Tracking Political Trends Around US Presidential Election Through Social Media Analysis

Kritika Garg Himarsh Jayanetti Kumushini Thennakoon

October 31, 2024

1 Introduction

In an era where social media platforms serve as both a public forum and a digital newsstand, understanding the evolving political landscape has become more complex and detailed. As the upcoming U.S. presidential election draws closer, platforms like X (formerly Twitter)[1] offer a rich, real-time view into the public's sentiments, concerns, and conversations. This project leverages data collected from the past few days using trending hashtags in the platform. we aim to use keyword analysis, sentiment analysis, and topic modeling techniques [1] powered by large language models (LLMs)[2] to draw conclusions on the current trends of the political background.

Through this analysis, we aim to capture public opinion, uncover emerging topics, and observe the voters' sentiments. Using recent data is essential for capturing the dynamics of the current political climate. Political discourse is highly fluid, especially as candidates campaign intensively, policies are announced, and significant events or news cycles shape public opinion in near real-time. Analyzing short-term data allows us to pinpoint these rapid shifts, helping to reveal immediate concerns, and spontaneous reactions, and possibly identify pivotal moments influencing voter behavior. For instance the presidential debates. Immediate data also enables the detection of micro-trends [3] that could signal larger upcoming shifts in voter priorities, enhancing the relevance of our analysis as the election nears.

Large language models (LLMs) are ideal for this task due to their capacity to handle vast amounts of text data and to interpret nuanced language, sentiment, and context—critical for an accurate portrayal of public opinion. By applying LLMs to Twitter data, we can effectively perform topic modeling, categorizing and prioritizing conversation themes to yield insights into the issues resonating most with voters. This study will not only provide timely insights into voter sentiments but also demonstrate the potential of social media as a valuable tool for real-time political trend analysis.

1.1 objectives

1. To collect and analyze the social media data using an X platform to identify emerging political trends, using hashtags, keywords, and user interactions as indicators.
2. To apply large language models (LLMs) in the text analysis process, aiming to improve the detection and interpretation of trends, sentiment, and underlying patterns in social media content related to political topics.
3. To develop predictive models for short-term trend evolution, focusing on political topics and using insights from LLMs to estimate how these trends might shift in the coming weeks.
4. To explore factors influencing trend longevity or short-lived spikes by examining variables like sentiment shifts, external political events, and influencer or media impact on social media discussions.
5. To identify potential applications and limitations of using LLMs for short-term political trend prediction in social media, providing insights for future work in predictive modeling in fast-evolving social contexts.

[1] https://x.com/

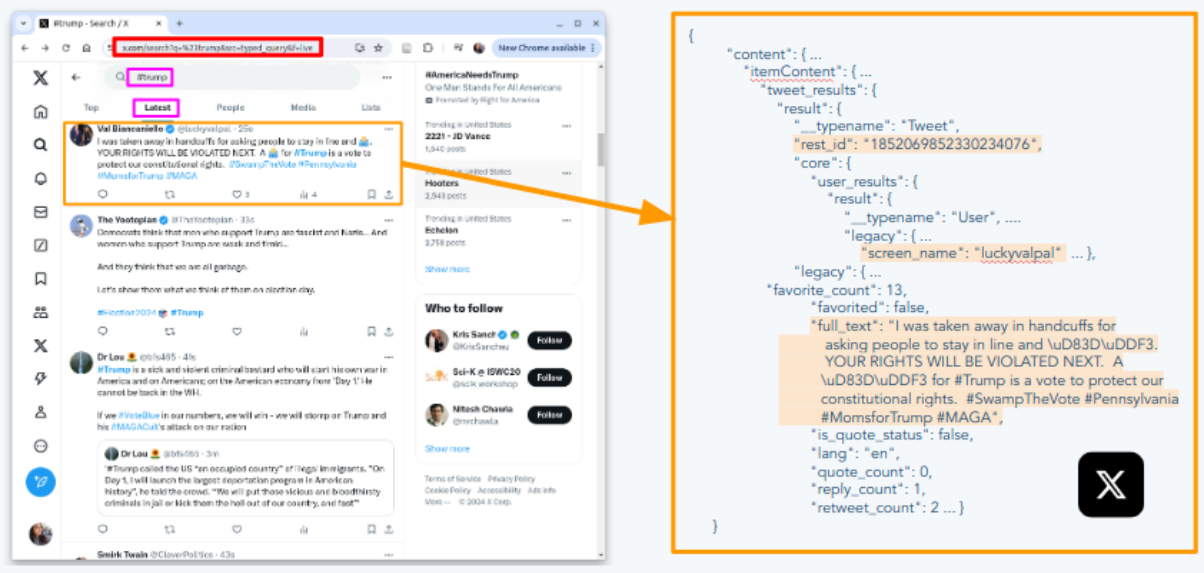

Figure 1: Extracting text data from Tweets

2 Methodology

2.1 Data Collection

In this study, we collected data from X (formerly Twitter) to gain insights into discussions and trends surrounding the U.S. presidential election in 2024. We used X because by observation we understood that X is the platform with the highest user engagement. We did look into platforms like Facebook where the election-related groups did not have much engagement. We focused on seven popular hashtags each representing critical keywords that capture various dimensions of election-related discourse.

1. #donald
2. #trump
3. #kamala
4. #harris
5. #presidentialelection
6. #pennsylvania
7. #russia

We used Selenium Wire Python library[2] to develop an automated web scraping [4] tool to gather tweets. We collected the Tweets in the time interval 2024-10-21 to 2024-10-27 to access the most recent data to come up with a better understanding of the current political situation around the presidential election [5].

2.1.1 Dataset

We were able to extract 3097 Tweets in total for all seven hashtags. Table 1 shows the number of data records we obtained for each Hashtag.

2.2 Data Pre-processing

In the data pre-processing phase, we refined the collected tweets to prepare them for analysis. The following steps outline the key transformations applied to ensure data quality and relevance:

[2]https://pypi.org/project/selenium-wire/

Table 1: Hashtags used for collecting data

Hashtag	Total records
#donald	45
#trump	589
#kamala	619
#harris	600
#presidentialelection	666
#pennsylvania	403
#russia	175

2.2.1 Removing Stopwords

Commonly used but non-informative words (e.g., "and," "the," etc.) were removed to focus on the meaningful content of each tweet. We used the Natural Language Toolkit (NLTK) [6] Python library to handle stopword removal, which helped streamline the text.

2.2.2 Cleaning Special Characters

We removed emojis, and links using regular expressions (regex) to reduce the noise within the dataset.

2.2.3 Removing Duplicates

To prevent bias from repeated content, duplicate tweets were removed, retaining only the first instance of each.

2.2.4 Extracting Hashtags

All hashtags were extracted separately, preserving these contextual keywords for future analysis. This enabled deeper insights into hashtag-specific trends and associations.

After applying these steps, we obtained a clean dataset with a total of 2,742 unique tweets, ready for further processing and analysis.

2.2.5 Obtaining Tweeted Date

We used `https://github.com/oduwsdl/tweetedat` project to obtain the tweeted date using the ID from the collected X posts. We merged the new "tweeted date" column to the created data frame.

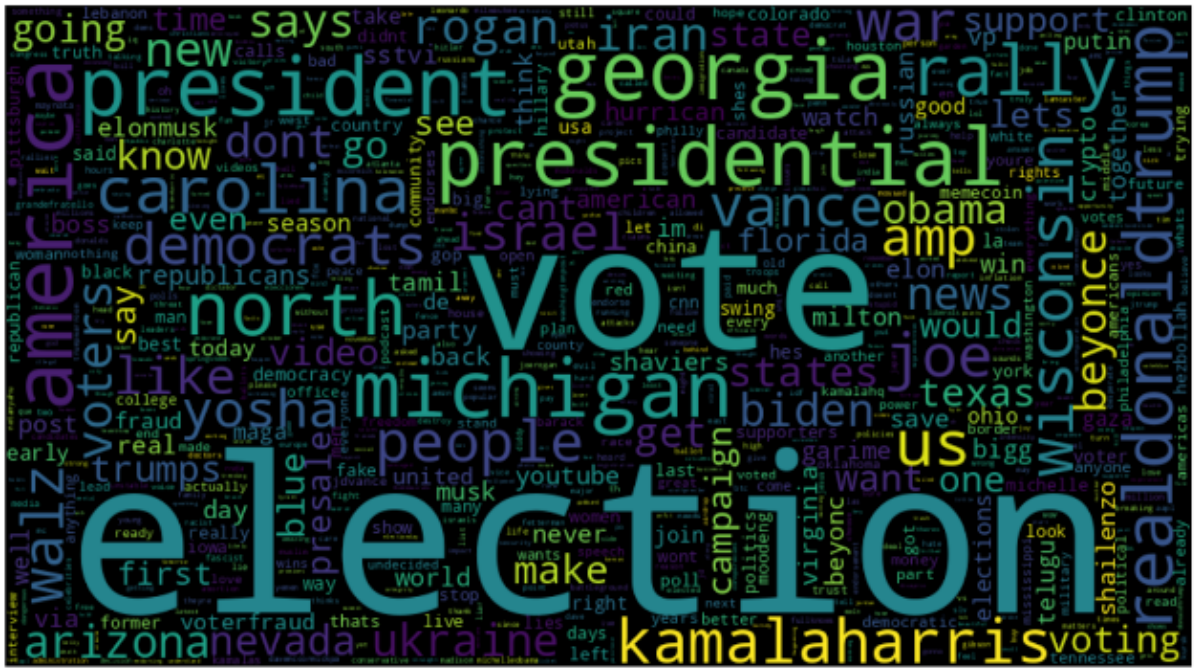

Figure 2: word cloud for overall word frequency

2.3 Tools and Techonologies

Language : Python 3 (version 3.10.12)
Operating System : Unix
Source Control : Git (GitHub)
Libraries : NumPy[3], scikit-learn[4], matplotlib[5], seaborn[6]
external projects : VADER [7] for sentiment analysis, Tweetedat [8] for extracting tweeted data
LLM : Llama 3 [9]

2.4 Exploratory Data Analysis

We did an exploratory data analysis to get a better understanding of the collected dataset. First, we checked the overall word frequency to see the keywords used by the public to express their opinion. Here we removed the words that we used as our hashtags. Then we observed the top 10 hashtags in the Tweets that we collected to understand the usage of hashtags by the public.

* Overall word frequency
* Overall hashtag top10
* Top engaging posts

2.5 Sentiment Analysis

For sentiment analysis on Twitter data, we used the VADER (Valence Aware Dictionary for Sentiment Reasoning) model, a tool specifically designed for the informal language and emotive expressions common in social media. VADER outputs four scores: positive, negative, neutral, and compound. The compound score, a normalized metric ranging from -1 (extremely negative) to +1 (extremely positive), captures the overall sentiment polarity by aggregating the intensities of each sentiment type. We selected the compound score for its ability to provide a holistic view of sentiment within each tweet, making it ideal for distinguishing between positive, neutral, and negative tones. Using a threshold of 0.05, we categorized tweets: scores above 0.05 indicated positive sentiment, scores below -0.05 signified negative

[3]https://numpy.org/
[4]https://scikit-learn.org/stable/
[5]https://matplotlib.org/
[6]https://seaborn.pydata.org/

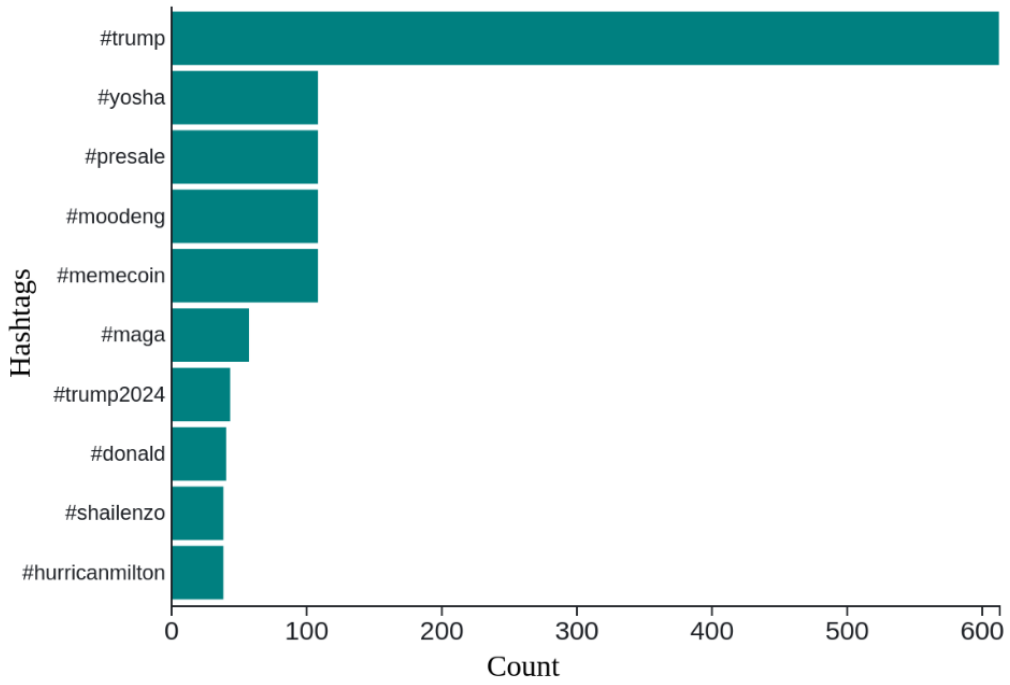

Figure 3: Overall top hashtags after removing the hashtags used for data collection.

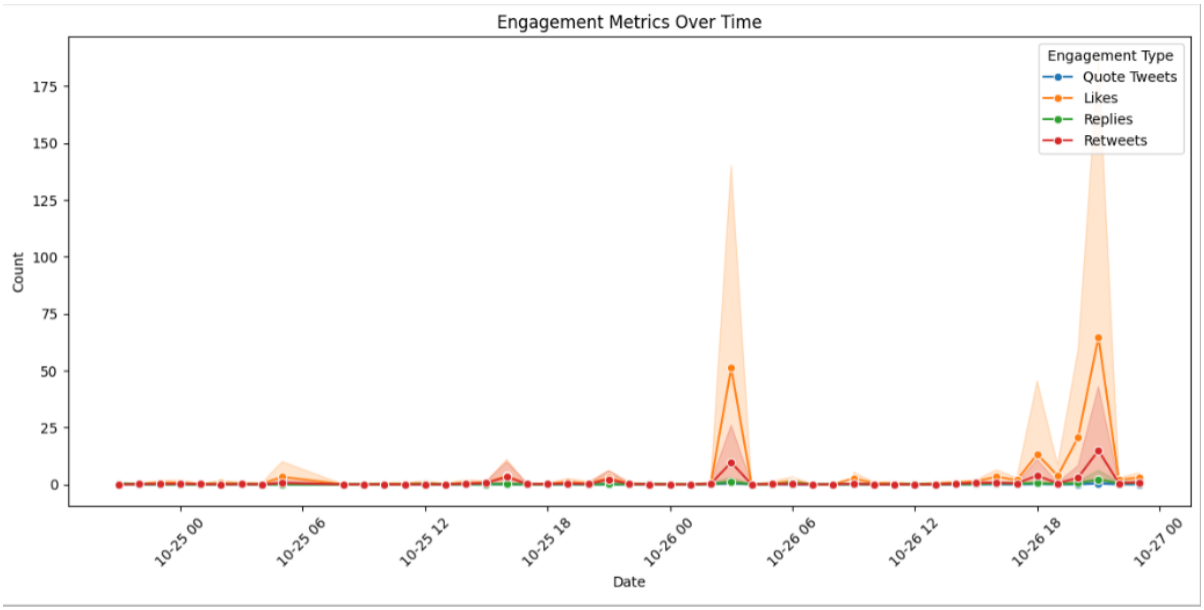

Figure 4:

sentiment, and scores within this range were considered neutral. This approach ensured a clear and balanced sentiment classification across our dataset.

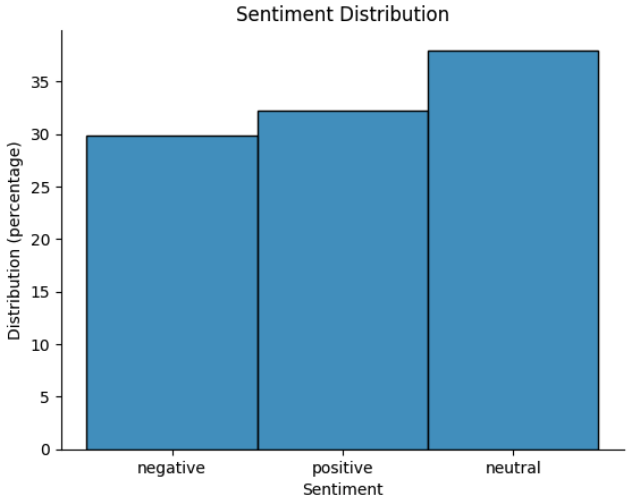

Figure 5: Add caption

*

2.5.1 Sentiment distribution

We analyzed the distribution of sentiment and found that it consists of 38% neutral (1041 out of 2742), 32% positive (884 out of 2742), and 30% negative (817 out of 2742). Having a higher percentage of neutral sentiment could suggest that a portion of the population is indifferent or ambivalent towards political issues, It could also mean that some individuals may not have formed strong opinions or may be undecided on specific topics. The relatively high percentage of positive sentiment (compared to negative sentiment) could indicate a general sense of optimism or support for certain political figures, policies, or events. However, it is also possible that positive framing or spin is being used to shape public opinion, especially by politicians and their supporters. While the percentage is lower, negative sentiment can be highly influential, especially if it comes from vocal and active users.

2.5.2 Sentiment for each hashtag

We conducted a sentiment analysis of tweets for specific hashtags. Given our limited dataset, not all charts yielded significant insights; however, we focused on #kamala (Figure 6) and #presidentialelection (Figure 7), as these provided more substantial analytical value.

Figure 6 shows the sentiment distribution of tweets over time for #kamala tweets. It shows the number of tweets categorized as positive, negative, and neutral in separate lines. This reveals that the sentiment of the tweets fluctuates over time. There are periods where positive sentiment dominates, followed by periods with a higher proportion of negative sentiment. The neutral sentiment appears to remain relatively stable throughout the period. If data on Donald Trump was available to us (as our data is limited), it would have been beneficial to analyze similar trends. This would have enabled a comparative sentiment analysis of various political figures and parties, offering insights into their relative popularity and support.

Figure 7 shows the sentiment distribution of tweets over time for #presidentialelection tweets, there are periods where positive sentiment dominates, followed by periods with a higher proportion of negative sentiment. There is a peak at 2:00 PM, October 25, 2024. We found out that there are trending election topics including tight polling in battleground states like Arizona and Wisconsin, rallies from candidates Harris and Trump, and comments from celebrities stirring debate. Election security concerns and international perspectives on a possible Trump victory were also widely discussed. Both of these major national polls released on October 25th show an extremely close race between Harris and Trump, with the candidates essentially tied. This indicates the 2024 presidential election remains highly competitive heading into the final days of the campaign.

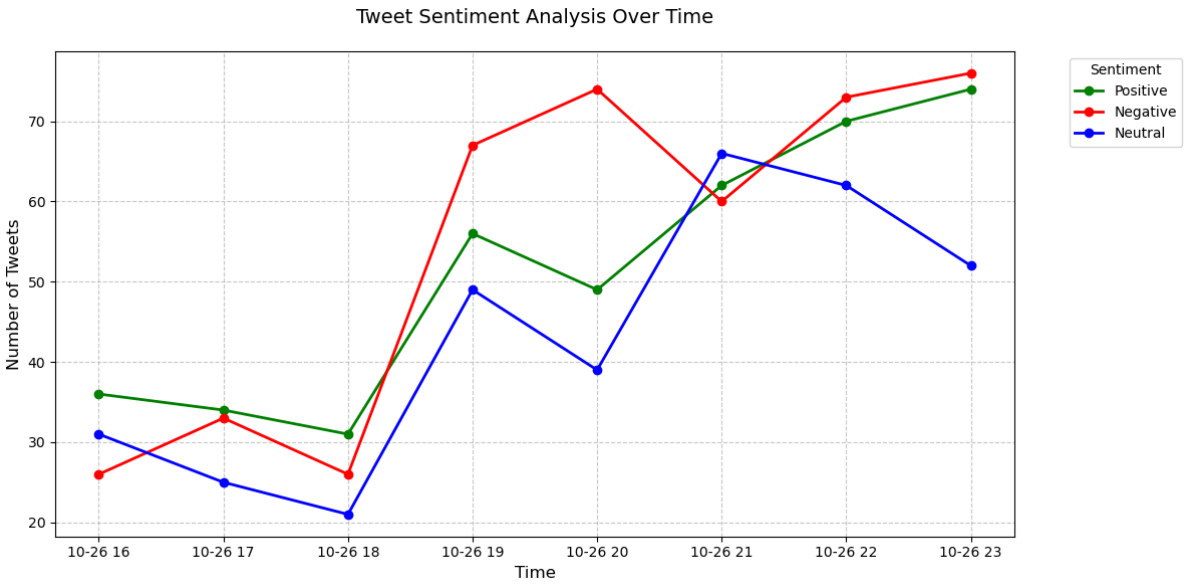

Figure 6: Add caption

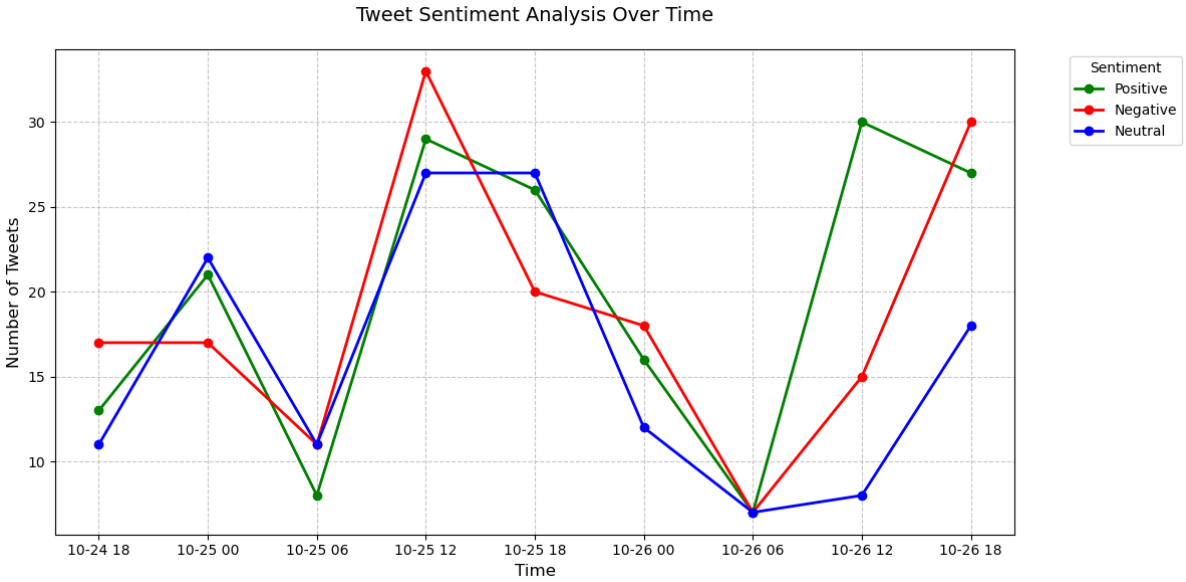

Figure 7: Add caption

2.5.3 Sentiment vs. Engagement

We also looked into engagement matrices by Sentiment. The heatmap in Figure 8 visualizes the distribution of engagement metrics (likes, retweets, replies, and quotes) across three sentiment categories: positive, neutral, and negative. Warmer colors (red and orange) indicate higher engagement, while cooler colors (light pink and white) indicate lower engagement. The heatmap shows that positive sentiment tweets generally receive the highest engagement, particularly in terms of likes and retweets. This suggests that positive content tends to resonate more with the audience and is more likely to be shared and liked. Neutral sentiment tweets show moderate engagement across all metrics, indicating that they attract attention but to a lesser extent than positive content. Negative sentiment tweets generally receive the lowest engagement, suggesting that they may not be as widely shared or liked. However, there is a noticeable spike in replies for negative tweets, which could indicate that they spark discussions and debates.

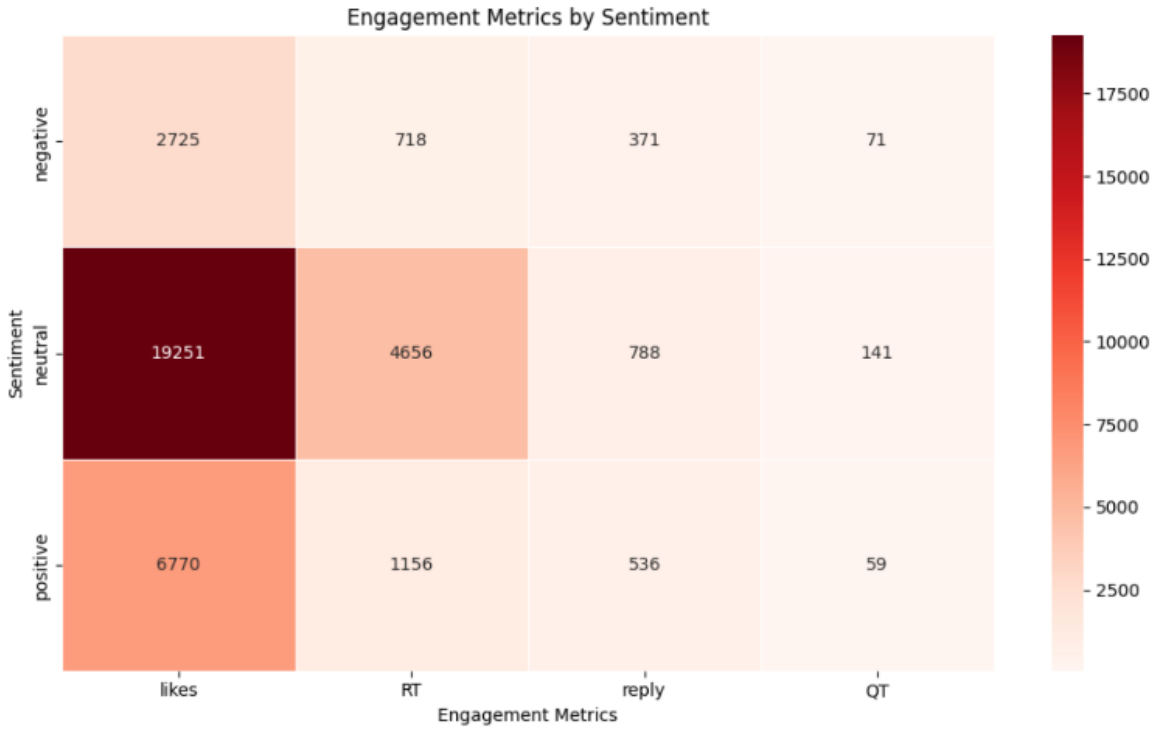

Figure 8: Add caption

2.6 Topic Modeling

We performed topic modeling on our Twitter data to identify trending topics and insights on the US presidential election. We first segmented the dataset into three categories based on sentiment analysis, which allowed us to understand the emotional context of the discussions before delving into the topics themselves. This dual approach enhances the interpretability of the results, as it links sentiment with thematic content, providing a more comprehensive view of public discourse.

To conduct the topic modeling, we leveraged recent advancements in natural language processing, particularly large language models (LLMs). We employed BERTopic, a modular topic modeling technique that utilizes LLMs to fine-tune topic representations. BERTopic effectively creates clusters and topics from the input data, forming a basis for deeper analysis. However, given the computational limitations associated with processing extensive document collections, we incorporated vector databases for efficient searching without requiring direct input of all documents to the LLMs. [7]

To enhance performance, we also implemented quantization techniques to reduce the size and computational demands of the LLMs. For instance, employing 8-bit quantization allows us to significantly decrease the memory footprint of models like Llama 2, facilitating faster and more manageable analyses. After creating the initial topics using BERTopic, we fine-tuned the quantized LLMs to distill and refine the information into more accurate topic representations.

This approach of performing topic modeling post-sentiment analysis not only uncovers emerging themes within the data but also contextualizes these themes within the emotional landscape of public opinion. By combining the strengths of BERTopic with quantized LLMs, we achieved a balance between efficient topic creation and nuanced topic representation, ultimately enabling more effective analysis of public sentiment surrounding the 2024 US election.

Before initiating the topic modeling process, we first downloaded the necessary pre trained large language models (LLMs) from the Hugging Face server. Specifically, we obtained the

`OpenHermes-2.5-Mistral-7B-GGUF and dolphin-2.7-mixtral-8x7b-GGUF` files. These files contain the essential models that will support our topic modeling task. Once the files are downloaded and saved in the current directory, we proceed to load our quantized LLMs using the llama_cpp library. This library is specifically designed to facilitate the operation of quantized LLMs, which have reduced size and computational requirements, making them more manageable for our analysis.

[7] https://xcelore.com/topic-modelling-with-quantized-llama-3-bertopic/

In our implementation, we utilized two representation models: KeyBERT and LlamaCPP. KeyBERT is a fast keyword extraction model, while LlamaCPP employs the quantized LLM we loaded earlier. To guide the LLM's response generation, we defined a prompt containing placeholders for the documents and associated keywords for each topic. During the execution of the model, the LLM fills these placeholders with the actual content, ensuring relevant and contextually appropriate outputs.

After training the BERTopic model, we can display the identified topics. Each topic is represented by a unique ID, accompanied by a list of the top words that characterize the topic. This structured approach allows for a comprehensive understanding of the themes present in the dataset, providing insights into public sentiment and discourse surrounding the 2024 US election.

3 Results and Discussion

3.1 Topic Analysis on Positive Topics

Figure 9 shows the trending topics related to the positive sentiment. The two main topics are as below:

1. "Politics and Encouraging People to Vote": This topic cluster likely includes tweets related to political campaigns, election news, voter mobilization efforts, and calls to action for people to participate in the democratic process.

2. "Yosha Crypto Presale": This topic cluster likely focuses on discussions and promotions related to the presale of a cryptocurrency known as Yosha. It includes tweets about the cryptocurrency, tokenomics, team, and investment opportunities. We also found these hashtags as shown in Figure 3

3.2 Topic Analysis on Negative Topics

Figure 10 shows the trending topics related to the negative sentiment. The three main topics are as below:

1. Politics and Violence: This topic highlights discussions surrounding political events that involve violence or threats of violence.

2 USA Politics: This cluster focuses on news and discussions related to US politics, potentially including debates, elections, or policy discussions.

3. Israel-Iran Conflict: This topic centers on tensions and conflicts between Israel and Iran, including military actions, political rhetoric, and regional implications.

4 Conclusions

5 Limitations

Data collection limitations and challenges

Using Large Language Models (LLMs) for topic modeling on election-related text data offers many insights but also presents notable challenges. One key limitation is bias within LLMs, as models trained on internet data may reflect existing political biases, potentially skewing results. Additionally, limited real-time relevance poses a challenge, as fixed training datasets may not include recent events, causing LLMs to miss timely election developments.

LLMs also struggle with specificity and contextual accuracy. They often generalize topics broadly, potentially overlooking the nuanced or specific issues in election discourse. This is further complicated by the presence of sarcasm, slang, and named entities on social media, which LLMs may misinterpret, leading to less accurate topic classifications.

Furthermore, computational limitations make analyzing large-scale social media data resource-intensive and can restrict scalability. Election-related data often contains noise, including spelling variations and bot-generated content, which can distort results if not properly filtered. Finally, ethical and interpretability concerns arise due to the black-box nature of LLMs. The lack of transparency can hinder the interpretability of results, while ethical concerns include the risk of inadvertently amplifying polarizing content.

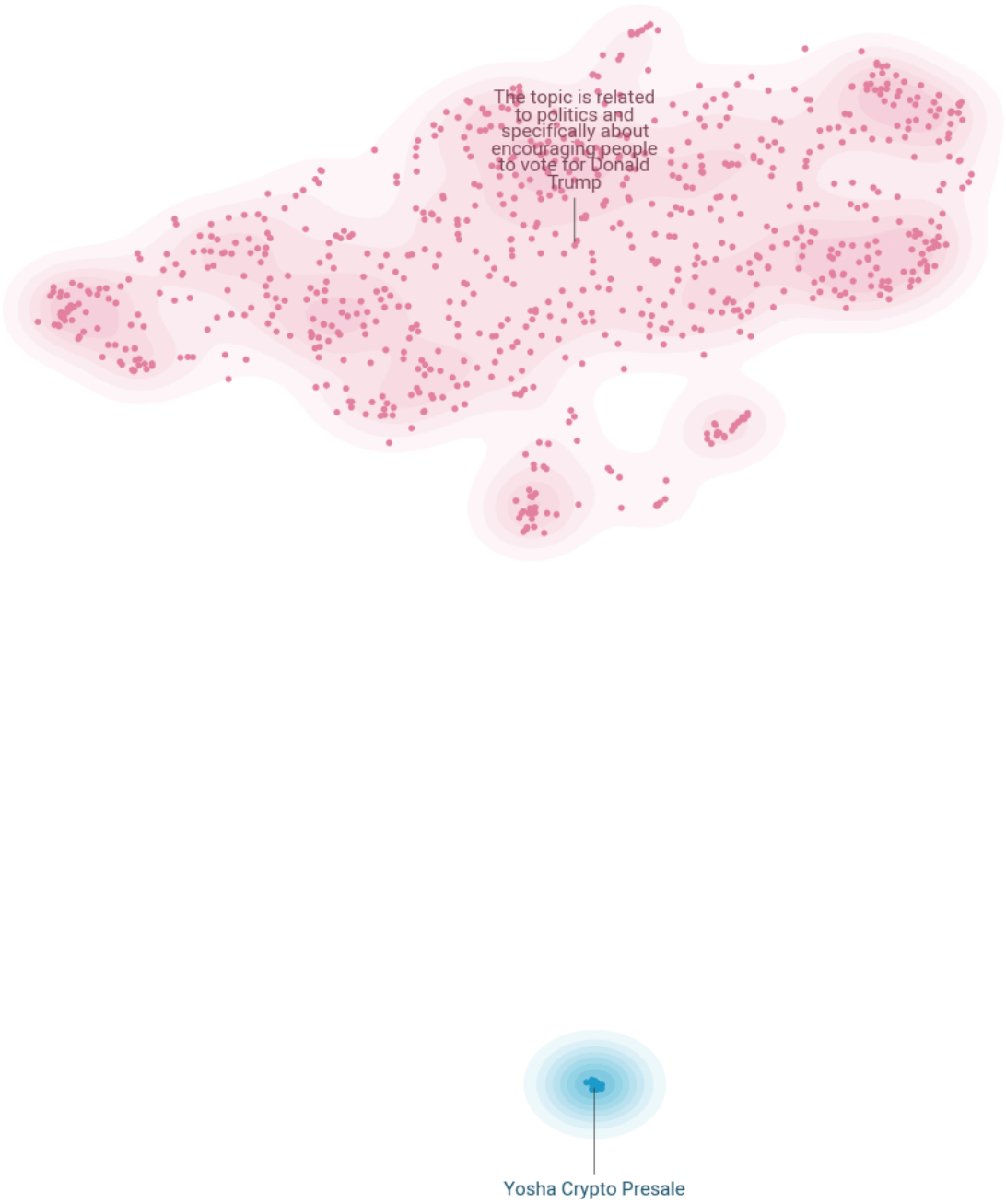

Figure 9: Add caption

6 Future Work

For future research, we aim to incorporate additional text data sources, specifically focusing on Reddit. Additionally, we will continue to gather more Twitter data to enhance our analysis.

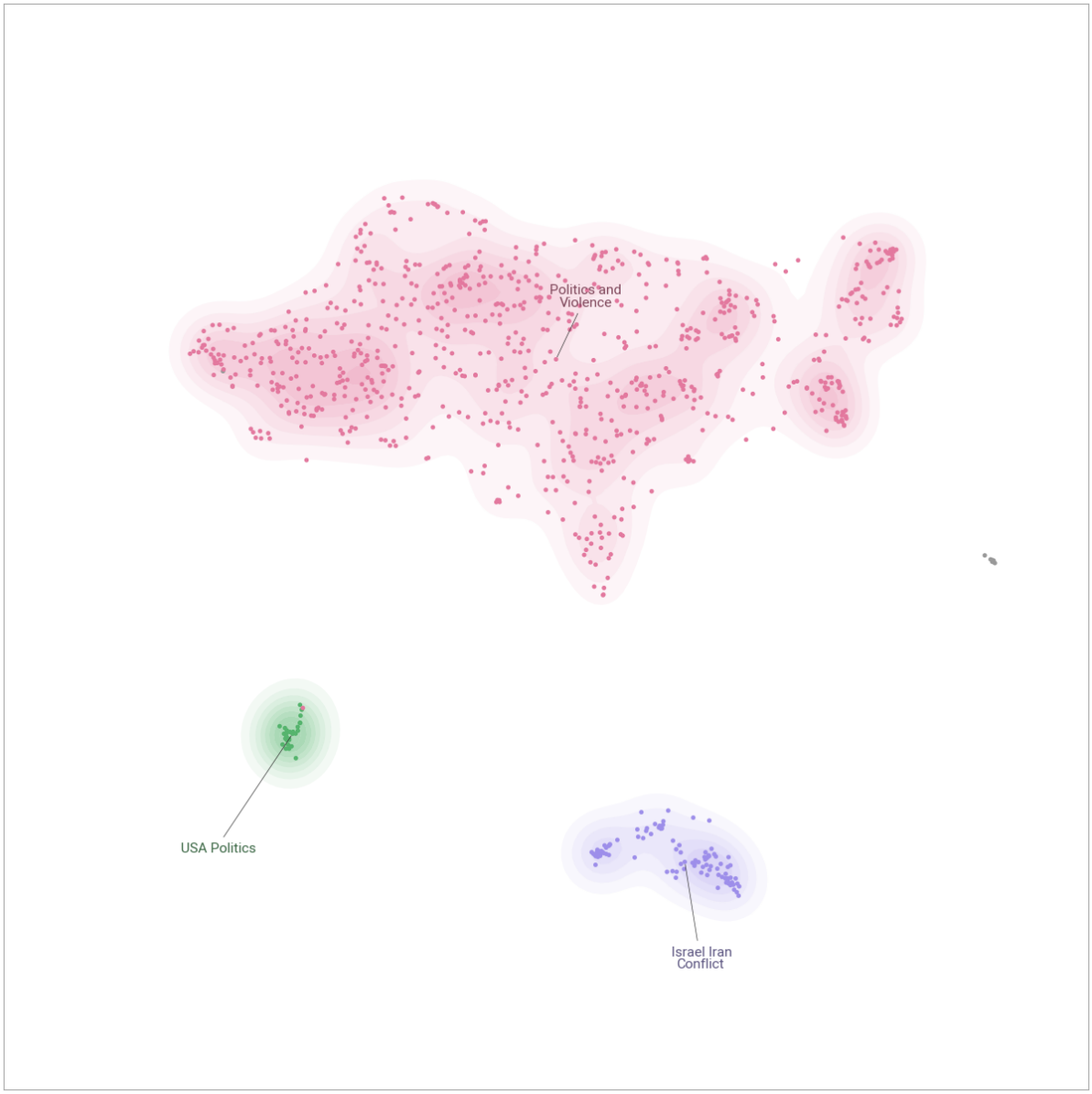

Figure 10: Add caption

References

[1] Vayansky, I. & Kumar, S. A. A review of topic modeling methods. *Information Systems* **94**, 101582 (2020).

[2] Minaee, S. *et al.* Large language models: A survey (2024). URL `https://arxiv.org/abs/2402.06196`. `2402.06196`.

[3] Penn, M. *Microtrends Squared: The New Small Forces Driving Today's Big Disruptions* (Simon and Schuster, 2018).

[4] Mitchell, R. *Web Scraping with Python: Collecting More Data from the Modern Web* (O'Reilly Media, Inc, 2018).

[5] BBC News. Covid: Twitter suspends Naomi Wolf after tweeting anti-vaccine misinformation. `https://www.bbc.com/news/world-us-canada-57374241` (2021).

[6] Bird, S. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 69–72 (2006).

[7] Hutto, C. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, vol. 8, 216–225 (2014).

[8] Sawood Alam, M. K., Nauman Siddiqui. GitHub - oduwsdl/tweetedat: TweetedAt tells the time of a tweet based on its tweet id — github.com. `https://github.com/oduwsdl/tweetedat` (2019). [Accessed 31-10-2024].

[9] Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).