Tracking Political Trends Around US Presidential Election Through Social Media Analysis

Kritika Garg Himarsha Jayanetti Kumushini Thennakoon

January 27, 2025

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 conclude 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/

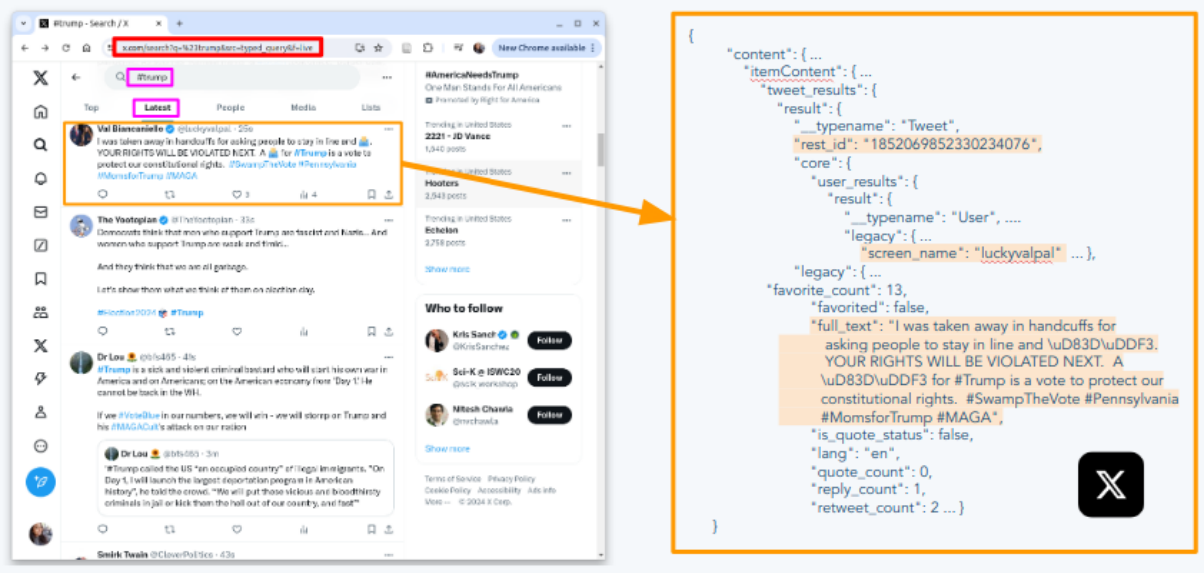

Figure 1: Extracting text data from Tweets

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 2024 US presidential election. Figure 1 shows the data extraction step from the Tweets.

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.

In the data pre-processing phase, we cleaned the collected tweets. First, we removed the Stopwords (e.g., "and," "the") to emphasize the meaningful content of each tweet, using the Natural Language Toolkit (NLTK) Python library[5]. Next, we removed the special characters, including emojis and links using regular expressions. We removed the duplicate tweets by keeping only the first instance of each

[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

tweet. In order to analyze the hashtags that were used, we extracted hashtags. Additionally, we extracted the date of the tweet by using TweetedAt[3]. After this cleaning and pre-processing stage, we were left with 2,742 unique tweets.

2.2 Tools and Techonologies

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

2.3 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. Figure 2 shows the word cloud for the overall word frequency, Figure 3 shows the top hashtags observed in the dataset, and Figure 4 shows engagement metrics over time.

2.4 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 sentiment, and scores within this range were considered neutral. This approach ensured a clear and balanced sentiment classification across our dataset. Figure 5 shows the sentiment distribution for overall data.

2.4.1 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

[3]https://github.com/oduwsdl/tweetedat
[4]https://numpy.org/
[5]https://scikit-learn.org/stable/
[6]https://matplotlib.org/
[7]https://seaborn.pydata.org/

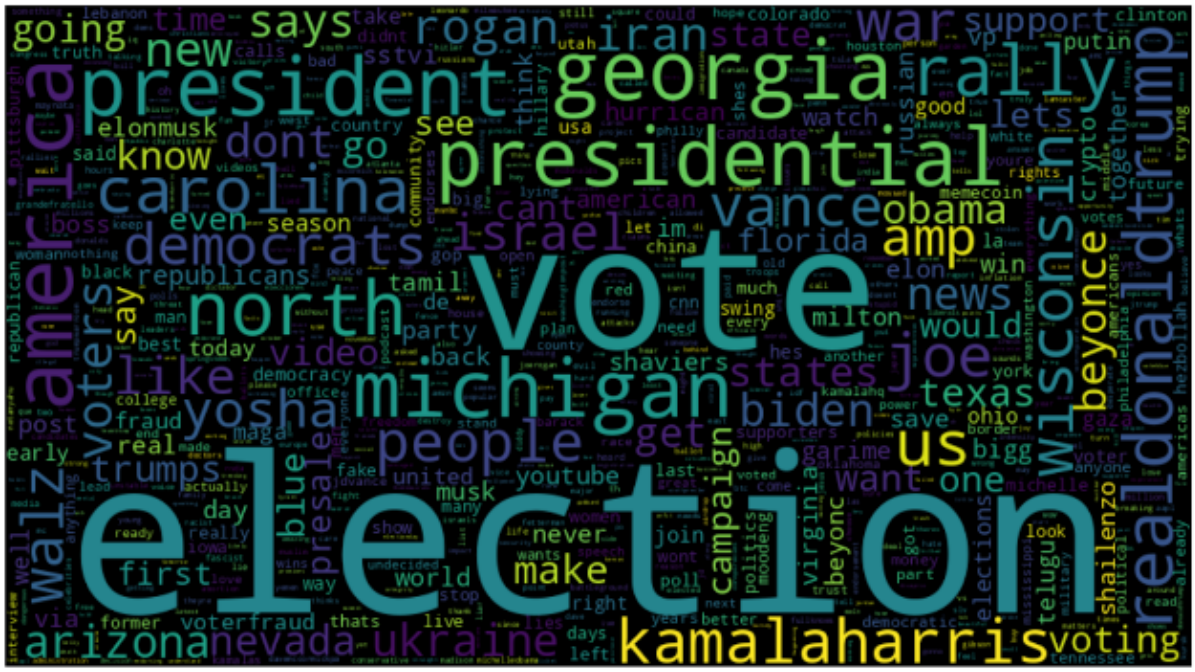

Figure 2: word cloud for overall word frequency

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

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.

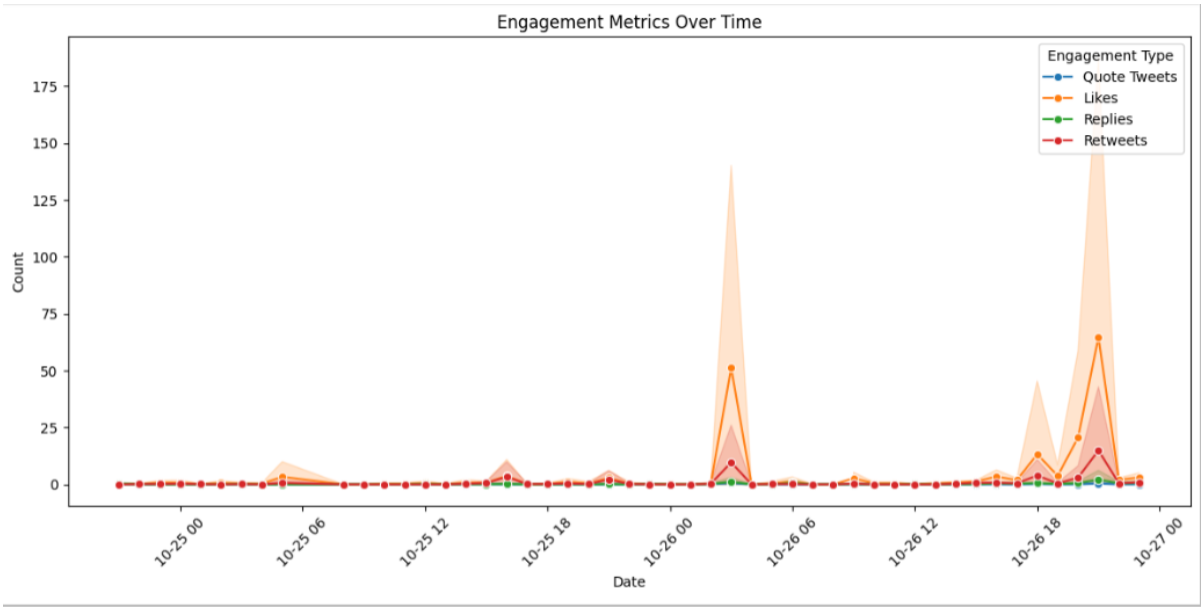

Figure 4: Engagement metrics over time.

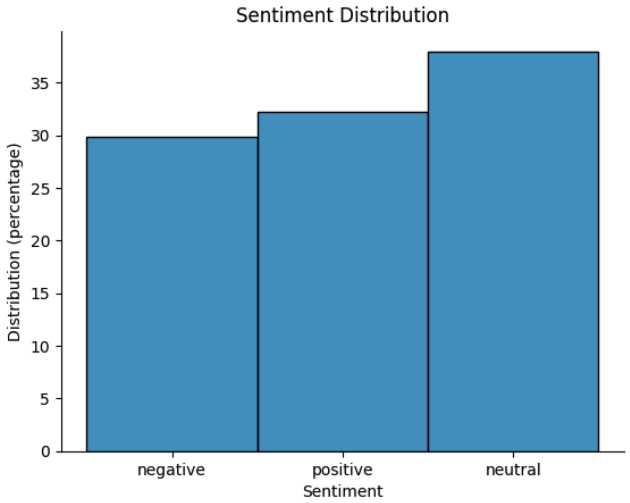

Figure 5: Analyzing the distribution of sentiment for the overall dataset.

2.4.2 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

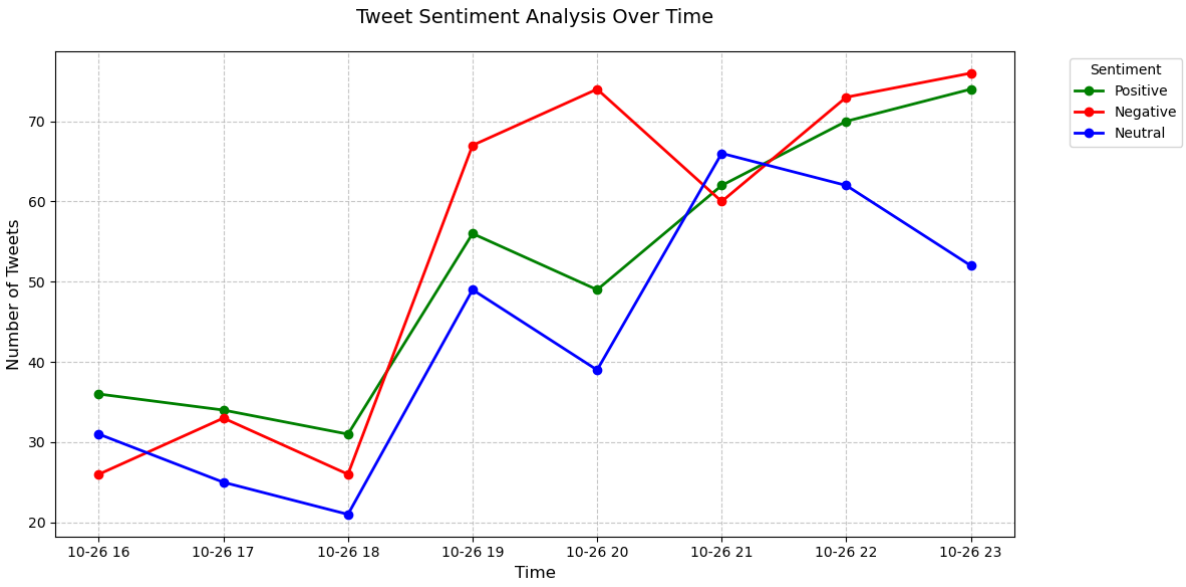

Figure 6: Sentiment distribution of aggregated tweets for #kamala and #harris hashtags over time.

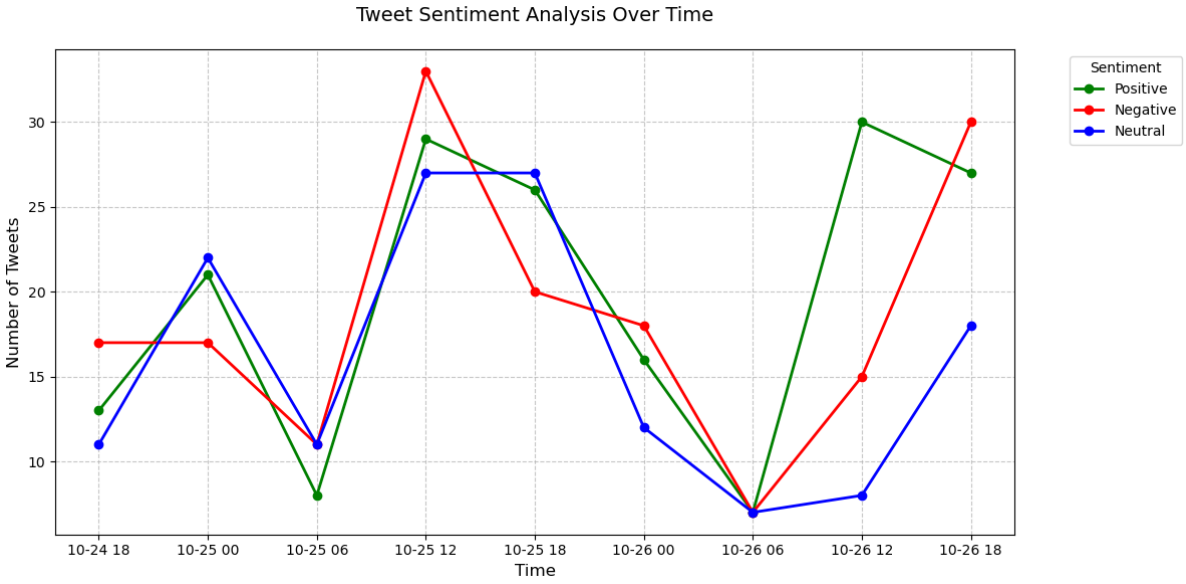

Figure 7: Sentiment distribution of tweets over time for #presidentialelections tweets.

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.

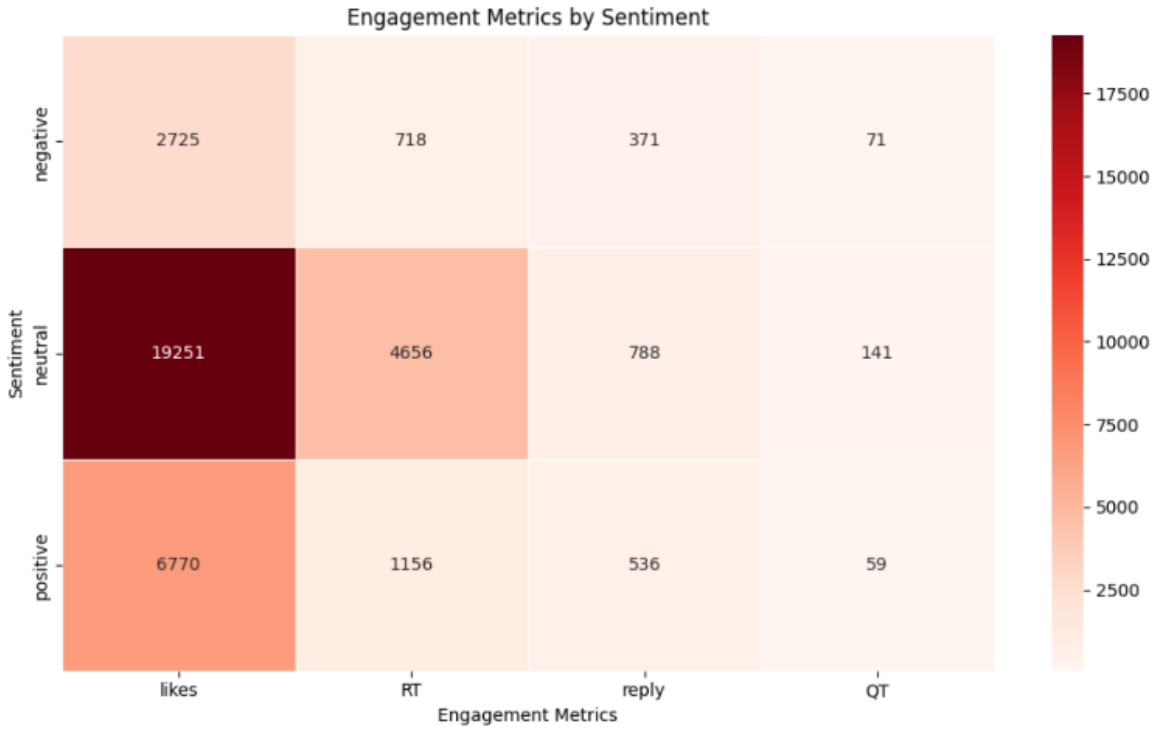

Figure 8: Distribution of engagement metrics across the 3 sentiment categories.

2.4.3 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.

2.5 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. [8]

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.

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

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.

We transform tweets into numerical representations using Sentence Transformers, specifically the BAAI/bge-small-en-v1.5 model. These models are particularly effective for clustering tasks as they efficiently generate document or sentence embeddings. Pre-calculating embeddings for each document accelerates exploration and enables quick iteration over BERTopic's hyperparameters when needed.

Once we obtain the numerical representations, we reduce their dimensionality. Clustering models often face challenges with high-dimensional data due to the curse of dimensionality. In BERTopic, UMAP is the default dimensionality reduction technique. UMAP is ideal because it preserves both the local and global structure of the dataset, retaining critical information necessary for clustering semantically similar documents.

We define the sub-models—UMAP for dimensionality reduction and HDBSCAN for clustering. The topic modeling process with BERTopic involves extracting embeddings using Sentence Transformers, reducing dimensionality with UMAP, clustering reduced embeddings with HDBSCAN, and fine-tuning topic representations using Llama-3-8b-instruct.

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

3.1 Topic Analysis on Positive Sentiment Tweets

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 Sentiment Tweets

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.

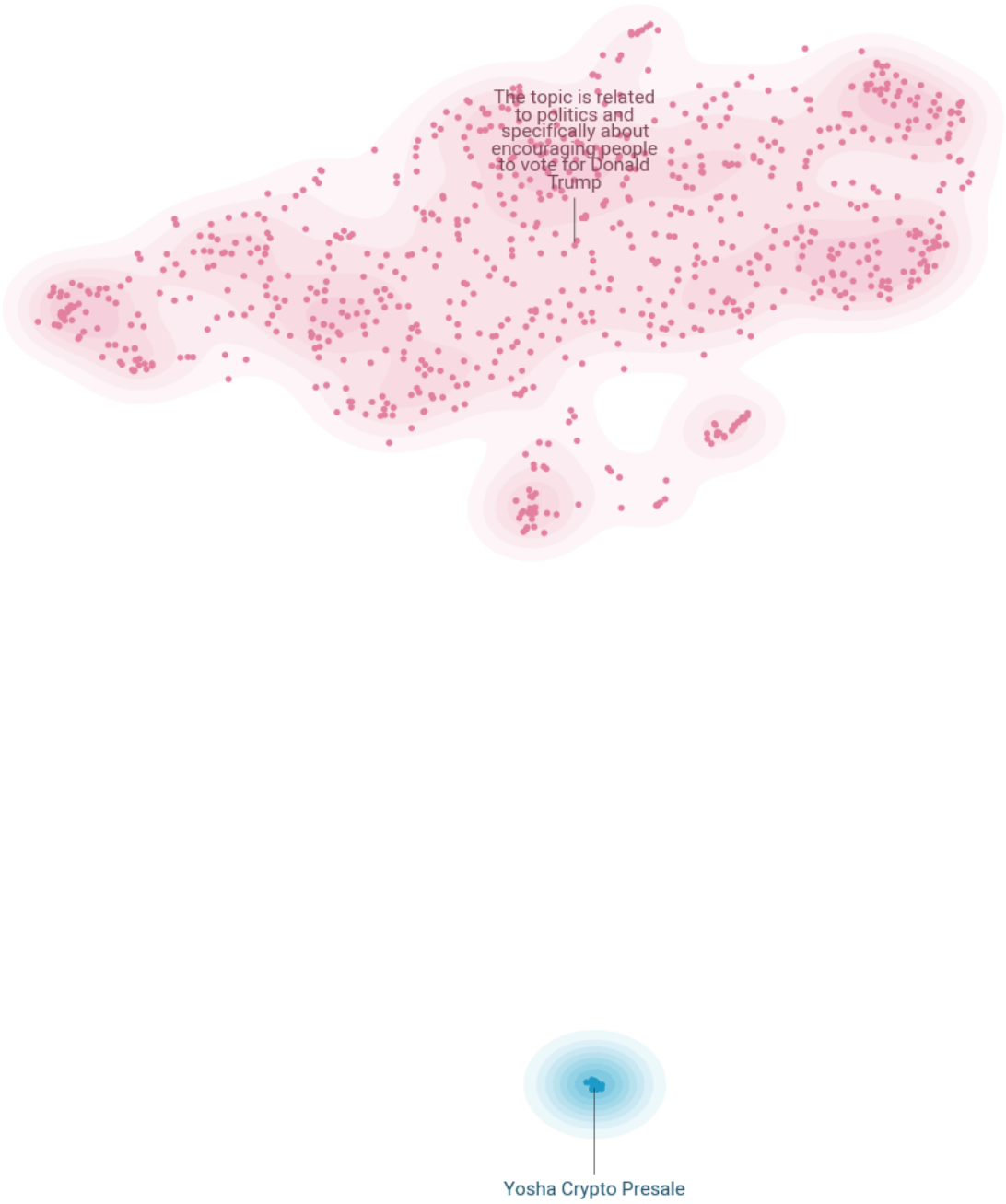

Figure 9: Trending topics related to the positive sentiment.

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

3.3 Topic Analysis on Neutral Sentiment Tweets

Figure 11 shows the trending topics related to neutral sentiment. The ten main topics are as below:

The topic analysis of neutral sentiment tweets revealed nine topics in the context of US election based discussions. The topics were ranging from 2024 US presidential candidates to other general news about the election. A significant portion of the conversation centers on Kamala Harris, including both direct discussions about her role as Vice President and broader public opinions, media narratives, and political debates surrounding her. Related discussions include her role in the election, her policies, and her impact on voter sentiment, often in comparison to former President Donald Trump.

Other discussions are more locally focused, especially on Philadelphia and Pennsylvania, where themes

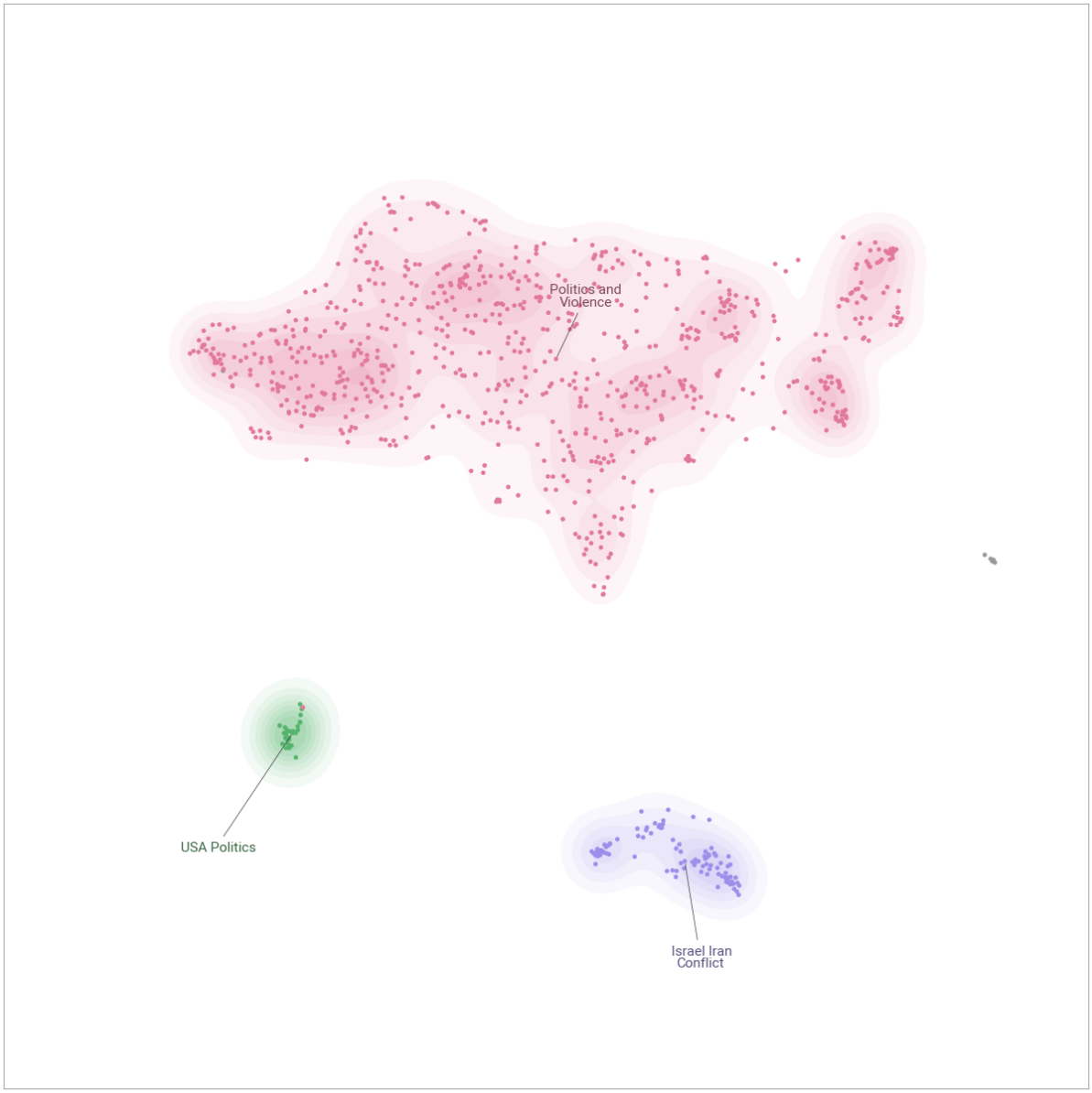

Figure 10: Trending topics related to the negative sentiment.

related to the history and cultural heritage of the region emerge alongside current political engagement. Support for Donald Trump, particularly among veterans, also stands out, with discussions often revolving around national security and veteran issues. Additionally, the analysis reveals conversations about the ongoing geopolitical tensions between Israel and Iran over missile technology, which, though international, is relevant due to its implications for U.S. foreign policy in an election context.

Overall, these topics reflect the blend of local, national, and international concerns that shape election discussions, providing insights into public sentiment, cultural identity, and political affiliations across multiple spheres.

4 Discussion

Some topics within the positive sentiments and negative sentiment data contain a large, cohesive cluster. This clustering behavior can arise because dense embeddings, especially from models like LLaMA, can create clusters that appear large due to their internal variance. UMAP further compresses distances in low-dimensional space, often amplifying internal similarity within each cluster. HDBSCAN then identifies these as single dense clusters instead of breaking them down. Also, The large cluster might

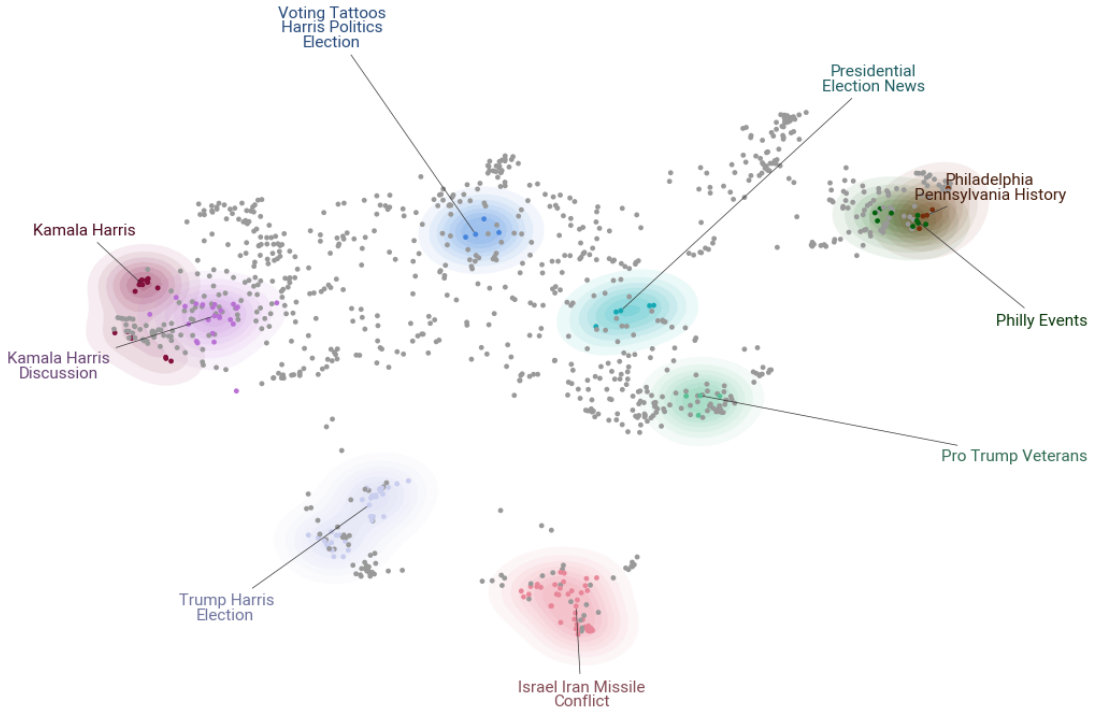

Figure 11: Trending topics related to neutral sentiment.

consist of topics that are semantically similar enough that HDBSCAN struggles to detect boundaries. This can be especially common with text data, where context variations can be minimal. We can adjust the hyperparameters for the submodels to identify natural sub-clusters. For example, Lowering the n_neighbors in UMAP can enhance local relationships in the data, potentially increasing differentiation in the embedding space for HDBSCAN. Also, Lowering min_cluster_size allows HDBSCAN to identify smaller clusters that might represent unique subtopics or niche patterns within the data.

However, Reducing min_cluster_size too much can lead to over-clustering, where HDBSCAN forms multiple small clusters that represent the same or very similar topics. For example, In topics within the neutral sentiments, we have two clusters "Kamala Harris" and "Kamala Harris Discussion" representing the same topic.

Given the relatively small size of our dataset—only 3,000 tweets—splitting it into three subsets likely hindered the model's ability to capture a comprehensive range of topics. Topic modeling, especially with election-related data, benefits from a larger volume of interconnected data points to better distinguish trends and subtopics. By fragmenting the dataset, we may have limited the model's capacity to identify trending topics across the entire collection, potentially resulting in an incomplete representation of the data's thematic diversity. A more robust approach might involve analyzing the entire dataset together to enhance topic detection and capture key trends more effectively. To address the limitations of working with smaller subsets, we reran the topic modeling pipeline on the entire dataset, yielding 36 distinct subtopics. As shown in Figure 12, visualizing the topics across the full dataset provides a clearer understanding of the data's thematic structure and reveals the relationships between different topics. This comprehensive view allows for more accurate identification of trending election-related themes and enhances our insights into the dataset's overall composition.

5 Conclusions

The analysis of social media data related to the 2024 U.S. presidential election provides important insights into how public opinion and voter behavior changed during this crucial time in American politics. We analyzed text data to understand the sentiments of the public and topic modeling to identify key themes and patterns in dataset collected based on popular hashtags.

Our findings show that social media has become a powerful tool for measuring public opinion, giving real-time feedback on how candidates are performing, their campaign strategies, and the issues that

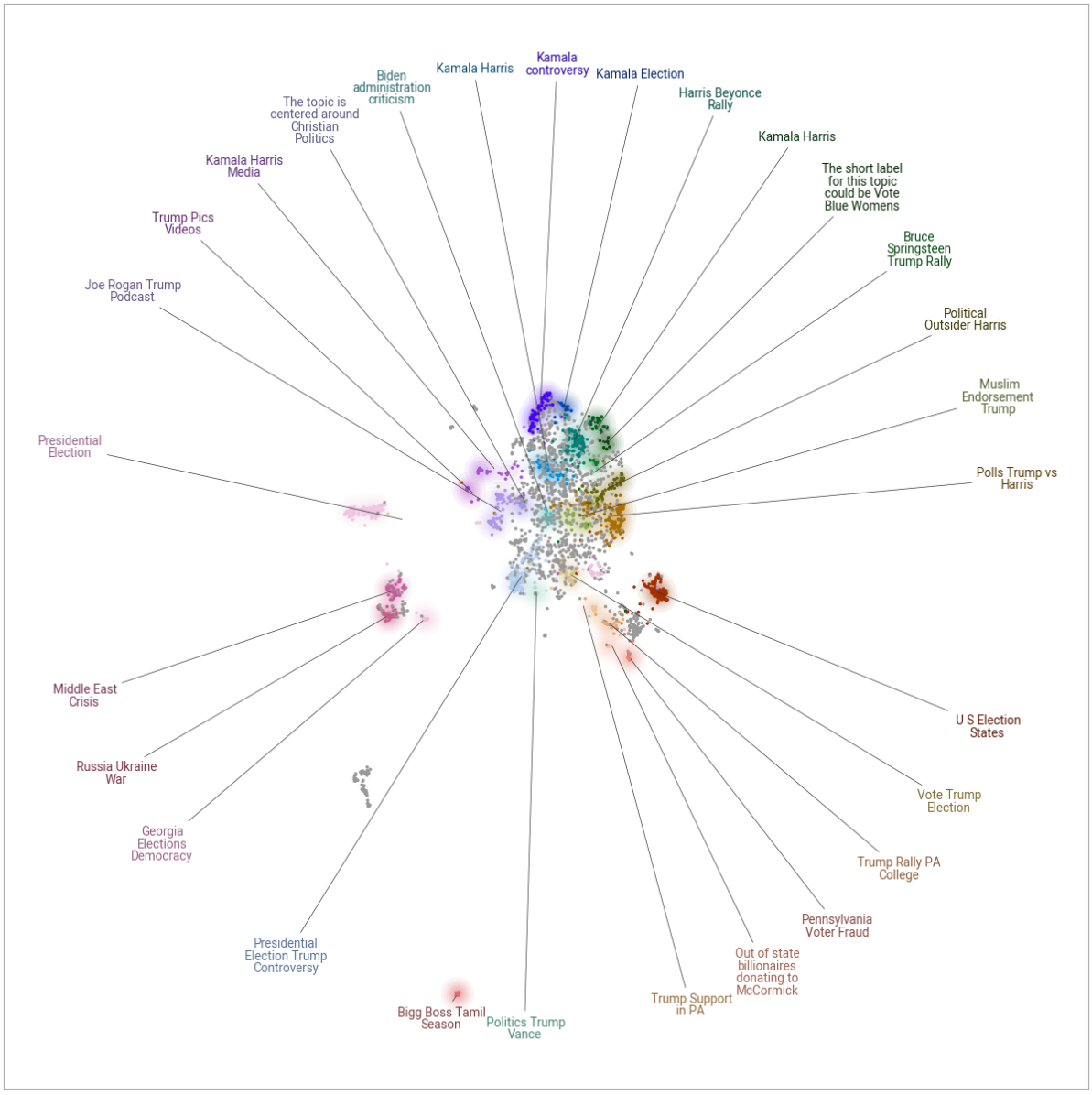

Figure 12: Current topics on Twitter related to the US election 2024.

matter most to voters. Discussions around specific hashtags not only revealed voter priorities but also changed based on significant events, debates, and announcements during the election. For example, when major media events, like candidate debates or important policy proposals, took place, we saw increased engagement on social media, showing how these moments directly influenced voter opinions.

In summary, this project highlights how essential social media is for understanding politics today and its potential to help shape campaign strategies and better understand voter behavior. Given the fast-changing nature of social media, it is crucial for researchers, political strategists, and policymakers to keep studying how social media affects political discussions and election outcomes. This study not only adds to our knowledge of digital political engagement but also paves the way for future research on how politics is evolving in the age of social media.

6 Limitations

The primary limitation of this analysis was the insufficient volume of data. Although our goal was to gather at least one week's worth of tweets, this was restricted due to challenges in web scraping and rate limiting. These constraints impacted the depth of our insights and the robustness of our findings. Given

more time and resources for the project, we could enhance data collection methods to obtain a more comprehensive dataset, allowing for a more detailed and accurate analysis of trends and sentiment over a sustained period.

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 [9]. LLMs also struggle with specificity and contextual accuracy. They often generalize topics broadly, potentially overlooking the nuanced or specific issues in election discoursee, thus, it is essential to evaluate and refine the use of LLMs to enhance their effectiveness and reliability in political discourse analysis.

7 Future Work

For future research, we aim to integrate additional data sources like Reddit, providing a broader spectrum of online public discourse. We also plan to continually gather more Twitter data, working within platform rate limits, to capture comprehensive and evolving election-related insights. Given the real-time nature of our data, future research could focus on developing a live dashboard or implementing ongoing trend analysis to track shifts in public sentiment and topic engagement as they happen. With more data, this approach could also support predictive modeling, allowing us to forecast potential voter behavior and sentiment trends based on current and emerging patterns. Such real-time tracking could be especially valuable for political strategists, offering actionable insights into public response to campaign events, announcements, and policy discussions, enabling a more dynamic and responsive understanding of election-related social media activity.

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

[6] 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).

[7] 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].

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

[9] Qu, Y. & Wang, J. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications* **11**, 1–13 (2024).