Electoralytics: Data-driven Political Insights

Rahul Patel, Prapti Trivedi, Tanya Garg
Department of Computer Science
New York University
Email: {rp3752, pt2295, tg2520}@nyu.edu

***Abstract*—Elections stand as pivotal events for nations, charting the course for future economic growth and prosperity. The substantial impact of these events prompts extensive discussions among diverse stakeholders on social media, with augmented coverage through new media and various channels, including polls, to gauge public opinions. This study undertakes a rigorous analysis and interpretation of public opinion data linked to the forthcoming 2024 US presidential elections. Employing robust data processing and advanced analytics, our application aims to furnish thorough insights into voter sentiment, preferences, and emerging trends. The study focuses on delivering accurate and in-depth analyses of the current data. This approach equips users with the tools to make informed decisions, fostering a profound understanding of the intricate electoral landscape and helps political strategists to understand the voter sentiment and plan strategies accordingly.**

***Index Terms*—Descriptive Analytics, MapReduce, Hive, Tableau, Sentiment Analysis**

I. Introduction

Elections are a fundamental aspect of democratic societies, providing citizens with the opportunity to choose their representatives and influence the direction of their government. This democratic process allows people to voice their preferences, participate in decision-making, and hold elected officials accountable. With the advent of social media and the increased digitization of information, elections have seen a significant transformation in terms of communication, information dissemination, and public engagement. Social media platforms such as Facebook, Twitter, Instagram, and others provide a global reach, enabling political candidates to connect with a broader audience. Monitoring social media conversations and news articles over time helps track changes in public opinion and response to campaign events or policy announcements. Various studies [5] [6] [7] and more, have used public options from the twitter platform. Nevertheless, the constraints in accessing public opinions on Twitter and Facebook have prompted a necessity to investigate alternative social media platforms such as Reddit.

News data plays a crucial role in the analysis of public sentiments on US presidential elections due to its significance in shaping public opinion and reflecting the prevailing political climate. News outlets are primary sources of information, influencing how individuals perceive political events, candidates, and issues. Analyzing news data allows researchers to understand the information that is being disseminated to the public, which in turn affects public sentiment. Not only this, but news data also captures events and developments related to the election, allowing for event-driven sentiment analysis. Understanding how news sources contribute to sentiment polarization is crucial for gaining insights into the broader political landscape. Analyzing news data alongside sentiment trends can provide predictive insights into potential shifts in public opinion. Early detection of sentiment changes in news coverage may serve as an indicator of evolving voter attitudes.

Another source, polls are conducted to gauge public opinion and predict the possible outcomes of an election. They serve as a snapshot of voter sentiment leading up to election day. Political candidates and parties use polling data to understand public preferences, identify key issues, and tailor their campaign strategies accordingly. This can include adjusting messaging, prioritizing certain policy areas, and targeting specific demographics.

Additional data sources for the examination and analysis of voter sentiments and polarization encompass census and geospatial data, offering insights into population characteristics crucial for comprehending voting patterns, demographic dynamics, and regional variations. Furthermore, historical election results and trends play a pivotal role in discerning patterns, identifying swing states, and understanding the evolution of voting behavior over time.

The objective of this paper is to understand the voter sentiments for the upcoming US Presidential elections 2024, which would help political strategists, politicians and political parties to get insight on how they need to strategies their campaign according to the public sentiment in news, social media and polls. These contributions are effective in answering the following questions:

- What is the nature of discussions surrounding US elections?
- Are there indications of polarization of voting preferences during a particular period?

II. Data Sources

A. Social Media - Reddit

Reddit serves as a diverse network of communities where individuals delve into their interests and passions. Users can concentrate on specific topics by posting content that is voted up or down based on relevance and user preference. Leveraging the unique structure of Reddit, this research employs it as a crucial data source for examining public discussions surrounding the US presidential elections.

The Python Reddit API Wrapper (PRAW) [1] is employed to extract discussions on specific topics, utilizing keywords

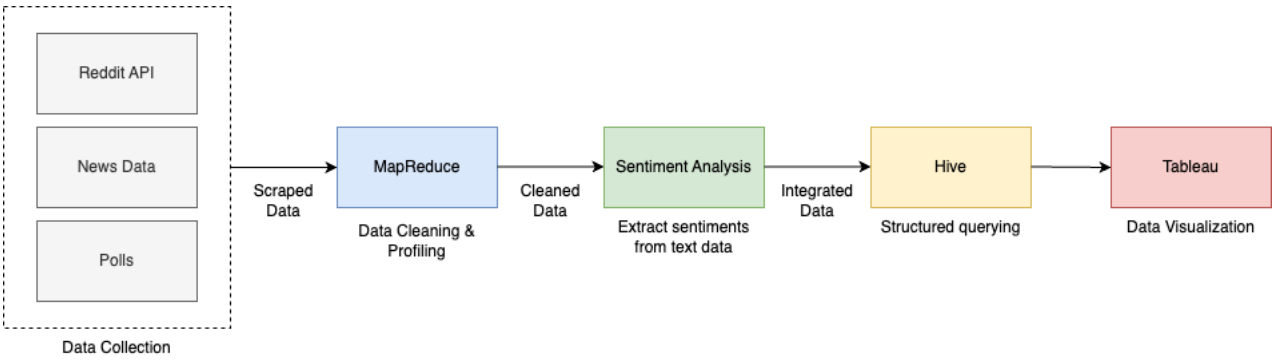

Fig. 1: Data analytics pipeline

like ”2024 US Presidential Elections”, ”Republican Party”, ”Democratic Party”, ”Trump” and ”Biden” This approach resulted in a substantial dataset of 4GB, encompassing posts and comments, providing a comprehensive foundation for the study's analysis.

B. News Data

News data is essential for understanding public sentiments in US presidential elections as it shapes opinions and reflects the changing political climate. This study uses news articles obtained through the Google News API [3] and the News API [2] from various reputable sources like ”The New York Times”, ”NBC News” and ”Reuters”. The combined data, estimated at 3-4 GB, includes titles and content based on keyword searches like ”2024 US Presidential Elections,” ”Democrats,” and ”Republicans” for the months of January to November 2023.

C. Polls

Poll data is crucial for understanding voter sentiments, predicting election outcomes, and guiding campaign strategies. It offers insights into key issues, demographics, and evolving trends, influencing media coverage and validating political approaches. As a real-time pulse on public opinion, poll data plays a pivotal role in shaping the narrative and strategies in US presidential elections. FiveThirtyEight is a prominent polling aggregation website that analyzes and aggregates polling data to provide comprehensive and statistically-driven insights into elections [4]. The 800MB data accumulates latest information about the general election polls from various polling agencies.

III. Methodology

As shown in Fig. 1, the project work is divided into 4 parts. First, the scraped data from all the sources go through a data exploration and cleaning pipeline wherein the inconsistencies in datasets were removed, and the datasets were analyzed and summarized based on their structure, quality, and content using MapReduce. Next, large-language models were used to extract sentiment scores from the text data based on the candidates and political parties favorability. Lastly, the datasets were integrated to carry out further analysis using Hive, along with the generation of descriptive analytics report using Tableau.

IV. Data Cleaning and Profiling

In this study, we used MapReduce which excels in handling large-scale data cleaning and profiling tasks by parallelizing computations across distributed systems. Its ability to process and analyze vast datasets in a scalable and fault-tolerant manner makes MapReduce an effective solution for tasks like identifying data inconsistencies, handling missing values, and generating comprehensive data profiles in big data environments.

A. Reddit

We employed the Reddit API to retrieve posts on the subject of 'US Presidential Elections 2024'. The gathered data encompassed key fields such as posting date, subreddit, title, content, and comments for each post. It is important to note that Reddit imposes constraints, limiting the retrieval to the most recent 250 posts and placing restrictions on the number of requests allowed within a specific time frame.

Next, MapReduce was used to perform data exploration and cleaning. Data exploration helped us to understand the data and we realized there were some anomalies in the comments which we needed to clean. Some of them were:

- Removal of posts and comments that had only emojis or website links since these were of no use to the analysis.
- 'Removed' or 'Deleted' keywords are used as placeholders when a comment is removed by the user or the moderator. Such texts served no purpose and were removed.

Next, data profiling was performed to gain insights into the quality of our data. Fig 2 shows the average frequency of Reddit posts related to 'US Presidential elections' posted every month, and we see that the posts have increased as the elections approach.

For keyword analysis, the dataset was tokenized into words to understand vocabulary and word usage. Fig 3 shows the most frequent words and other words of interest.

A comparison of the appearance of certain parties (Republicans vs Democrats) and Presidential candidates (Donald Trump and Joe Biden), as shown in Table I.

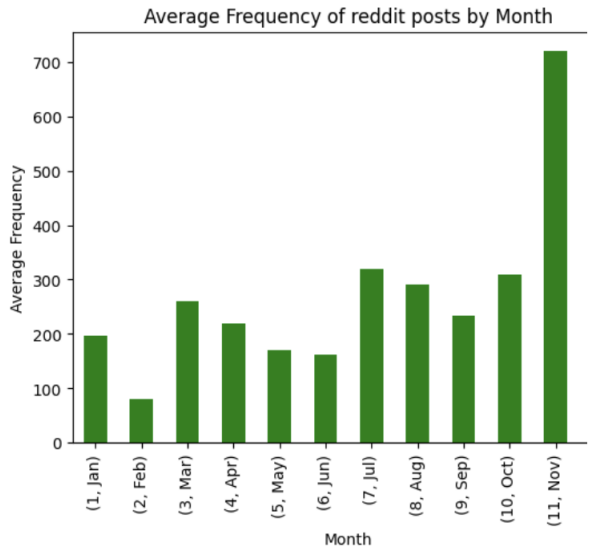

Fig. 2: Average frequency of Reddit posts

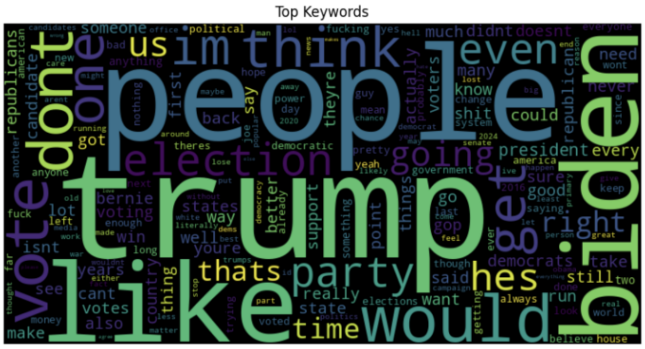

Fig. 3: Top keywords in Reddit comments

B. News Data

This analysis includes news articles obtained from Google News API and News API to extract data related to 'US Presidential Elections 2024'. The extracted data fields include the date of publication, media sources, and headlines.

Next, data exploration was performed using MapReduce to understand the data and find out if there were any anomalies. Based on this, the data was cleaned. Some of the steps include:

- Removal of duplicate entries, for example: multiple instances of the same title due to the occurrence of top news during multiple weeks.
- Handling of multiple line entries: By default, MapReduce considers a line as one record. Instead of separating data based on line breaks, a delimiter was introduced only when "date" or "source" key terms appeared in a line, ensuring that data entries taking more than one line were not broken off at the end of the line.

Further, data profiling was done to gain insights into the structure, quality, and content of the data. Fig 4. shows an estimate of the count of relevant articles from top media sources. To understand the vocabulary and word usage, the most frequent words were identified (fig 5.), and a comparison of the appearance of certain parties (Republicans vs Democrats) and Presidential candidates (Donald Trump and Joe Biden), as shown in Table II.

TABLE I: Keywords-of-interest Comparison

Keyword	Count
Trump	63903
Biden	39153
Republicans	13096
Democrats	12924

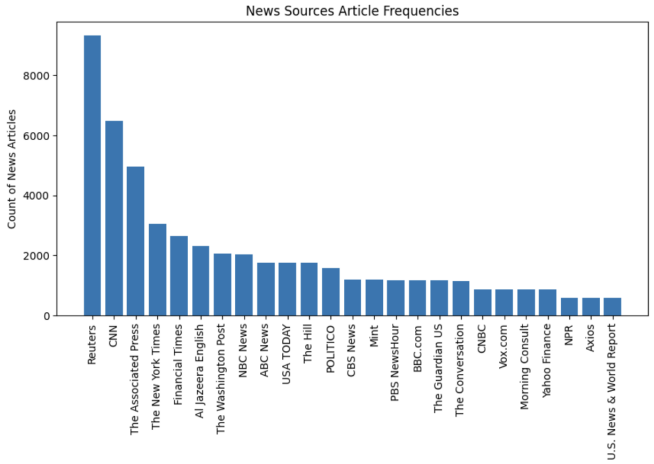

Fig. 4: Count of news articles from top news sources

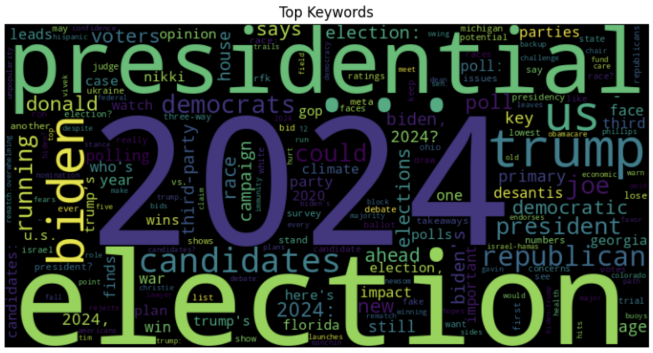

Fig. 5: Top keywords in news articles

C. Polls Data

The poll data extracted from FiveThirtyEight included - candidate details, poll count, region, polling organization details etc.

In the data cleaning phase, we carried out the following steps using MapReduce:

- Removal of duplicate rows

TABLE II: Keywords-of-interest Comparison

Keyword	Count
Trump	17301
Biden	15394
Republicans	6159
Democrats	4943

- Removal of rows with missing values or inconsistent data types
- Aggregation of candidate scores over the different questions asked during a poll.
- Change in date format to UNIX supported format, for compatibility with Hive for downstream analysis.

There are various ways through which polls can be conducted. Some of the popular means used during the 2024 election polls are shown in Fig 6, with online channels being the most popular means of conducting polls.

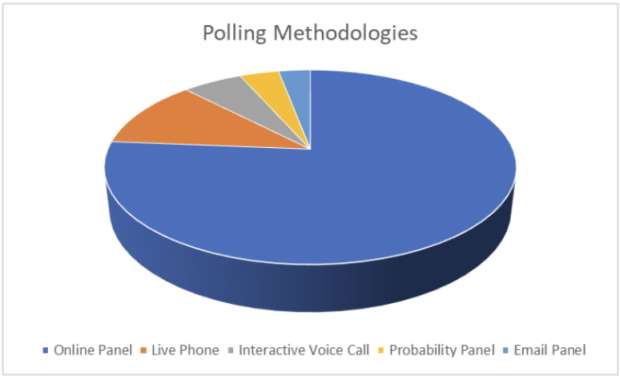

Fig. 6: Polling Techniques

Some of the most prominent polling agencies are shown in Table III, based on the number of polls conducted this election season.

TABLE III: Top Polling Agencies

Polling Agency	Count
Morning Consult	369
Emerson	82
YouGov	67
Redfield & Wilton Strategies	51
Echelon Insights	45

V. Sentiment Analysis

Analyzing public sentiment involves gauging the emotional tone in a text to comprehend how people feel and perceive certain matters. Various methods, such as machine learning, natural language processing, and lexicon-based approaches, can be applied for this purpose. Traditional sentiment analysis typically categorizes output as positive, negative, or neutral, offering broad scores that lack specificity regarding opinions on specific candidates or political parties.

Larger language models, like GPT-3.5, have proven effective in sentiment analysis due to their capacity to understand contextual nuances and intricate meanings in diverse language expressions. These models offer valuable insights into public sentiments on various topics. Employing strategic prompt engineering techniques enables targeted sentiment analysis for more specific insights.

In this study, we employed OpenAI's GPT3.5 chat completion API [10] to provide favorability scores for each news article and Reddit comment with respect to a specific political party and candidate. The prompt used for this analysis is shown in Fig. 7. Here, DATA refers to the text content from the news article or Reddit data.

```
Given this input sentence: [DATA]
Give a score between 0-1 for the following 4 attributes:
1. Favorability of Trump
2. Favorability of Biden
3. Favorability of Democrats
4. Favourability of Republicans

Answer these in the following form between the following tags:
1. <trump_favour> and </trump_favour>
2. <biden_favour> and </biden_favour>
3. <dem_favour> and</dem_favour>
4. <repub_favour> and </rep_favour>
```

Fig. 7: Prompt for sentiment analysis

VI. Analytics

A. Hive

Hive is an SQL-like query language system built on top of Hadoop, that provides a mechanism to manage and query large datasets stored in Hadoop's distributed storage [8].

The public opinion sentiment data from Reddit and News articles were queries to get an average sentiment score for the four terms: Trump's favorability, Biden's favorability, Republican's favorability, and Democrat's favorability, grouped as per date.

An example of the query used to extract average sentiment score from news is shown in fig 8.

```
SELECT
    date,
    AVG(trump_favorability) AS trump_favorability,
    AVG(biden_favorability) AS biden_favorability,
    AVG(republicans_favorability) AS republicans_favorability,
    AVG(democrats_favourability) AS democrats_favourability
FROM
    news
GROUP BY
    date
ORDER BY
   date;
```

Fig. 8: Prompt for sentiment analysis

The resulting table was used to create visualizations using Tableau.

B. Tableau

Tableau plays a pivotal role in data analytics by providing a user-friendly platform for visualizing and interpreting complex datasets, enabling analysts to uncover actionable insights and make data-driven decisions efficiently [9].

We connected Tableau to Hive using the Cloudera ODBC driver. For this study, we created visualizations for the most recent sentiment scores for Reddit and news data, and the polling favorability scores from the polling data.

On comparing the most popular politicians running for the elections, as shown in fig 9, we see that in all three analyses, Joe Biden is leading Donald Trump, and the average favorability score is between 0.4 to 0.5. Reddit particularly shows a significant bias toward Biden.

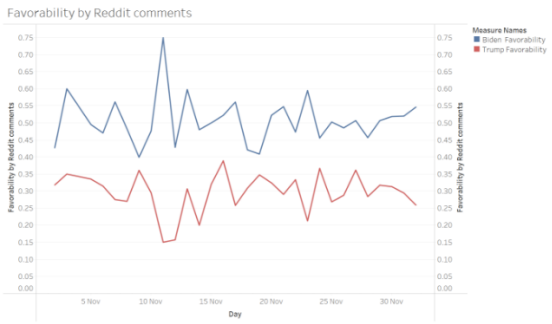

Fig A. Reddit: Candidate favorability

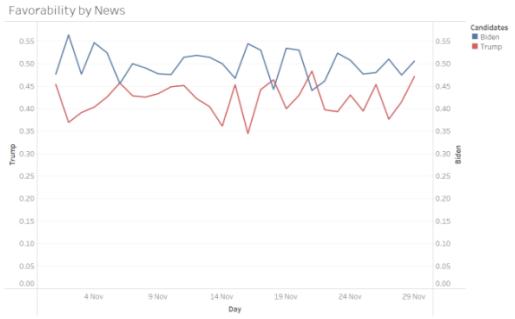

Fig B. News: Candidate favorability

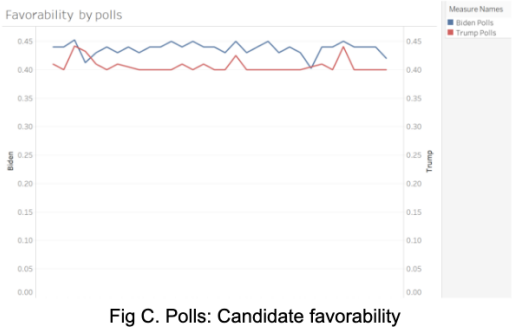

Fig C. Polls: Candidate favorability

Fig. 9: Candidate favorability (Blue: Biden; Red: Trump)

A similar analysis for political parties is shown in figures 10. There is no clear winner like in the case of candidates, however, we do observe some amount of bias towards the Republicans in the news data.

Fig. A: Reddit: Political Party favorability

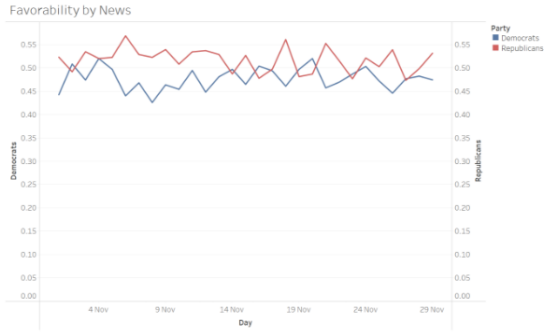

Fig. B: News: Political Party favorability

Fig. C: Polls: Political Party favorability

Fig. 10: Political Party favorability (Blue: Democrats; Red: Republicans)

VII. Conclusion

In conclusion, this paper delves into the examination of public discourse on social media, and added media coverage via various channels, including polls, and news data. The study provides valuable insights into voter sentiment for the impending 2024 US presidential elections. By employing advanced data processing and analytics, the research equips users with the tools to make well-informed decisions and fosters a comprehensive understanding of the intricate electoral landscape. The observed bias towards the Republicans in

news data adds a layer of insight, contributing to the broader understanding of public sentiment. This collective analysis not only addresses the nature of discussions surrounding US elections but also sheds light on potential indications of voting preferences' polarization during specific periods. Ultimately, these contributions aim to empower political strategists, politicians, and parties in crafting effective campaign strategies aligned with the evolving public sentiment in news, social media, and polls.

The current work involves performing analysis on the batch data. This can be extended to streaming data by getting daily insights on voter sentiments. This can be done by setting up a pipeline that can scrape and clean data, perform sentiment analysis, and give out favorability scores in real-time.

REFERENCES

[1] PRAW Documentation, *Python Reddit API Wrapper (PRAW) Documentation*, https://praw.readthedocs.io/en/stable/

[2] News API, *News API - A JSON API for live news and blog articles*, https://newsapi.org/

[3] GoogleNews Python Package, *PyPI - The Python Package Index*, https://pypi.org/project/GoogleNews/

[4] FiveThirtyEight Polls, *FiveThirtyEight - Polls*, https://projects.fivethirtyeight.com/polls/

[5] Gian M. Fulgoni, Andrew Lipsman, and Carol Davidsen, "The power of political advertising: Lessons for practitioners: How data analytics, social media, and creative strategies shape US presidential election campaigns," *Journal of Advertising Research*, vol. 56, no. 3, pp. 239–244, 2016.

[6] Purva Grover, Arpan Kumar Kar, Yogesh K Dwivedi, and Marijn Janssen, "The untold story of USA presidential elections in 2016-insights from Twitter analytics," in *Digital Nations–Smart Cities, Innovation, and Sustainability: 16th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2017, Delhi, India, November 21–23, 2017, Proceedings 16*, pp. 339–350, 2017, Springer.

[7] Sri Hari Deep Kolagani, Arash Negahban, and Christine Witt, "Identifying trending sentiments in the 2016 US presidential election: A case study of Twitter analytics," *Issues in Information Systems*, vol. 18, no. 2, pp. 80–86, 2017.

[8] The Apache Hive Project, *Apache Hive*, https://hive.apache.org/

[9] Tableau Software, *Tableau*, https://www.tableau.com/

[10] OpenAI, *OpenAI API Documentation*, https://platform.openai.com/docs/api-reference/chat