Analyzing the Relevance of Reddit Political Threads in the United States 2024 General Elections.

Anushka Pandey, Confidence Oguebu

DATA 297, Special Topics: Natural Language Processing

12/18/2024

Abstract

In today's digital age, online communities like Reddit have become the town hall of the internet, featuring intensive discussions especially related to politics, engaging various users who contribute through upvotes and comments. Recently, the political landscape of the United States has experienced significant changes, likely influenced by the widespread adoption of social media. In light of this, our project explores the dynamic world of Reddit, with a focus on threads about the U.S. elections. With thousands of posts spanning various subreddits, this dataset represents a unique collection of political opinions, voter sentiments, and emerging trends. We aim to uncover hidden patterns by analyzing sentiment shifts, identifying key topics, and exploring the impact of online discussions on real-world political outcomes.

To investigate this pattern, we scraped posts from US election-focused subreddits from 2021 to 2024. We attempt to find their political inclinations through the content of each reddit post. We use BERTopic and LDA topic models to observe the shift in political discussions over time. We also use AWS comprehend sentiment analysis API to identify sentiments associated with each leader, and finally we try to predict sentiment based on reddit self-text and mentioned leaders using Naive Bayes, Logistic regression, Random Forest and Bidirectional LSTM and MLP dense layer models. Our findings reveal a notable surge in leader mentions, particularly for Kamala and Trump, in the months leading up to the election. Additionally, our top-performing algorithms demonstrated higher accuracy when applied to political conversations compared to non-political comments. This highlights the importance of filtering our data to focus on subreddits and mentions within the relevant political discourse.

Introduction

Recent studies have shown some correlation between political inclination and psychological factors such as mode of communication, life ideologies, and information processing. Sylwester K et al. (2015) in their research argued that liberals and conservatives in the United States exhibit notable differences in their attitudes toward non-political activities, including art and leisure. These psychological differences are reflected in the language used on social media by Republicans and Democrats and the differences appear to be distinct enough to suggest that political affiliation to some large extent has an influence on an individual's social media communication style. In this research, we focus on Reddit-style discussions of political views. An interesting objective here is to analyze the shifting trends in sentiments associated with prominent leaders relevant to the United States 2024 general elections.

Previous studies have analyzed the language used by 'red' and 'blue' supporting individuals on twitter, under the assumption that most followers of the official Twitter accounts of the Republican and Democratic parties would have conservative and liberal views, respectively (Guimaraes, A et al. 2019). A closely related study examines the traits and anomalies of political discussions on Reddit, revealing the richness and complexity of online forum conversations in both context and conversational dynamics. Building on this, our project hypothesizes that shifts in public opinion

about key political leaders significantly shape the overall sentiment toward those leaders, potentially influencing election outcomes on a broader scale. We implement and fine-tune a range of machine learning and deep learning models, including transformer-based architectures like BERT, to perform a comparative analysis for sentiment classification. Additionally, we utilized pretrained BERTopic and Latent Dirichlet Allocation (LDA) techniques for topic modeling.

Literature Review

Hofmann et al. (2021) present a comprehensive dataset comprising over 600 political subreddits spanning 12 years, designed to study political discourse on Reddit. This study integrates textual and network analysis to examine polarization, ideological leanings, and engagement trends. The authors employ natural language processing (NLP) techniques and network graph analysis to capture large-scale patterns in how political communities on Reddit interact and evolve. While this research provides a foundational understanding of macro-level dynamics in political discourse, it does not address leader-specific narratives or employ advanced topic modeling approaches for granular analysis. While Hofmann et al. offer valuable insights into subreddit interactions, their analysis lacks focus on specific political leaders and does not employ advanced topic modeling techniques such as BERTopic to uncover nuanced temporal trends. This study explores the evolution of discourse related to individual political figures, a key focus of this research.

Papakyriakopoulos et al. (2020) investigate the influence of Reddit's upvote and downvote mechanisms on political discourse. Analyzing 155 million comments across 55 political subreddits, the authors classify discussions into deliberative, civic, and demagogic categories, emphasizing how voting mechanisms amplify polarization and ideological bias. This study highlights the role of engagement metrics in shaping online discussions and provides a robust framework for examining the tone and quality of discourse. However, it does not address temporal changes or the emergence of specific leader-centric narratives. While the study offers critical insights into engagement dynamics, it does not explore temporal shifts in political narratives or leader-specific discourse. Furthermore, it does not incorporate advanced NLP methodologies, such as contextual embeddings or topic modeling, to analyze evolving themes.

Dataset Description

The dataset consists of posts sourced from Reddit, a widely used social media platform for user-generated content and discussions, spanning the years 2021 to 2024. These posts were collected using the Reddit API through the Python Reddit API Wrapper (PRAW). A keyword-based filtering approach was employed to ensure relevance, targeting discussions centered on specific themes and political events. The dataset emphasizes data from selected subreddits, such as r/politics and r/news, recognized for their high levels of political engagement. It includes metadata such as title (summarizing the post), selftext (the body of the post containing user-shared information), upvotes (indicating post popularity), num_comments (reflecting engagement levels), flair (categorizing the post), created_utc (timestamp in UTC), permalink (unique URL), and subreddit (community of interest). Additionally, hashtags and matched_keywords highlight topics or themes associated with the posts, while normalized scores (scaled_upvotes and scaled_num_comments) facilitate comparative analysis. Temporal features such as day_of_week and time_of_day were derived from timestamps, and word_count provides insights into content length. To ensure robust text analysis, a processed_selftext field was generated by applying preprocessing techniques like tokenization, lemmatization, and stop-word removal. This dataset is structured to enable detailed temporal, thematic, and engagement

analyses, providing a rich source for understanding public discourse during the specified timeline.

Data Pre-processing

To ensure the dataset was suitable for analysis, a thorough data cleaning process was undertaken to enhance accuracy, reduce noise, and maintain consistency. The following steps outline the cleaning and preprocessing methods applied:

1. Handling Missing Data: Posts with missing values in critical fields such as title or selftext were reviewed. Entries with both fields missing were removed, as they lacked meaningful content. Posts with missing auxiliary fields such as flair or hashtags were retained, as they were not critical to the analysis.
2. Removing Duplicate Entries: Duplicate posts were identified and removed using the permalink field, a unique identifier for each Reddit post. This ensured that no post was analyzed more than once, reducing redundancy in the dataset.
3. Filtering Irrelevant Posts: Posts were filtered based on keyword matching during extraction to ensure relevance to the research focus. Any post that did not match the predefined keywords was excluded from the dataset.
4. Standardizing Timestamps: The created_utc field, which contains the timestamp of post creation, was converted from UNIX format to a standardized datetime format. From this, additional temporal features were derived, including day_of_week and time_of_day, which categorized posts by the day and period of posting (e.g., morning, afternoon).
5. Text Preprocessing: The textual data in the selftext field was preprocessed to prepare it for topic modeling and sentiment analysis. First, tokenization was performed to split the text into individual tokens, such as words or phrases. Next, all text was converted to lowercase to ensure uniformity. Stop-word removal was then applied, using the NLTK stop-word list to eliminate common words like "the" and "and" along with custom stop words such as "dont," "said," and "nbsp" to further refine the text. Punctuation and special characters were removed to clean up the data, followed by lemmatization using the WordNet Lemmatizer to reduce words to their base or root forms, such as converting "running" to "run." The processed text was stored in a new column, *processed_selftext*, which served as the foundation for downstream analyses.

Exploratory Data Analysis (EDA)

Word clouds were generated to visualize frequently mentioned words, revealing key topics such as "vote," "people," "election," and "trump," which dominated discussions and highlighted a focus on civic engagement and political figures. The prominence of "kamala harris" in the word cloud indicated her significant presence and relevance in the analyzed text. Additionally, n-grams analysis of unigrams and bigrams provided further context by uncovering common phrases. Words like "trump," "state," "vote," and "election" were among the most frequent, emphasizing the dataset's political and electoral themes, while the presence of "democrat" and "republican" reflected bipartisan discussions. Descriptive statistics on engagement metrics, such as upvotes and num_comments, were also analyzed to identify posts with the highest community interaction. A positive correlation between upvotes and comments suggested that highly upvoted posts tended to generate more engagement, while outliers with high comment counts but fewer

upvotes pointed to controversial or debate-driven topics.

Methodology

This study employs a systematic methodology to analyze Reddit posts discussing political topics, focusing on topic modeling, temporal analysis, and sentiment evaluation. The overall goal is to identify prominent themes and track shifts in narratives surrounding political leaders over time. The methodology is divided into distinct stages, as outlined below:

Topic Modelling

Topic modeling is basically an unsupervised machine learning strategy which has the ability of checking a set of documents, identifying words and uncovering the patterns within them and consequently cluster word bunches and comparative expressions that best characterize a set of corpora. Some of the most popular techniques are the Latent Dirichlet Allocation (LDA) and the BERTopic. In this research, we will be employing these two techniques for our topic modelling task and compare their performance across different leaders.

BERTopic

The study utilized BERTopic, a state-of-the-art topic modeling algorithm, to extract latent themes from the Reddit dataset. BERTopic leverages contextual embeddings, enabling it to capture semantic nuances in textual data.Topic modeling was applied to the entire dataset to identify overarching themes in political discourse. Key steps included:

- Preprocessing: The processed_selftext field was used as the input, which had been cleaned during the data preparation stage by removing stop words, punctuation, and irrelevant terms.
- Embedding Model: The all-MiniLM-L12-v2 model from SentenceTransformers was employed to generate sentence embeddings, providing semantic-rich representations of text.
- BERTopic Parameters: The algorithm was initialized with parameters such as min_topic_size=5 to ensure sufficient data in each topic and nr_topics=10 to achieve a manageable level of granularity.

To analyze the prominence of specific political leaders in Reddit discussions, the second phase of the analysis focused on tracking mentions of selected leaders over time. The analysis utilized named entity recognition and temporal mapping to highlight shifts in discourse surrounding prominent figures.

- Named Entity Recognition (NER): Extracted names of political leaders using the spaCy NER model (en_core_web_sm), focusing on entities labeled as "PERSON" within the processed_selftext field.Filtered posts containing relevant leaders, ensuring that only leader-specific discussions were analyzed.
- Temporal Mapping: Grouped posts by month using the created_utc timestamp to create time_bin for chronological tracking.Quantified leader mentions per time bin for trend visualization.
- Leader Filtering: Defined a list of relevant leaders: Donald Trump, Kamala Harris, Joe Biden, Elizabeth Warren, and Bush.Normalized names to lowercase for consistency and filtered the data accordingly.

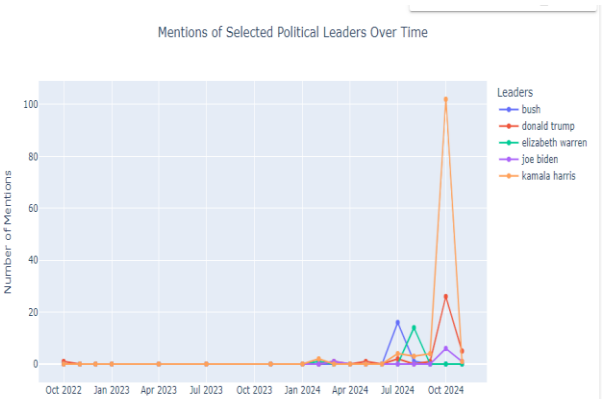


Fig 1: BERTopic Modelling on Leader Mentions

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic model that captures the implicit topic structure from a collection of documents. It is a three-level hierarchical Bayesian model consisting of word, topic, and document layers. To achieve more targeted topics, we refine our dataset to include a relevant corpus containing mentions of the leaders of interest. We then apply the LDA model to this filtered corpus. Following the topic modeling process, we generate word clouds to highlight frequently occurring words, which are subsequently used for labeling the topics. At this stage, we will further purify the results using the intersection of the leader mentions from our corpus and the results from our model. For this project, we experimented with scikit-learn's Latent Dirichlet Allocation model and genism's LDA pretrained model and we opted for the gensim pretrained model because of its topic quality, hyperparameter customization, and topic interpretability. We employed Gensim's pre-trained LDA model to identify the dominant topics within our data on a monthly basis. To prepare the text for analysis, we converted it into both Count vectors and TF-IDF vectors. Since the topics derived from TF-IDF vectors more effectively capture the central themes of the corpus, we chose to use TF-IDF vectors for our analysis. We choose num_topics=5 and we label topics based on the most frequently occurring words in the topic which appeared to be "Maga", "Biden", "Harris", "Trump", and "Trump" again as the dominant word in the final topic. Our findings indicate that, similar to the results from BERTopic analysis, the leader Kamala experienced a notable surge in mentions between the months of August and October, with a more significant spike in October as can be seen in Fig 2 below:

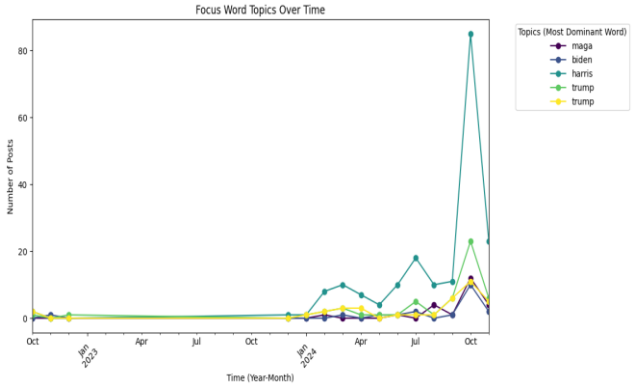


Fig 2: LDA topic Modelling on Leader mentions

The top 5 frequent words for each topic and their frequencies:
0: 0.518"maga" + 0.373*"joe" + 0.069*"biden" + 0.032*"trump" + 0.005*"kamala"*
1: 0.532"biden" + 0.287*"trump" + 0.162*"harris" + 0.016*"kamala" + 0.002*"joe"*
2: 0.535"harris" + 0.366*"kamala" + 0.073*"trump" + 0.019*"joe" + 0.005*"biden"*
3: 0.297"trump" + 0.144*"harris" + 0.144*"biden" + 0.139*"kamala" + 0.138*"maga"*
4: 0.965"trump" + 0.029*"biden" + 0.002*"kamala" + 0.002*"joe" + 0.001*"harris"*

Sentiment Trend Analysis and Prediction

In our research, we experimented with various pretrained sentiment analysis models to automatically assign sentiments to mentioned leaders. Initially, we tested the DistilBERT base uncased model, but its sentiment tagging proved ineffective, often assigning arbitrary positive or negative sentiments based solely on isolated words in the text. This approach was irrelevant to our analysis. To address this, we refined the model by prompting it to focus specifically on the leaders of

interest, isolating their mentions, and assigning positive, negative, or neutral sentiments to the surrounding text. However, the model failed to distinguish sentiment variations and instead assigned neutral sentiments to all texts and leader mentions across the corpus. We then tested the BERT Base Uncased model, which primarily assigned negative sentiments to mentions of leaders. Next, we used the AWS Comprehend custom API, and the results were more varied and insightful, covering sentiment labels such as "negative," "positive," "neutral," and "mixed." Figures 4 and 5 below display the sentiment trend analysis for Trump and Harris over time. Given the complexity of the dataset and the challenge of accurately capturing sentiments related to leaders, most of the sentiments were categorized as neutral. This is reflected in the noticeable spike in neutral sentiments in the graphs below.

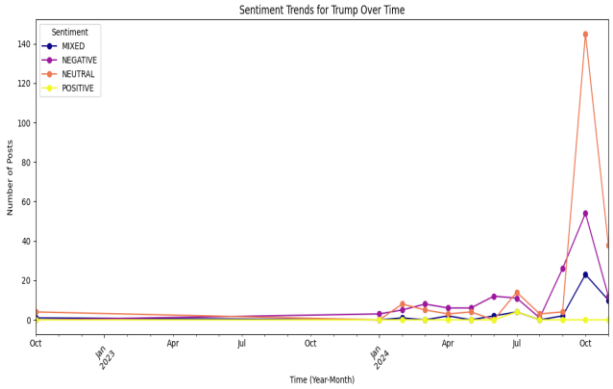


Fig 3: Sentiment trend analysis for Trump over time

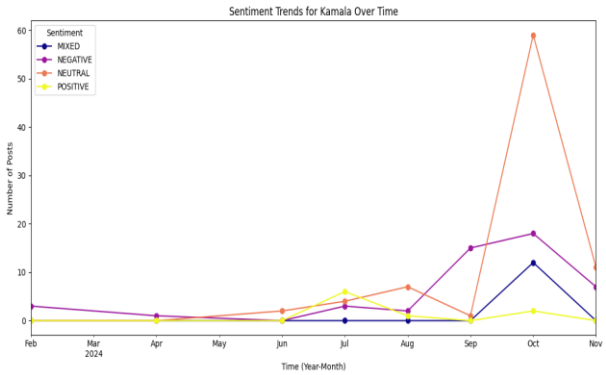


Fig 4: Sentiment trend analysis for Harris overtime

Leader Sentiment Prediction

We applied multiple machine learning models to predict sentiment labels associated with leaders mentioned in the text. Specifically, we compare the performance of these models in predicting sentiments within our dataset. Before feeding the data into the models, we first label-encode both the leader mentions and the sentiment labels. Next, we transform the processed text into vectors using Count Vectorizer and TF-IDF Vectorizer, setting max_features=5000 for both methods. Next, we combine text and extracted leaders into one input vector before feeding it to our model. We observed that the count vectors performed better in predictions across the three different models we explored- Naive Bayes, Logistic regression and random forest. Hence, we will proceed with comparing the results using TF-IDF vectors for training across these three models. We also utilized pre-trained Word2Vec embeddings from Google News (300-dimensional vectors) to create word embeddings for training. However, models trained with Count Vectors outperformed those trained with these word embeddings in terms of prediction accuracy. We split our dataset into 80% for training and 20% for testing.

Experiment and Results

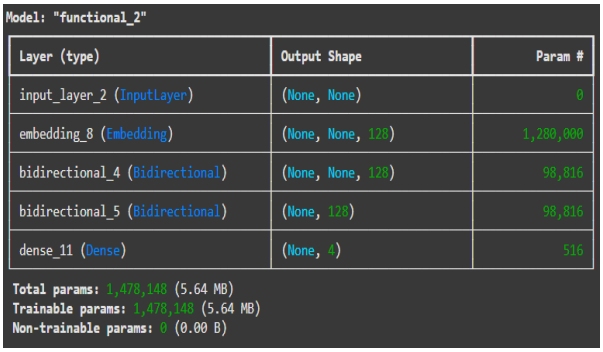
Random Forest: For this project, we utilize the pre-trained sklearn's Random Forest classifier. We set the hyperparameters n_estimators= 100, max_leaf_nodes= 5, max_depth= 10, random_state=42. We train on 80% of our dataset and test on the remaining 20%. The random forest model achieved a 90% performance accuracy on the test set.

Logistic Regression: Next, we utilized the pre-trained sklearn's Logistic Regression classifier. We set the max_iter=1000 and used 80% of our data for training. The logistic regression trained on the count vector inputs achieved 86% accuracy on the test set predictions.

Naive Bayes: We also used sklearn's pre-trained Multinomial Naïve Bayes model for training. However, the Naïve Bayes model had the lowest performance, achieving an accuracy of 75% on test set predictions when using Count Vectors with max_features=5000.

To identify the best-performing models, we also experimented with deep learning approaches, including Bidirectional LSTMs and Multi-Layer Perceptrons (MLPs).

Bidirectional LSTM: We implemented the Bidirectional LSTM models using two bidirectional LSTM layers and one output dense layer with a 'Softmax' activation (see Fig 5 for model summary). This model was trained on our dataset with a 32-batch size and 5 epochs and after evaluation, it recorded an accuracy of 83.5 % and a loss of 0.511 on the test set.



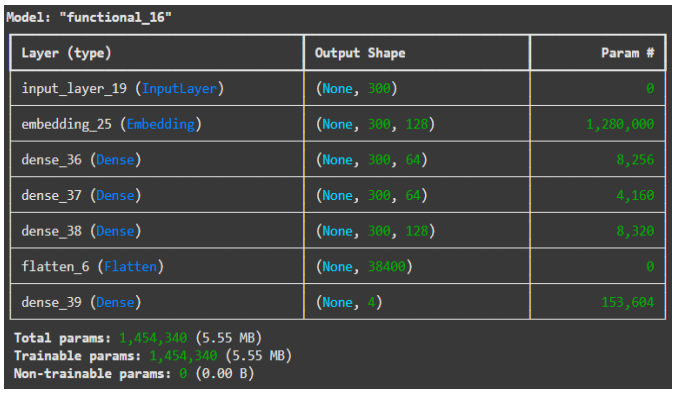
Model: "functional_2"

Layer (type)	Output Shape	Param #
input_layer_2 (InputLayer)	(None, None)	0
embedding_8 (Embedding)	(None, None, 128)	1,280,000
bidirectional_4 (Bidirectional)	(None, None, 128)	98,816
bidirectional_5 (Bidirectional)	(None, 128)	98,816
dense_11 (Dense)	(None, 4)	516

Total params: 1,478,148 (5.64 MB)
Trainable params: 1,478,148 (5.64 MB)
Non-trainable params: 0 (0.00 B)

Fig 5: Bidirectional LSTM model summary

MLP Dense Layers: Using three dense layers of 64, 64 and 128 units with a RELU activation, and an output dense layer with a 'Softmax" activation (see Fig 6 for model summary), we trained the MLP model on the training set using a 32-batch size and 5 epochs. The model achieved 81.6% accuracy and a loss of 0.416 when evaluated on the test set.



Model: "functional_16"

Layer (type)	Output Shape	Param #
input_layer_19 (InputLayer)	(None, 300)	0
embedding_25 (Embedding)	(None, 300, 128)	1,280,000
dense_36 (Dense)	(None, 300, 64)	8,256
dense_37 (Dense)	(None, 300, 64)	4,160
dense_38 (Dense)	(None, 300, 128)	8,320
flatten_6 (Flatten)	(None, 38400)	0
dense_39 (Dense)	(None, 4)	153,604

Total params: 1,454,340 (5.55 MB)
Trainable params: 1,454,340 (5.55 MB)
Non-trainable params: 0 (0.00 B)

Fig 6: MLP dense layers model summary

The table below shows the summary of models and their accuracies when evaluated on the test set:

Model name	Train Accuracy	Test Accuracy	Test Loss
Naïve Bayes	—	75%	—
Logistic Regression	—	86%	—
Random Forest	—	90%	—
Bidirectional LSTM	90.5%	83.5%	0.511
MLP	94.9%	81.6%	0.416

Table 1: Model performance comparison table

Conclusion

In conclusion, discussions on political topics revealed several key themes. Core conversations centered around elections, leadership, and governance, with keywords like "trump," "republican," "state," "vote," and "election" highlighting the focus on electoral processes and political decision-making. Leadership discussions gravitated toward major figures such as Kamala Harris, Joe Biden, and Donald Trump, emphasizing their campaigns and influence. Civic engagement emerged as a significant theme, underscoring voter responsibility and legislative processes through terms like "leadership," "senator," "ballot," and "state." Representation and inclusivity were also prominent, with mentions of "vance," "latino," "luchador," and "black" reflecting identity politics.

The emotional climate of the discussions was marked by public anxiety and anticipation, evidenced by keywords like "nervous," "election," "vote," and "win."

Leader-specific trends revealed significant spikes in mentions of Kamala Harris and Donald Trump during late 2024, coinciding with the election period and reflecting heightened public interest in their campaigns. Temporal dynamics showed that leader mentions remained low and stable throughout 2022–2023, with notable increases emerging in 2024, aligning with the electoral campaign timeline. A comparative analysis for both the BERTopic and LDA models highlighted Kamala Harris as surpassing other leaders in mentions, indicating strong public focus on her campaign or related events. Meanwhile, leaders like Joe Biden and Elizabeth Warren maintained relatively steady engagement with occasional minor spikes, suggesting ongoing relevance but less dramatic shifts in attention.

Despite the challenges of sentiment analysis due to the complexity of the dataset, we effectively identified leaders and associated sentiments using the AWS Comprehend API. Its ability to capture nuances and perform context-aware analysis proved invaluable. Among the machine learning models, the Random Forest model significantly outperformed others in predicting sentiments associated with each leader in the corpus, even though the 'Neutral' sentiment was the most prevalent in our dataset. Overall, while this analysis offers valuable insights into political discussions and the temporal sentiment patterns associated with leader mentions, it does not provide sufficient evidence to reliably predict the broader outcomes of the 2024 general elections.

Limitations

Lack of access to post comments: The inability to access post comments posed a significant challenge to our analysis. Posts often contained lengthy text covering various topics, whereas comments would have been shorter, more direct, and better reflected individuals' actual perspectives on political issues and preferences.

Automatic sentiment tagging: Given the limited timeframe for this project, we were unable to manually annotate the sentiment labels. As a result, the sentiment labels associated with leader mentions in our analysis may not accurately reflect the true sentiments expressed in the text.

Insufficient data: Due to computational constraints and the nature of our dataset, we had to reduce its size and filter it to include only posts mentioning the leaders of interest. Access to more data would likely have provided broader insights, enabling us to draw stronger and more generalized conclusions from our analysis.

References

1. Guimaraes, A., Balalau, O., Terolli, E., & Weikum, G. (2019). Analyzing the Traits and Anomalies of Political Discussions on Reddit. Proceedings of the International AAAI Conference on Web and Social Media, 13(01), 205-213. https://doi.org/10.1609/icwsm.v13i01.3222
2. Parsons, J., Schrider, M., Ogunlela, O., and Ghanavati, S., “Understanding Developers Privacy Concerns Through Reddit Thread Analysis”, arXiv e-prints, Art. no. arXiv:2304.07650, 2023. doi:10.48550/arXiv.2304.07650.
3. Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of research trends using LDA based topic modeling. Global Transitions Proceedings, 3(1), 298-304.

4. Sylwester K, Purver M. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. PLoS One. 2015 Sep 16;10(9):e0137422. doi: 10.1371/journal.pone.0137422. PMID: 26375581; PMCID: PMC4574198.

5. Papakyriakopoulos, Orestis, Serrano, Juan Carlos Medina and Hegelich, Simon. "Political communication on social media: A tale of hyperactive users and bias in recommender systems." Online Social Networks and Media 15 (2020): 100058.

6. Hofmann, V., Schütze, H., & Pierrehumbert, J. B. (2022). The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 1259-1267.

7. Al-garadi, M. A.; Varathan, K. D.; Ravana, S. D.; Ahmed,E.; Shaikh, G. M.; Khan, M. U. S.; and Khan, S. U. 2018.Analysis of online social network connections for identifi-cation of influential users: Survey and open research issues.ACM Comput. Surv.51(1):16:1–16:37

Appendix

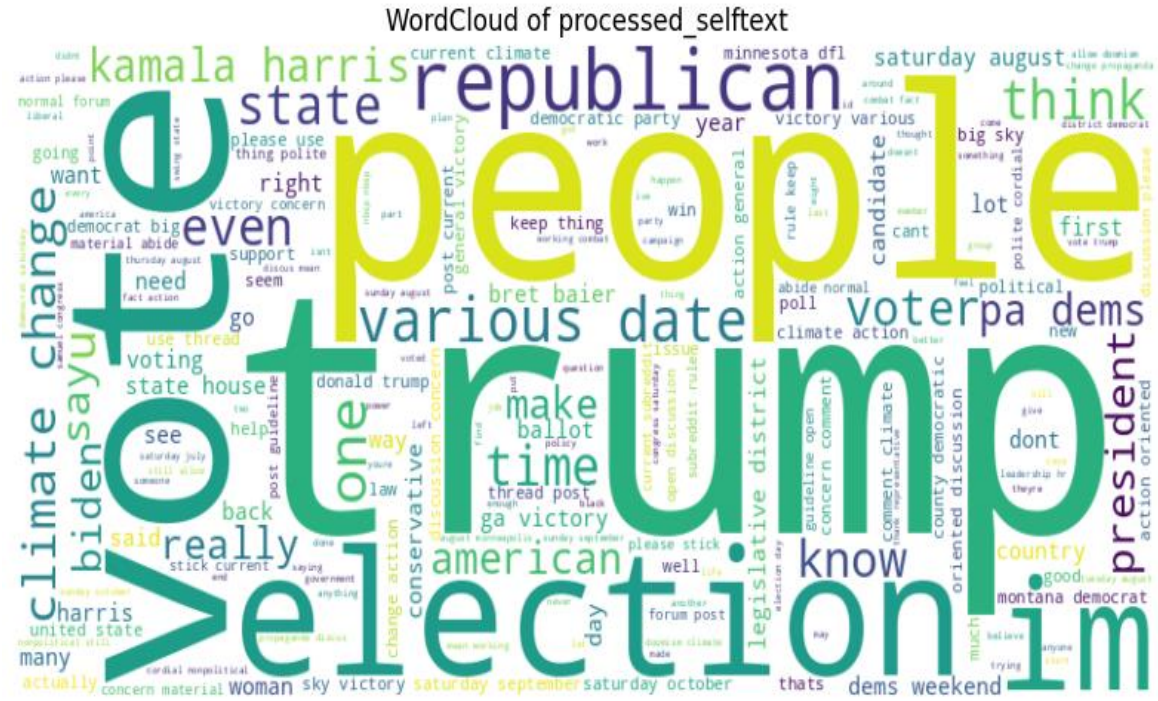


Fig 1: Word Cloud of the processed post text, showing most frequently occurring words

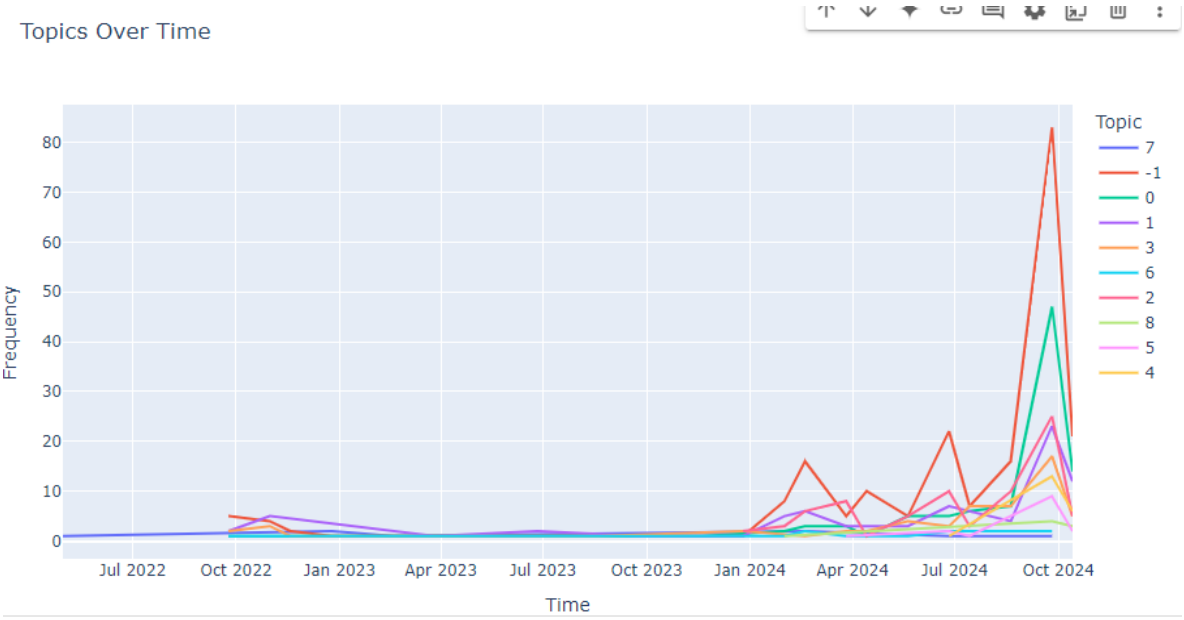


Fig 2: BERTopic model showing shift in topic mentions over time.

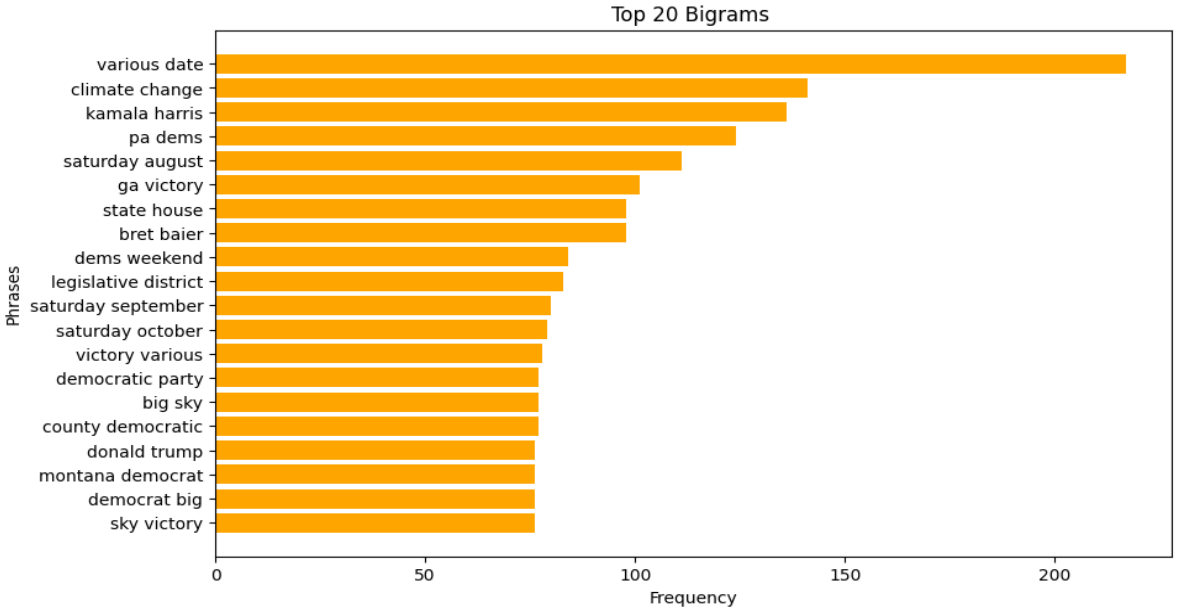


Fig 3: Top 20 Bigrams

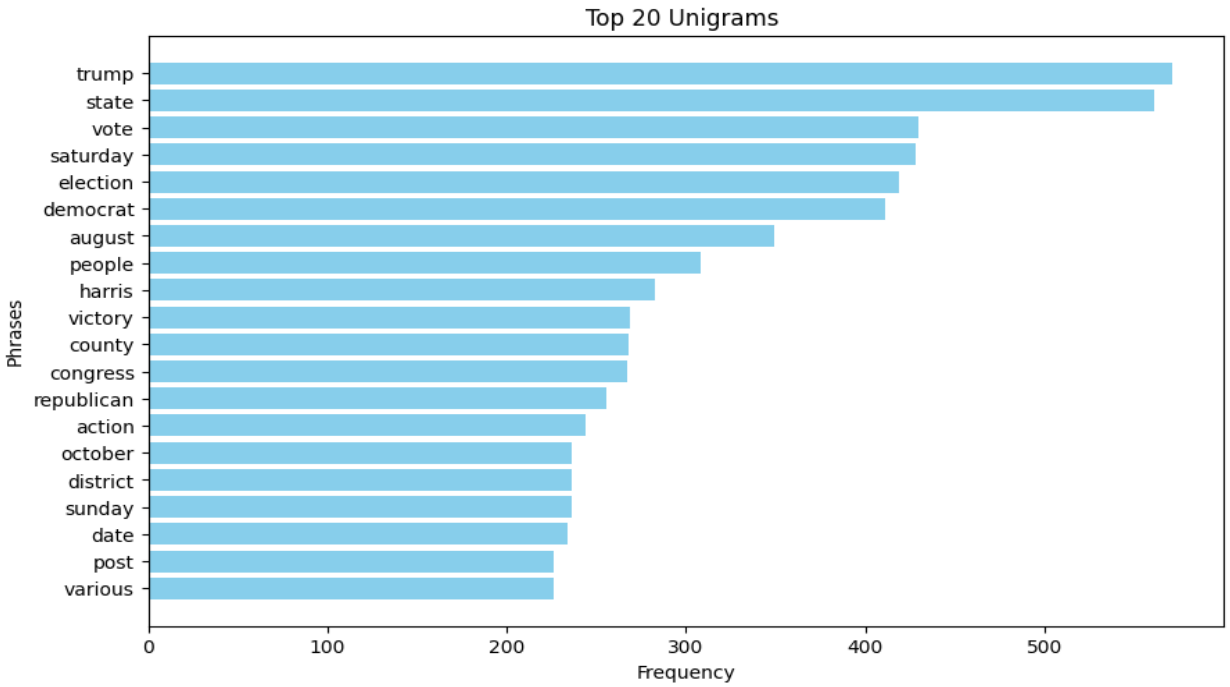


Fig 4: Top 20 unigrams