION)** |
| **BIDEN** | 51% | 49% |
| **TRUMP** | 50% | 50% |
| **KENNEDY** | 65% | 35% |
| **WEST** | 60% | 40% |
| **STEIN** | 60% | 40% |

Table-2 Subjectivity for last 30 days

This subjectivity data reflects the sentiment distribution within Reddit posts discussing five politicians. The percentages are divided into 'Neutral' and 'Positive' categories for two distinct datasets: the top 200 trending posts and posts from the last 30 days.

For the top 200 trending posts, the subjectivity data shows a relatively balanced distribution between neutral and positive sentiments for most politicians, with the neutral sentiment being slightly more dominant. Biden, Trump, and West have neutral sentiment percentages in the mid to high fifties, indicating that more than half of the top posts about them are general statements without a strong opinion. Kennedy's data is nearly evenly split, while Stein stands out with a higher positivity rate at 56.03%, suggesting that posts about Stein tend to contain more positive opinions than neutral statements.

A graph of different colored rectangles

Description automatically generated

Figure-9 Subjectivity plot for top 200 posts

When we look at the subjectivity data from the last 30 days, the distribution shifts notably for Kennedy, who now has a significant majority of neutral sentiment at 65%, a change that suggests a more objective or information-focused discussion. West and Stein also see their neutral sentiment percentages increase to 60%, which could imply a shift towards more fact-based dialogue or a decrease in emotionally charged discussions. Biden and Trump, however, maintain a relatively even split between neutral and positive sentiments.

A graph of different colored bars

Description automatically generated  
 Figure-10 Subjectivity plot for last 30 days

Overall, this data suggests that while there is a significant level of engagement with these political figures, much of it does not carry a strong sentiment, either because the discussions are more informational in nature or because Reddit users are reserved in their expression of clear opinions. However, the variance between the top trending posts and those from the last 30 days also indicates fluctuating engagement patterns, which could reflect changing public interests or responses to current events. The higher positivity in the case of Stein, in particular, may warrant further investigation to understand the drivers behind this more favorable sentiment.

The subjectivity results from comparing the top 200 trending Reddit posts with those from the previous month reveal a close distribution between neutral and positive sentiments. Neutral posts typically contain general statements, lacking a clear personal opinion or emotional content, while positive posts are indicative of favorable opinions toward the politicians. The tendency for neutral sentiment to be slightly more prevalent suggests that users often share information or perspectives without a strong personal stance, which is a critical insight for those monitoring public discourse.

The polarity analysis for the top 200 trending Reddit posts and the posts from the last 30 days provides a revealing snapshot of public sentiment toward five politicians. In the analysis of the top 200 posts, Biden leads with a neutral sentiment at 62.28%, closely followed by West with an identical percentage, and Trump with a slightly lower 60.90%. Kennedy and Stein present a more varied picture; Kennedy has a significantly lower neutral percentage at 55.89%, while Stein has even less, at 50.72%. Notably, Stein has the highest negative sentiment at 33.81%, indicating a considerable amount of critical opinion in the discussions surrounding this individual.

|  |  |  |  |
| --- | --- | --- | --- |
| **TOP 200 POSTS** | **NEUTRAL** | **POSITIVE** | **NEGATIVE** |
| **BIDEN** | 62.28% | 24.57% | 13.13% |
| **TRUMP** | 60.90% | 22.63% | 16.46% |
| **KENNEDY** | 55.89% | 20.08% | 24.01% |
| **WEST** | 62.28% | 16.22% | 21.49% |
| **STEIN** | 50.72% | 15.45% | 33.81% |

Table- 3 Polarity for top 200 days

A graph of different colored bars

Description automatically generated  
 Figure-10 Polarity plot for top 200 days

When shifting the focus to the data from the past month, the trends alter slightly. Both Biden and Trump see a neutral sentiment of 58%, but Trump's negative sentiment is slightly higher than Biden's, at 23%. Kennedy and West show an increase in neutrality, at 70%, with Kennedy having a quarter of the sentiment being negative. Remarkably, Stein's entire sentiment in the last month's data is categorized as neutral, with no positive or negative sentiments recorded, which is an unusual trend that might suggest a lack of strong sentiment-driven discussions or could be indicative of data anomalies.

|  |  |  |  |
| --- | --- | --- | --- |
| **LAST MONTH** | **NEUTRAL** | **POSITIVE** | **NEGATIVE** |
| **BIDEN** | 58% | 24% | 18% |
| **TRUMP** | 58% | 19% | 23% |
| **KENNEDY** | 70% | 5% | 25% |
| **WEST** | 70% | 15% | 15% |
| **STEIN** | 100% | 0% | 0% |

Table- 4 Polarity for last 30 days

A graph of different colored squares

Description automatically generated  
Figure-11 Polarity plot for top 200 days

This polarity data underlines the fluctuating nature of public sentiment. While neutrality dominates the overall sentiment, there are significant variations in positive and negative sentiments across different politicians and time frames. These variations could reflect the dynamic nature of political discourse on Reddit, offering crucial insights for political campaign strategists. The apparent critical stance towards certain politicians in recent discussions, particularly in the trending posts, could signal an urgent need for campaign strategy adjustments. Overall, the data underscores the importance of continuously monitoring online sentiment to effectively engage with the electorate and manage political messaging.

Conclusions and Discussions

The polarity and subjectivity data from Reddit posts provide intriguing insights into the public sentiment surrounding five politicians during a critical election cycle. By analyzing the top 200 trending posts and those from the past month, we can discern distinct sentiment patterns that have significant implications.

In terms of polarity, the data reveals that neutral sentiments are predominant across both datasets, which might suggest that Reddit users prefer to discuss politics in a more informative or detached manner, rather than expressing strong opinions. However, when looking at the trending posts, Stein stands out with an exceptionally high negative sentiment, indicating that recent discussions are more critical of this politician. For Biden and Trump, the presence of substantial negative sentiment, particularly in Trump's case, could point to controversial or divisive political positions that resonate with the Reddit community.

Turning to subjectivity, the close balance between neutral and positive sentiments across most politicians indicates a community engaged in a mix of sharing information and expressing supportive opinions. Stein's distinct lead in positive sentiment among the top 200 posts suggests a strong base of supportive users, while the comparatively lower positive sentiment for West could imply less enthusiasm or support within the discussions.

When examining sentiment over time, it is notable that Stein's sentiment in the last month's posts is entirely neutral, a stark contrast to the previous trending data. This could indicate a shift in the public discussion, possibly due to changes in the political landscape or a response to recent events. The data for Kennedy shows an interesting dichotomy with a high level of neutral sentiment, yet also a notable degree of negative sentiment, implying that while discussions may be factual, there is also a significant critique present.

Conclusions drawn from the sentiment analysis reveal the fluid nature of online political discourse. The prevalence of neutral sentiment suggests that while Reddit is a platform for political engagement, users may often share or discuss content without strong emotional investment. However, the presence of both positive and negative sentiments, and their fluctuations over time, highlights the platform's role in shaping and reflecting the broader political narrative.

For political strategists and campaign teams, these findings emphasize the importance of active sentiment monitoring to understand the electorate's pulse. A high neutral sentiment could be an opportunity for shaping opinion, while negative sentiment could signal areas requiring attention or redressal. The ability to respond to these sentiment trends could be pivotal in steering campaign messages and strategies, especially in the agile and rapidly changing environment of social media.

Acknowledgment

We are profoundly grateful to Professor Saman Kumarawadu for his invaluable guidance and unwavering support throughout our research. His data analysis expertise deepened our understanding and enriched our approach to studying public sentiment through social media data. Professor Kumarawadu's detailed feedback and academic support were pivotal in refining our methodologies and enhancing our outcomes. His dedication and willingness to engage deeply with our work have impacted our educational and professional growth.

We also thank Syracuse University for providing us with the necessary facilities and resources to conduct our research. The university's access to advanced research tools and technologies significantly facilitated our data collection and analysis, enabling us to understand data mining topics comprehensively. The supportive academic environment at Syracuse has been instrumental in our project's success, and we are thankful for the continuous encouragement from the university community."

References

[1] Tunggawan and Y. E. Soelistio, "Bayesian Twitter-based prediction on 2016 U.S. presidential election," doi: 10.1109/IC3INA.2016.7863019.,

[2] Ibrahim et al. (2015) Ibrahim M, Abdillah O, Wicaksono AF, Adriani M. Buuzzer “Detection and sentiment analysis for predicting presidential election results in a Twitter nation” 2015.pp.1348-1353.

[3] Oyebode & Orji (2019) Oyebode O, Orji R, “Social media and sentiment analysis: the Nigeria presidential election”, 2019.pp.0140-0146.

[4] Mahmood et al. (2013) Mahmood T, Iqbal T, Amin F, Lohanna W, Mustafa A, “Mining Twitter big data to predict 2013 Pakistan election winner”, 2013.pp.49-54.

[5] Bansal & Srivastava (2018) Bansal B, Srivastava S. “Predicting election with hybrid topic-based sentiment analysis of tweets”, Doi:10.1016/j.procs.2018.08.183.

[6] Borra, Erik, and Bernhard Rieder. "Programmed method: Developing a toolset for capturing and analyzing tweets," pp. 262-278.

[7] Omuya, Erick Odhiambo, et al. "Sentiment analysis on social media tweets using dimensionality reduction and natural language processing."

[8] Neri, Federico, et al. "Sentiment Analysis on Social Media." IEEE, 2012.

[9] Singh, Aaryan, et al. "Sentiment Analysis on User Feedback of a Social Media Platform."