Understanding Public Perceptions of the US Parties through Reddit Comments

*Liu Zhi Wei
Dept. of Information Management
National Taiwan University

*Chun **
Dept. of Information Management
National Taiwan University

*Shu **
Dept. of Information Management
National Taiwan University

*Chen **
Dept. of Information Management
National Taiwan University

*Lee **
Dept. of Information Management
National Taiwan University

***Abstract*—Our project leverages Reddit data to analyze emotional fluctuations and focus areas of U.S. Democrats and Republicans before and after the 2024 election. By employing sentiment analysis and unsupervised learning techniques, the research provides insights into how political events influenced emotional trends and thematic discussions, offering valuable perspectives for understanding electoral behaviors.**

***Index Terms*—U.S. election, sentiment analysis, unsupervised learning, clustering**

I. Research Background

A. Research Objectives

Social media platforms have grown into powerful channels for political communication and voter mobilization. While Twitter has been widely studied, Reddit offers unique advantages due to its topical subreddits, longer text format, and community-driven moderation. Focusing on the 2024 U.S. election—a period marked by polarized rhetoric and heightened public scrutiny—this study harnesses sentiment analysis and unsupervised learning to examine how Democrats and Republicans expressed emotions and engaged with key campaign issues online. By comparing data-driven sentiment scores with the thematic groupings revealed by clustering methods, the research provides a multi-faceted understanding of the topics and emotional responses that defined each party's online discourse. These insights hold value for political strategists, sociologists, and communication researchers seeking to better comprehend the interplay between digital platforms and electoral behaviors.

B. Experimental Data

- **Data Source:** Comments were sourced from two publicly available Kaggle datasets. These datasets, which are updated daily, comprise Reddit posts and comments relevant to Republicans and Democrats, respectively, based on keyword-matching in designated subreddits. To ensure data quality, we filtered out non-English entries and removed any comments that did not explicitly relate to political discussions. Basic text preprocessing steps (e.g., lowercasing, removing special characters) were applied to standardize the input for sentiment analysis.
- **Timeframe:** From July 1, 2024, to December 1, 2024.
- **Data Division:** Comments were categorized based on party affiliation into two groups: Democrats and Republicans.

II. Methodology

A. Sentiment Analysis

We experimented with both the TextBlob and VADER Sentiment models to perform sentiment analysis and calculate sentiment scores. After reviewing multiple related studies and conducting manual evaluations, we found that VADER provided superior performance. Therefore, VADER was selected as the primary sentiment analysis model in our final approach.

The VADER sentiment model calculates sentiment scores based on a pre-built lexicon that associates words and phrases with their respective sentiment intensity. For each input text, VADER assigns three primary scores:

- Positive score: Indicates positive emotion.
- Negative score: Indicates negative emotion.
- Neutral score: Indicates no significant emotion.

In addition, VADER provides a compound score, which is a normalized metric ranging from -1 (most negative) to +1 (most positive). This score is derived from a summation of the valence scores of individual words in the text, adjusted according to grammatical and syntactical rules.

Visualization: We conducted our analysis in a Jupyter Notebook environment using Python. After computing VADER scores for each comment, we aggregated sentiment results on a daily basis, allowing us to observe trends across the election cycle.

B. Classification Models

To classify Reddit comments based on political affiliation (Democrats or Republicans), we utilized the Natural Language Toolkit (NLTK) for text preprocessing, including stemming and stopword removal. Several machine learning models were employed for this task, including Support Vector Machines (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN).

*Authors are sorted according to their student IDs.

Each model was trained and tested using features extracted from the preprocessed dataset, including word frequencies and sentiment scores, to maximize classification accuracy. Below is a summary of the models and their respective characteristics:

- **Support Vector Machines (SVM):** We utilized both linear and radial basis function (RBF) kernels for SVM models to explore the impact of different kernel types on performance.
- **Naive Bayes (NB):** Naive Bayes is computationally efficient and works well for text classification tasks where the features represent word counts or TF–IDF scores. NB serves as a baseline model for our classification experiments.
- **K-Nearest Neighbors (KNN)**
- **Feature Selection:** For all models, we used TF–IDF vectors to represent text features.

Model Comparison: Each classifier was evaluated using precision, recall, F1-score, and accuracy metrics to ensure a thorough assessment of their performance.

C. Unsupervised Learning

In addition to sentiment analysis, we applied unsupervised learning techniques to identify recurring themes within the Reddit comments and to analyze how these themes differed between Democrats and Republicans. The overall goal was to cluster semantically similar texts together, extract relevant keywords, and discover topic-level patterns that might not be immediately evident from sentiment scores alone.

Topic Modeling Techniques: For topic modeling on the two datasets, KMeans clustering and Latent Dirichlet Allocation (LDA) were used. These methods grouped comments into predefined clusters or topics based on vectorized text features, such as bag-of-words or TF-IDF representations. The clustering process revealed distinct themes within the comments, helping to uncover underlying patterns and key discussion points in both Democratic and Republican conversations.

After performing topic modeling on both datasets, we extracted high-weight terms to identify different topic interests based on political groups. To achieve this, we approached the process in various ways, with the goal of selecting keywords and tracking topic trends for different groups. However, our process evolved over multiple attempts to refine how keywords were selected and grouped:

- **Attempt 1: TF–IDF on Entire Dataset**
  We first computed TF–IDF scores for all comments in the six-month period, intending to pick out the top keywords in each month. However, since the inverse document frequency (IDF) was computed over the *entire* dataset, the resulting keywords tended to be too generic and provided minimal insight into evolving monthly or topic-specific trends.
- **Attempt 2: TF–IDF + Chi-Square by Month**
  To achieve more granular differentiation, we introduced a chi-square feature selection step and re-scoped the IDF calculation to focus on the specific six-month window. Treating each month as a separate class, we identified words that showed significant association with particular months. This method better captured time-sensitive discussions and revealed more contextual keywords tied to real-world events.
- **Attempt 3: Subreddit-Based Discrimination**
  Given that Reddit communities are segregated by shared interests, we segmented data into two overarching categories based on political subreddits.
- **Attempt 4: Topic Grouping and Manual Labeling**
  To address the need for broader thematic categories (e.g., *foreign policy*, *gender issues*), we manually aggregated related keywords under umbrella topics. By mapping individual terms such as *Israel*, *Palestine*, and *Iran* to a *Middle East* topic, we captured higher-level narratives. This final stage enabled us to conduct frequency analysis of each *topic* rather than just isolated words, making it easier to interpret the political emphasis Democrats and Republicans placed on issues like human rights, military aid, or social values.

All clustering and topic modeling were performed using Python libraries including `scikit-learn` (for KMeans and TF–IDF vectorization) and `gensim` (for LDA). Key considerations included:

- **Preprocessing:** As with the sentiment analysis step, comments were cleaned and standardized (e.g., lowercasing, removal of punctuation and stopwords).
- **Hyperparameter Selection:** The number of clusters (K) and topics were tuned experimentally based on interpretability and model coherence.
- **Manual Validation:** Clusters and topics were inspected by domain experts to confirm that groupings made conceptual sense and were not dominated by noise or irrelevant text.

III. Results and Discussions

A. Sentiment Analysis Results

We plotted daily sentiment counts (positive, negative, and neutral) and aligned them with five key political events. **The figure on the top represents Democrats, and the bottom one represents Republicans.** These events were selected based on their potential impact on public opinion and the significant media attention they attracted.

1) **Trump Got Shot**

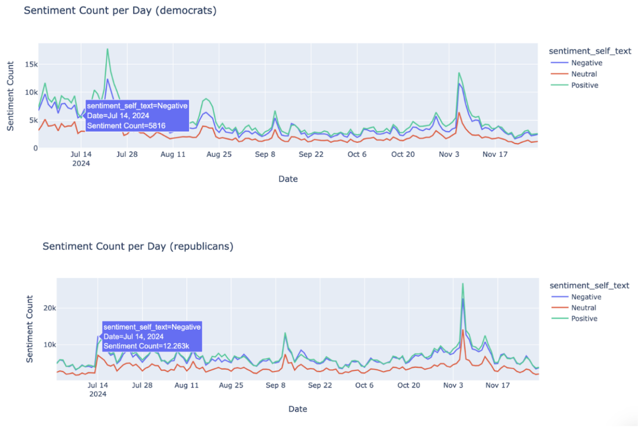

Fig. 1. 1st peak

2) Biden Withdraws from the Election

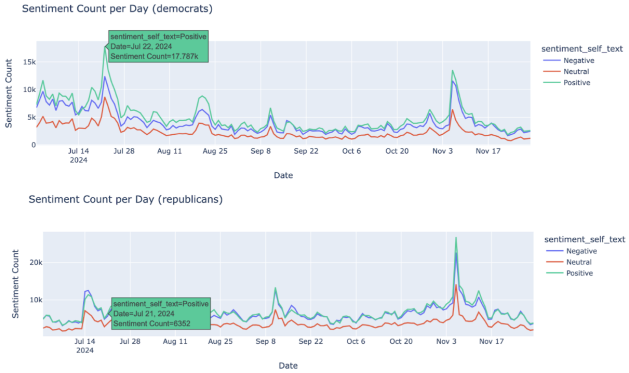

Fig. 2. 2nd peak

3) Kamala Declares Candidacy

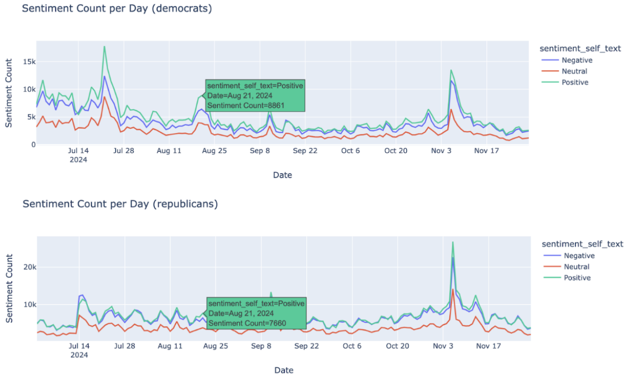

Fig. 3. 3rd peak

4) U.S. Candidate TV Debate

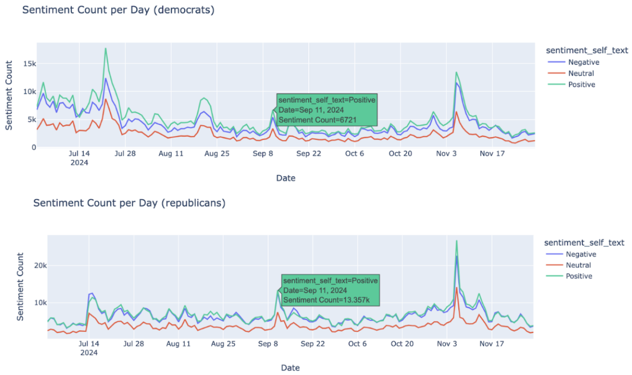

Fig. 4. 4th peak

5) Voting Day

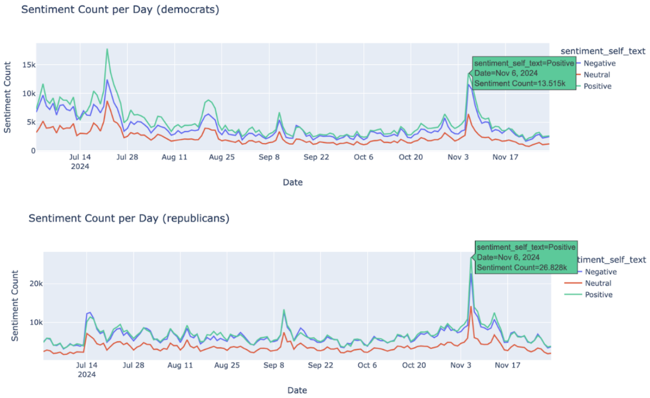

Fig. 5. 5th peak

a) Daily Sentiment Trends at Key Events: For each event date, we calculated the proportion of positive, negative, and neutral comments in the Democratic and Republican datasets separately. Surprisingly, despite the gravity of events like the shooting involving former President Trump and the withdrawal of President Biden, our data did not indicate pronounced spikes in extreme sentiment. Although we observed marginal increases in either positive or negative posts around the event dates, these fluctuations were not as substantial as one might have expected from such high-profile occurrences.

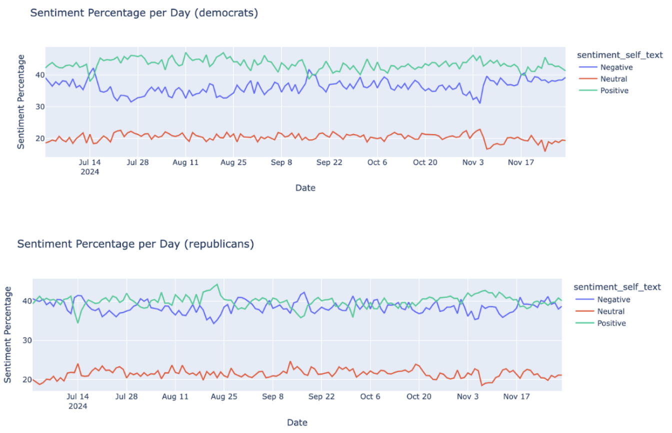

Fig. 6. Overall sentiment percentage

b) Overall Timeline Sentiment Distribution: In addition to examining specific events, we plotted the overall timeline of daily sentiment percentages for both Democrats and Republicans. While certain dates showed modest deviations—particularly following significant press coverage or political announcements—the day-to-day distribution of sentiment scores remained relatively stable. For instance, our initial assumption was that the Trump shooting or Biden's announcement to withdraw from the race would trigger marked emotional shifts, especially within Democratic discussions. Contrary to these expectations, the proportion of positive and negative posts showed only minor variations, suggesting that the Reddit communities we analyzed did not exhibit large-scale swings in emotional valence.

c) Interpretation and Possible Explanations: There are multiple factors that could explain the smaller-than-anticipated changes in sentiment:

- **Platform Dynamics:** Reddit users, compared to broader social media audiences, may engage in more measured discussions, resulting in less dramatic sentiment changes on a day-to-day basis.
- **Data Scope:** Although the events were highly publicized, not all subreddits discussed them with the same intensity. Comments may have been distributed among various threads, diluting overall sentiment spikes.
- **Polarization and Echo Chambers:** If individuals primarily engage in like-minded subreddits, their baseline sentiment toward opposing parties might already be either predominantly negative or positive, leaving little room for sharp shifts.

In summary, while we did observe some localized reactions to major campaign milestones, the broader analysis of daily positive, negative, and neutral comments indicates that sentiment proportions remained relatively constant over the study period. This finding underscores the nuanced nature of online discourse—where even major events do not always translate into large, measurable fluctuations in sentiment at the community level.

B. Classifier Testing

This section presents the performance evaluation of various classifiers on the dataset.

SVM with Linear Kernel

```
              precision    recall   f1-score
democrats       0.54        0.29      0.38
republicans     0.64        0.83      0.72
accuracy                              0.62
macro avg       0.59        0.56      0.55
weighted avg    0.60        0.62      0.59
```

SVM with RBF Kernel

```
              precision    recall   f1-score
democrats       0.58        0.25      0.35
republicans     0.64        0.88      0.74
accuracy                              0.63
macro avg       0.61        0.56      0.54
weighted avg    0.62        0.63      0.58
```

K-Nearest Neighbors (KNN)

```
              precision    recall   f1-score
democrats        0.42       0.37      0.39
republicans      0.60       0.66      0.63
accuracy                              0.54
macro avg        0.51       0.51      0.51
weighted avg     0.53       0.54      0.53
```

Naive Bayes Classifier

```
              precision    recall   f1-score
democrats        0.63       0.13      0.22
republicans      0.62       0.95      0.75
accuracy                              0.62
macro avg        0.62       0.54      0.48
weighted avg     0.62       0.62      0.53
```

General Trends:

- Across all models, Republicans were consistently classified more effectively than Democrats, as evidenced by higher recall and F1-scores.
- The Naive Bayes classifier exhibited the most significant imbalance, with Democrats having extremely low recall.
- SVM with RBF kernel performed the best overall, achieving the highest accuracy and relatively balanced results compared to other models.
- KNN struggled the most with overall performance, as evidenced by its lowest accuracy and F1-scores across both classes.

C. Unsupervised Learning Results

In addition to sentiment analysis, unsupervised learning techniques were employed to uncover recurring themes and linguistic patterns within the comments. These methods provided a deeper understanding of how discussions varied between Democratic and Republican comments on reddit.

Topic Modeling Insights: KMeans clustering and LDA were employed to group comments into semantically similar categories, uncovering meaningful patterns in the data. To determine the optimal number of clusters, the Elbow Method was applied, as shown in the figure below. The graph indicates that $k = 9$ is an appropriate choice, where the inertia begins to level off, balancing interpretability and computational efficiency.

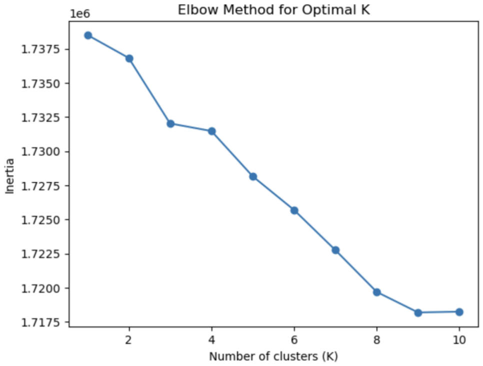

Fig. 7. Optimal k

The Democratic discussions were grouped into nine clusters, reflecting diverse topics:

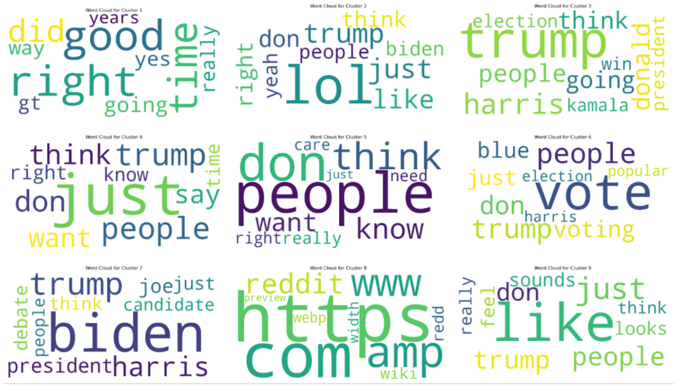

Fig. 8. Democrats' K-Means result

- Cluster 1 focused on general reflections on time and political direction (e.g., “time,” “good,” “right”).
- Cluster 3 highlighted elections, leadership, and political figures such as Trump, Biden, and Harris.
- Cluster 6 revolved around voting and civic engagement, emphasizing terms like “vote,” “Trump,” and “people.”

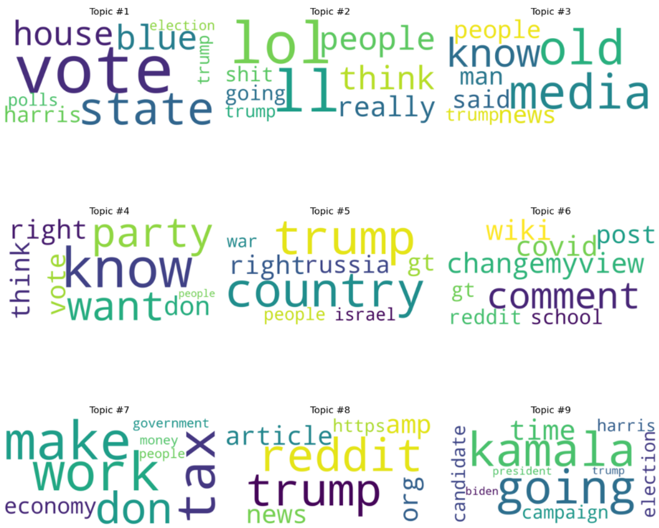

Fig. 9. Democrats' LDA result

- Topic 1: Focused on voting and political parties. Key terms include ”vote,” ”state,” ”house,” and ”blue.”
- Topic 7: Economy and labor. Words like ”make,” ”work,” ”tax,” and ”economy” suggest economic policies and employment concerns.
- Topic 9: Campaigns and candidates. Key words like ”kamala,” ”time,” ”going,” and ”campaign” indicate discussions on political figures and election strategies.

The Republican discussions were also grouped into nine topics:

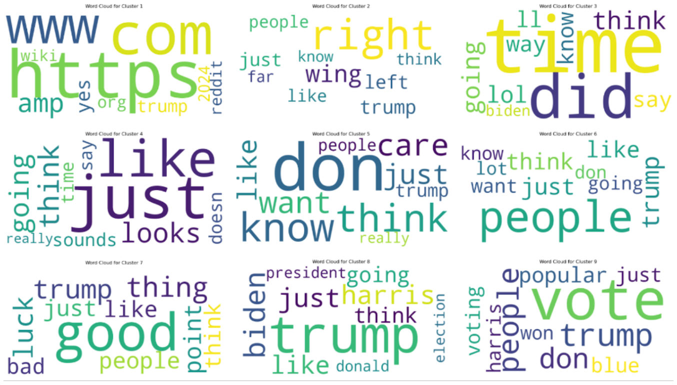

Fig. 10. Republicans' K-Means result

- Cluster 1: Online discussions and references to external links, e.g., “reddit,” “wiki,” “org.”
- Cluster 3: Conversations involving political figures and events, with terms like “Biden,” “Trump,” “going,” and “did.”

- Cluster 9: Themes centered on elections and voting, featuring keywords like “won,” “Harris,” “voting,” and “Trump.”

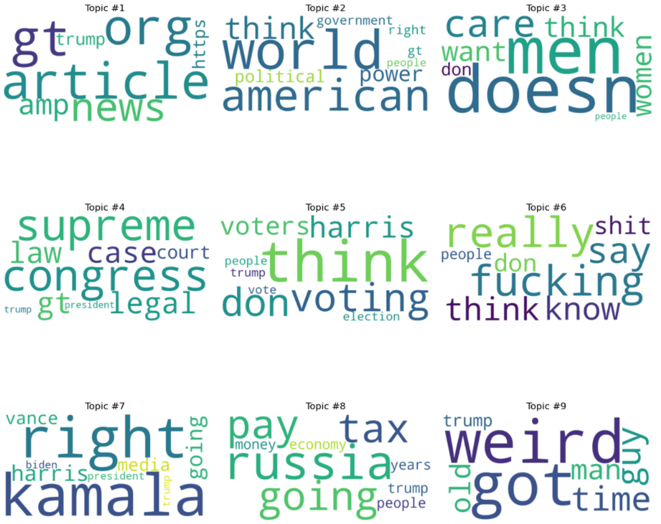

Fig. 11. Republicans’ LDA result

- Topic 2: Focused on global politics and American perspectives, highlighted by terms such as ”world,” ”American,” and ”political.”
- Topic 5: Centered on political engagement and elections, emphasizing voter participation and the electoral process.
- Topic 8: Revolved around economic policies and international relations, with key terms like ”tax,” ”Russia,” and ”going.”

Summary: LDA provides a broad thematic overview, capturing high-level topics and general themes, while K-means offers focused clustering of specific interest areas. Both models highlight shared discussions, such as elections and political figures, but differ in strengths: LDA excels at nuanced topics (e.g., legal and economic issues), whereas K-means effectively identifies meta-discussions and distinct clusters. Together, these methods help uncover key discussions on both Democratic and Republican comments on Reddit.

D. Keywords Extracting

To uncover distinct topic trends among political groups, multiple keyword extraction methods were explored:

1) **TF–IDF on the Entire Dataset:** This method calculated the TF–IDF scores for keywords across the entire dataset.

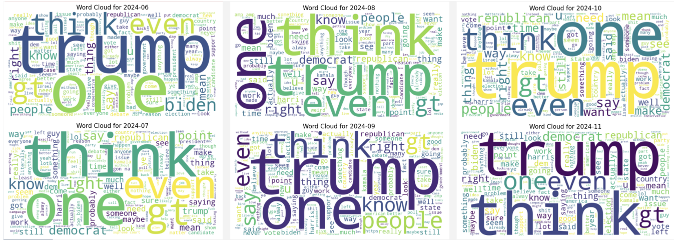

Fig. 12. Naïve attempt on both parties

However, the results lacked specificity, as the dataset’s concentration over six months resulted in generic words with minimal topic differentiation.

2) **TF–IDF + Chi-Square by Selected Months:** By applying chi-square for word selection and then narrowing the TF–IDF scope to six months, greater differentiation emerged. This approach highlighted keywords associated with specific events.
   - **When Biden stepped down in July.** We can see the keywords focused on “Biden,” “Step,” and “Candid.”

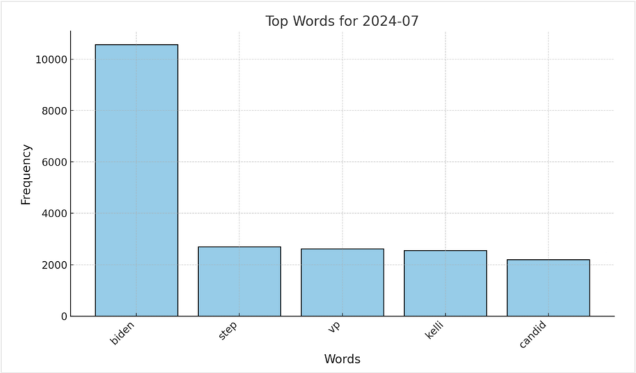

Fig. 13. Keywords for 2024-07

   - **When Tim Walz was nominated as a vice-presidential running mate,** where Josh Shapiro was the other possible vice-presidential running mate. **All of these events occurred** in August.

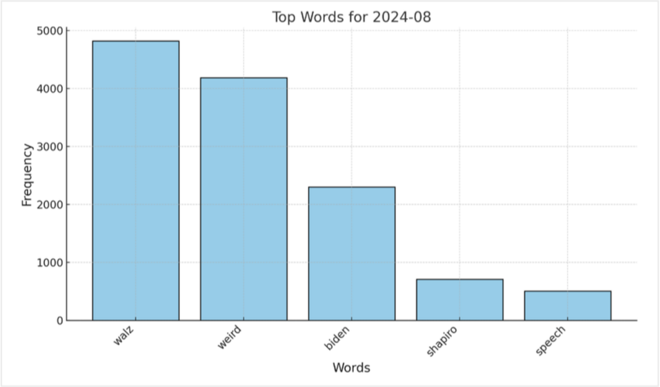

Fig. 14. Keywords for 2024-08

3) **Subreddit-Based Discrimination:** We segmented data into two overarching categories based on political subreddits: Democrats (e.g., `r/democrats`, `r/VoteDEM`) and Republicans (e.g., `r/Conservative`, `r/AskThe_Donald`). We then examined the frequency of previously identified keywords in each partisan subreddit.

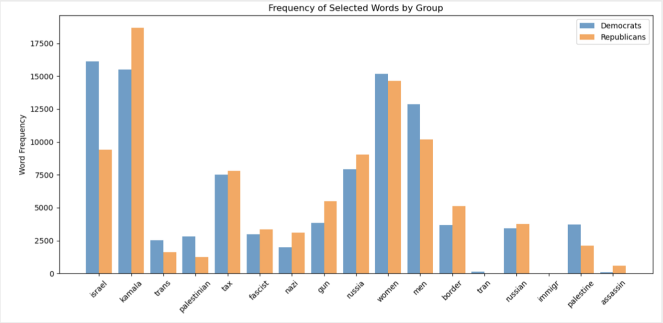

Fig. 15. Subreddit-Based Discrimination

This approach surfaced certain party-specific language patterns but still fell short in providing distinctly cohesive topics (e.g., mentions of international conflicts scattered among multiple keywords, such as "palestine" and "israel," which should be aggregated into one topic).

4) **Topic-Based Grouping:** Manual aggregation of keywords into topics allowed more nuanced insights into party inclinations.

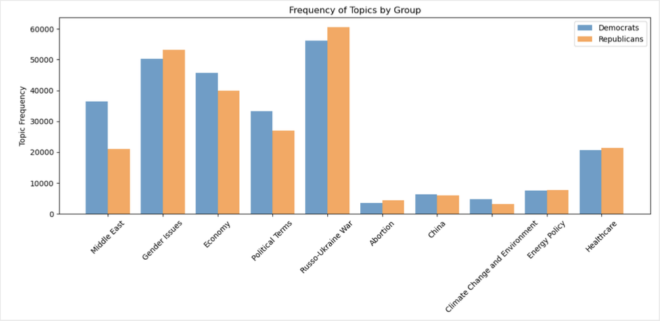

Fig. 16. Grouping by topic

Middle East

- **Democrats:** The higher frequency of discussion suggests a focus on human rights issues, international diplomacy, and the implications of U.S. foreign policy in the Middle East, such as the Israel-Palestine conflict or humanitarian crises in the region.
- **Republicans:** The relatively lower focus might stem from emphasizing U.S. security and counter-terrorism strategies, such as relations with countries like Iran or Afghanistan, but in a more specific, targeted way.

Gender Issues

- **Republicans:** The frequency may indicate a focus on traditional gender roles or resisting progressive narratives, possibly discussing gender issues in the context of culture wars or criticism of "woke" ideologies.

Russo-Ukraine War

- **Republicans:** Their higher frequency may reflect concerns about the financial costs of U.S. involvement, military aid, or critiques of the Biden administration's handling of the war. Isolationist elements within the party could contribute to this interest.
- **Democrats:** Discussions likely center on supporting Ukraine as a stand for democracy and human rights against authoritarianism. There may be criticism of previous administrations' ties with Russia.

IV. Conclusion

A. Analysis and Future Aspects

This study highlights the contributions and significance of using sentiment analysis and keyword clustering to reveal the emotional dynamics and focal points of the two major U.S. political parties during election periods. By employing sentiment analysis and unsupervised learning methods, the research successfully uncovers the differences in emotional changes and key areas of interest between Democrats and Republicans across various events. These findings not only demonstrate the potential of data-driven methods in political studies but also provide critical insights into voter attitudes and public opinion dynamics. Such discoveries are instrumental in gaining a more comprehensive understanding of the drivers of political behavior during elections, offering valuable references for policymakers and academic researchers alike.

The Intersection of Technology and Sociology: This research underscores the importance of integrating technological tools with sociological analysis to better understand election dynamics.

B. Future Directions

1) **Data Diversity and Integration:** Future studies could incorporate more social media platforms (e.g., Twitter or Facebook) and news media to enhance the comprehensiveness and diversity of the analysis.
2) **Cross-Cultural Comparisons:** Researching electoral processes in other countries could explore whether similar methodologies are applicable across different political and cultural contexts.
3) **Refined Segmentation and Audience Analysis:** Further investigations could delve into the latent characteristics of commenters (e.g., gender, age, geographical location). Combining such insights with user segmentation data can facilitate a deeper exploration of how different demographic groups react to events and whether these responses are influenced by local economic, cultural, or media environments. This approach would provide a more holistic understanding compared to binary classification.

REFERENCES

[1] Mohd Zeeshan Ansari, et al., ”Analysis of Political Sentiment Orientations on Twitter,” *Procedia Computer Science*, Volume 167, 2020, Pages 1821-1828.

[2] Kaggle, ”Public Opinion on Republicans,” [Online]. Available: https://www.kaggle.com/datasets/asaniczka/public-opinion-on-republicans-daily-updated/data.

[3] Kaggle, ”Public Opinion on Democrats,” [Online]. Available: https://www.kaggle.com/datasets/asaniczka/public-opinion-on-democrats-updated-daily/data.

[4] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support Vector Method for Novelty Detection. In Proceedings of the 12th International Conference on Neural Information Processing Systems (Denver, CO) (NIPS’99). MIT Press, Cambridge, MA, USA, 582-588.https://dl.acm.org/doi/10.5555/3009657.3009740.