Analysis of Political Discussion on Reddit(r/PoliticalDiscussion): Engagement, Sentiment, and Ideological Leaning in Pre-2024 US Presidential Elections Discourse

Introduction

Social media has transformed how people engage with politics, particularly in the context of public discussions and opinion formation. Platforms like Reddit have grown as digital forums where users exchange ideas, pose questions, and debate on political matters, especially around high-stakes events such as elections. Unlike traditional media, social media allows users to interact with content directly, expressing support, criticism, and polarized opinions in real time. These platforms offer rich sources of unstructured data, including post content, user interactions, and engagement metrics, which can provide insights into public sentiment and ideological trends.

Reddit is particularly significant for studying political discussions due to its community-driven structure, where subreddits dedicated to specific topics facilitate targeted discussions. Subreddits like r/PoliticalDiscussion are forums where users of varying political ideologies discuss policies, candidates, and pressing national issues. Given the role of social media in influencing public opinion, analyzing Reddit discussions can reveal key insights into public sentiment toward candidates and parties, as well as shifts in ideological leanings, especially during critical election periods.

This study aims to analyze Reddit posts from political discussion-focused subreddits over some months preceding the 2024 U.S. presidential election. By exploring metrics like engagement (upvotes and comments), sentiment, and ideological leaning of questions posed in discussions, this research seeks to offer insights into the tone and focus of public discourse during the lead-up to an election

Literature Review

Recent studies have increasingly turned to social media analysis to investigate political discourse, public sentiment, and ideological leanings during elections. Platforms like Twitter and Reddit have been used to gauge public sentiment, with some research indicating that sentiment expressed on social media can align with election outcomes (Chaudhry 2021). However, ("A. Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period." 2023) found that sentiment alone may not reliably predict voting outcomes, highlighting the complex nature of social media's influence on public opinion. Temporal analyses, such as those conducted by ("Temporal Characteristics of Reddit Discussions Across Three Political Events", n.d.), reveal patterns in how Reddit communities engage with political content over time, with different factual communities joining discussions at distinct stages.

In addition to sentiment analysis, researchers have explored more fine-grained approaches to detect political bias and polarization on social media. ("Political Bias and Factualness in News Sharing across more than 100,000 Online Communities" 2021) analyzed ideological variance in Reddit's communities, finding that right-leaning groups often engage more with biased sources. Meanwhile, advanced unsupervised techniques, such as those proposed by ("Fine-grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning" 2022), have used deep learning to predict political leanings based on social media content, clustering users into ideological groups based on text analysis. This study aims to build on these methodologies, applying sentiment and ideological analysis to understand the political dynamics on Reddit during the 2024 U.S. election period.

Research questions

This research investigates the effectiveness of automated political bias detection in digital content through an ensemble machine learning approach. Our primary objective is to develop and evaluate a system that can accurately classify political bias across different forms of written content, from formal news articles to informal social media discussions. We specifically examine how effectively an ensemble model combining Support Vector Machines, Naive Bayes, and Convolutional Neural Networks can identify and categorize political leanings. This question is particularly relevant given the increasing need for automated tools to help readers navigate the complex landscape of political media. Additionally, we investigate the linguistic features and patterns that characterize different political orientations, seeking to understand how language use varies across the political spectrum. The research also explores the generalizability of models trained on formal news articles when applied to more casual, user-generated political discussions, addressing the practical challenge of creating robust classification systems that work across different types of political discourse. These questions are approached through a comprehensive analysis of labeled news articles from AllSides and Reddit political discussions, combining traditional machine learning techniques with deep learning approaches to capture both explicit and subtle indicators of political bias.

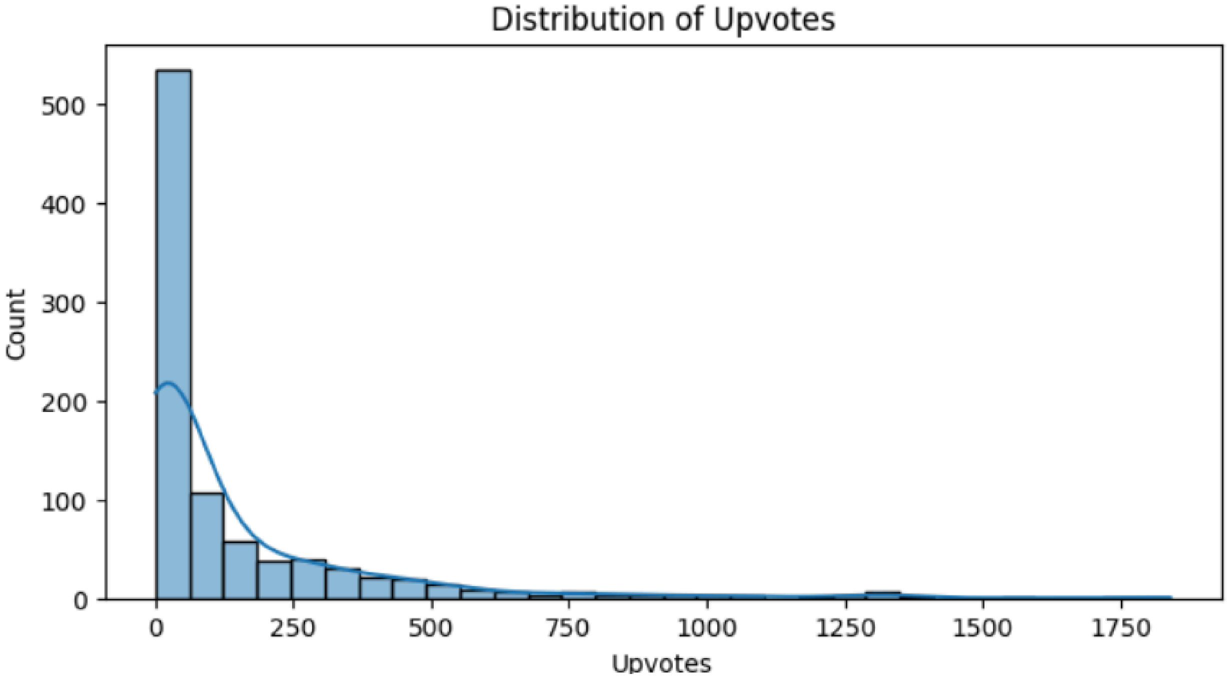
Datasets Description and analysis

Our study utilizes two primary datasets: the AllSides news article collection and Reddit political discussions.

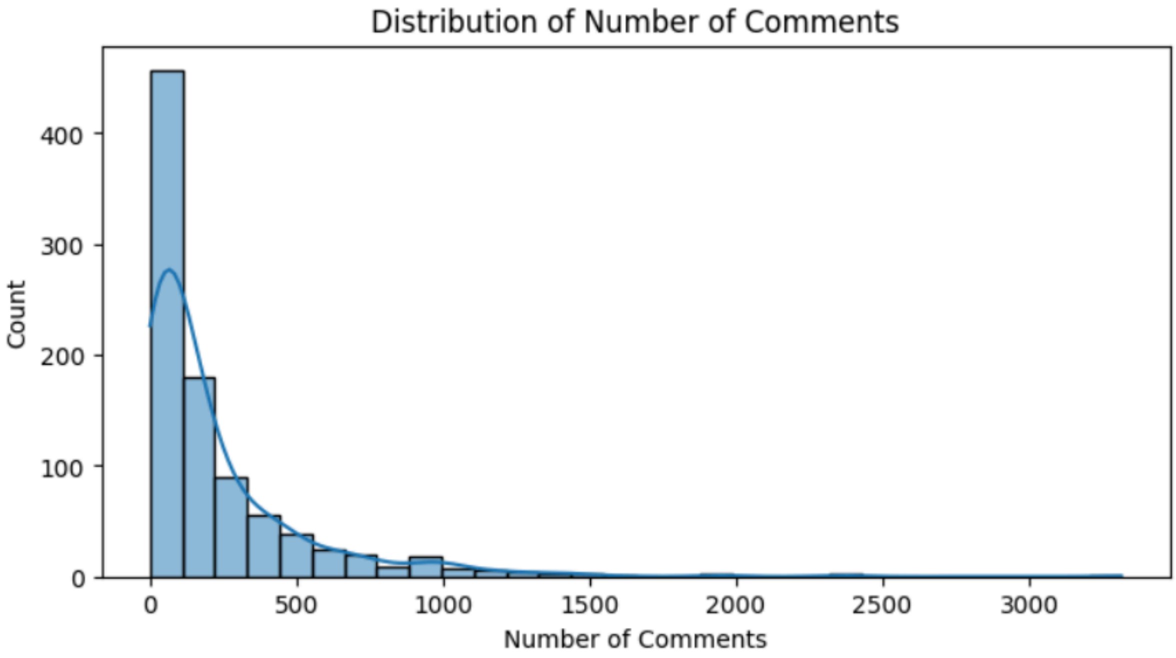
A. Reddit data Crawling and its Exploratory data analysis:

Data for this study was collected from Reddit using the PRAW (Python Reddit API Wrapper) library. The dataset comprises approximately 900 posts from the subreddit r/PoliticalDiscussion, gathered from July 15, 2024, to November 3, 2024. Each post includes various metadata fields: `post_id`, `title`, `content`, `created_utc` (timestamp), `subreddit`, `num_comments` (number of comments), `upvotes`, `upvote_ratio` (ratio of upvotes to total votes), and `author`.

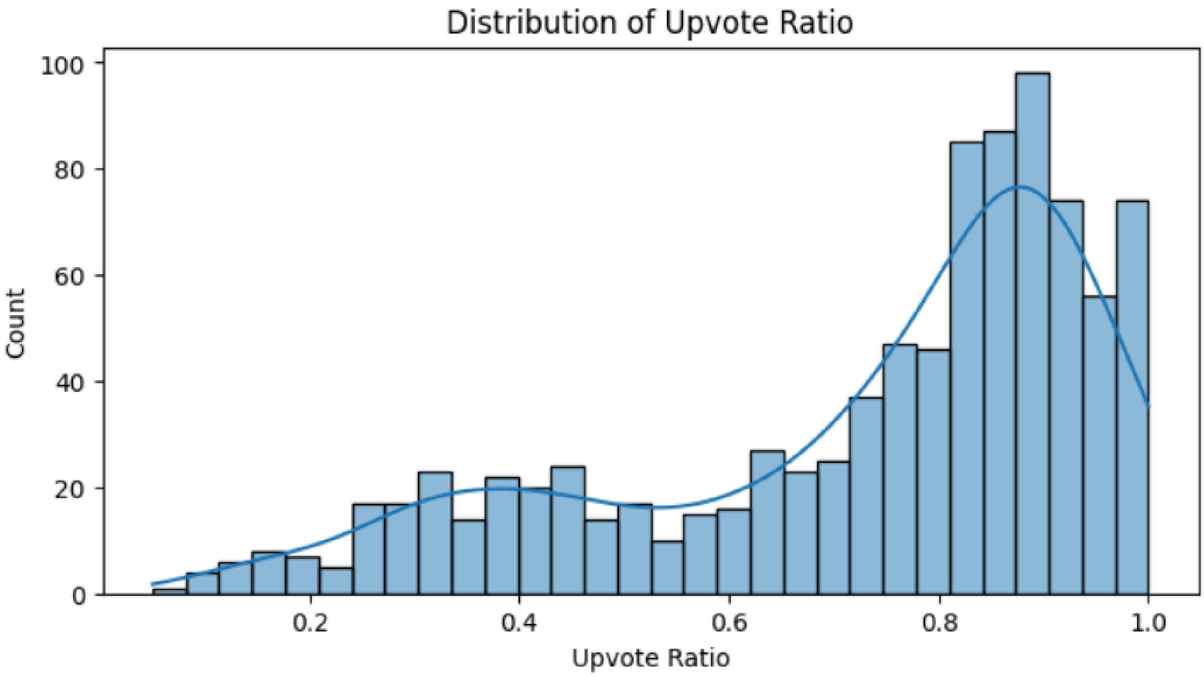
The dataset focuses on the engagement metrics and textual content of each post. Since Reddit posts in political subreddits are often posed as questions, the dataset is particularly suitable for analyzing public inquiry, sentiment, and the framing of political issues.



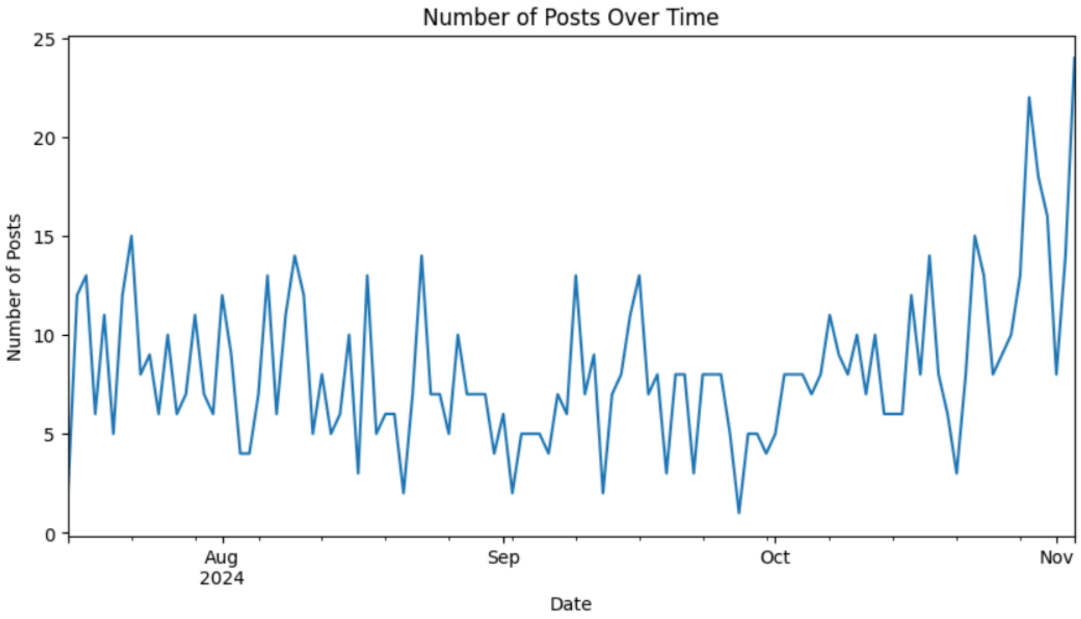
Right-Skewed Distribution: The majority of posts receive a relatively low number of upvotes, while a small portion receives a significantly larger number of upvotes. This indicates a long tail in the distribution. There are likely outliers present in the data, which are the posts with exceptionally high upvote counts.



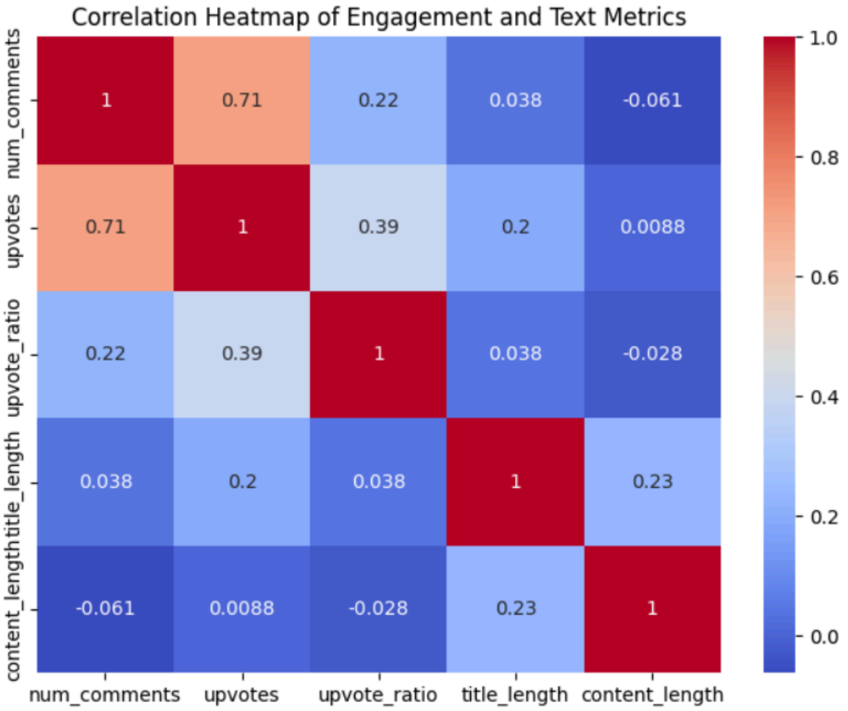
Right-Skewed Distribution: Similar to upvotes, the majority of posts receive a relatively low number of comments, while a small portion receives a significantly larger number of comments. There are likely outliers present in the data, which are the posts with exceptionally high comment counts.



The distribution is roughly unimodal, with a peak around 0.8-0.9. This indicates that most posts receive a moderate to high level of positive engagement.

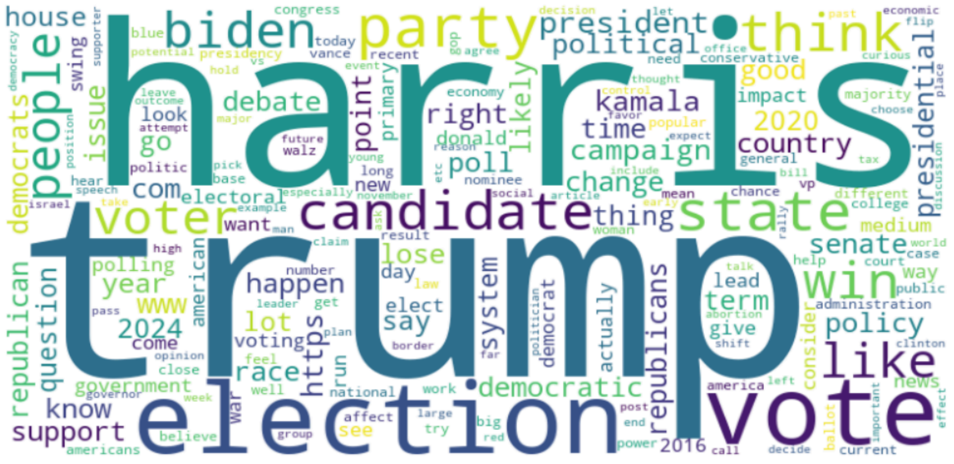


The number of posts fluctuates over time, there are several instances of peaks where the number of posts significantly increases, followed by troughs where the number decreases. There is no clear upward or downward trend exception being around late October to early November which is a few weeks before elections. The fluctuating pattern in post frequency suggests that the platform's activity is influenced by various factors, such as news events, holidays, or other external factors that can drive spikes in activity.



The strong positive correlation between upvotes and comments indicates a symbiotic relationship. Posts that generate more comments are likely to attract more upvotes, and vice versa. This suggests that engagement, as measured by both comments and upvotes, is a key driver of post popularity.

This is the world map for the entire corpus,



B. AllSides Ideological labeled data

This dataset comprises 17,362 articles from the AllSides platform, representing a diverse spectrum of political viewpoints. The dataset exhibits a natural distribution across political orientations, with 23% (3,996 articles) classified as left-leaning, 45% (7,803 articles) as center-aligned, and 32% (5,563 articles) as right-leaning. While showing mild imbalance, this distribution effectively mirrors the real-world landscape of digital political media.

The articles in our dataset encompass a wide range of political discourse, from domestic policy debates to international relations, economic policies, and social issues. AllSides' systematic rating methodology has expertly labeled each piece for political bias, providing a reliable foundation for our machine learning approaches.

Methodology

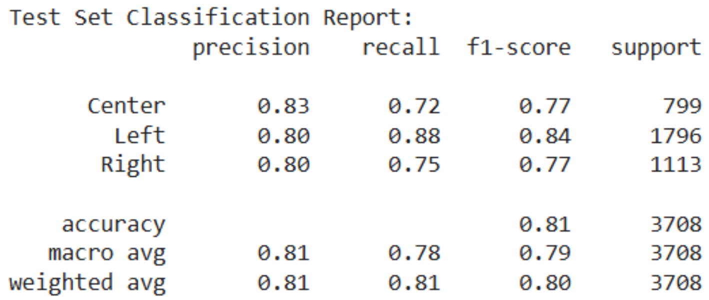
Text Preprocessing and Feature Engineering

Our analysis begins with a comprehensive text preprocessing pipeline to standardize and clean the textual data. This process includes converting text to lowercase, expanding contractions (e.g., "don't" to "do not"), and removing URLs, email addresses, and special characters through regular expressions. The text then undergoes tokenization and stop word removal using NLTK's English stop word list, eliminating common words that carry minimal discriminative value. Finally, lemmatization is applied using NLTK's WordNetLemmatizer to reduce words to their base forms, ensuring different inflections of words are treated uniformly. This standardized text serves as the foundation for our feature extraction process, where we employ both TF-IDF vectorization for traditional machine learning models and word embeddings for deep learning approaches. The processed text maintains essential semantic information while reducing noise and dimensionality, creating a robust foundation for our ensemble classification system.

Support Vector Machine (SVM)

Our implementation of the Support Vector Machine forms a crucial component of the ensemble system, specifically designed to handle the complex nature of political text classification. The model employs a linear kernel, chosen for its effectiveness in high-dimensional text data and computational efficiency. The feature engineering process utilizes TF-IDF vectorization with a carefully selected maximum of 5,000 features, striking a balance between capturing important term relationships and maintaining computational feasibility. This vectorization implements sublinear scaling and L2 normalization, ensuring that the feature space appropriately represents term importance while managing the impact of varying document lengths.

The SVM's hyperparameter optimization process involved extensive grid search cross-validation, focusing primarily on the regularization parameter (C). We explored values ranging from 0.1 to 10.0, using 5-fold cross-validation to ensure robust parameter selection. To address the mild class imbalance in our dataset, we implemented balanced class weights, effectively adjusting the model's sensitivity to different political categories. This configuration proved particularly effective at capturing subtle distinctions between political orientations while maintaining generalization capability.

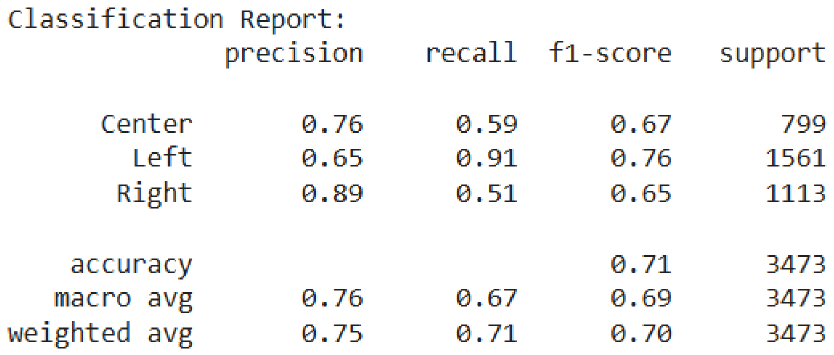


```
Test Set Classification Report:
              precision    recall  f1-score   support

      Center       0.83      0.72      0.77       799
        Left       0.80      0.88      0.84      1796
       Right       0.80      0.75      0.77      1113

    accuracy                           0.81      3708
   macro avg       0.81      0.78      0.79      3708
weighted avg       0.81      0.81      0.80      3708
```

Multinomial Naive Bayes

The Multinomial Naive Bayes classifier serves as our second base model, chosen for its probabilistic approach to text classification and proven effectiveness in document categorization tasks. Our implementation focuses on careful hyperparameter optimization, particularly the smoothing parameter (alpha), which plays a crucial role in handling the sparse nature of text data. Through a comprehensive grid search, we evaluated alpha values ranging from 0.1 to 2.0, along with different prior probability configurations, to find the optimal balance between model flexibility and reliability.

The model processes features using the same TF-IDF vectorization scheme as the SVM, ensuring consistency in feature representation across the ensemble. We enhanced the basic implementation by incorporating feature selection based on chi-squared testing, which helps identify the most discriminative terms for political bias classification. The model's probability estimates undergo calibration to improve prediction confidence, particularly important for the ensemble's voting mechanism. This probabilistic foundation provides valuable insights into classification confidence and complements the other models' decision boundaries.

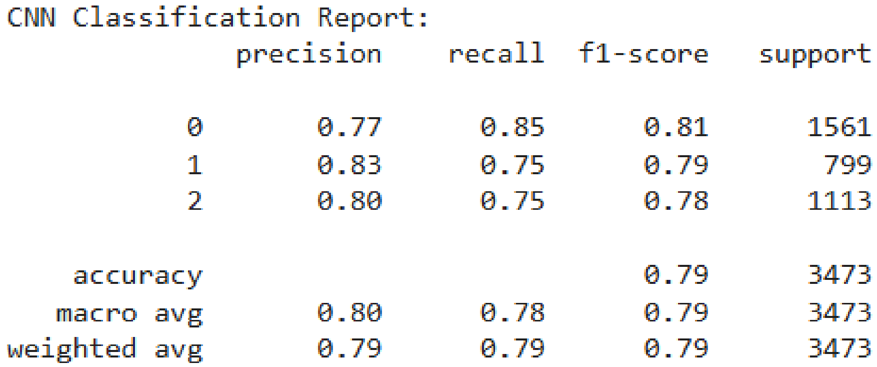


```
Classification Report:
              precision    recall  f1-score   support

      Center       0.76      0.59      0.67       799
        Left       0.65      0.91      0.76      1561
       Right       0.89      0.51      0.65      1113

    accuracy                           0.71      3473
   macro avg       0.76      0.67      0.69      3473
weighted avg       0.75      0.71      0.70      3473
```

Convolutional Neural Network (CNN)

Our CNN architecture represents the deep learning component of the ensemble, designed to capture sequential patterns and contextual relationships in political text. The model begins with an embedding layer utilizing pre-trained GloVe word embeddings (100 dimensions), providing a rich semantic foundation for text representation. The sequence length is dynamically set based on the 95th percentile of text lengths in our dataset, optimizing the balance between information retention and computational efficiency.

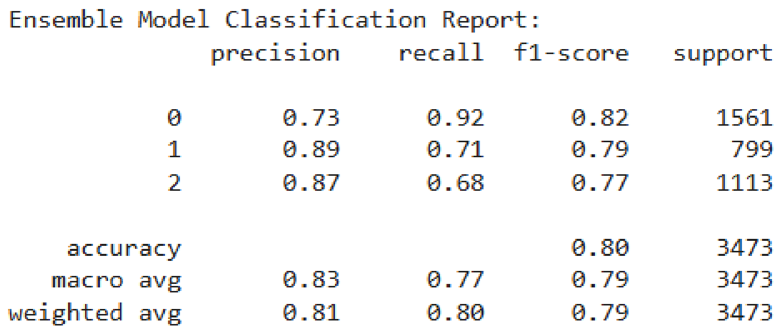
The network architecture consists of two convolutional layers, each with 128 filters and a kernel size of 5, followed by max pooling operations. This design enables the model to identify relevant n-gram patterns and their hierarchical relationships. The global max pooling layer condenses these features before feeding them through dense layers with dropout regularization (0.2 rate), helping prevent overfitting. The final layer employs softmax activation for three-class prediction. The training utilizes the Adam optimizer with a learning rate of 0.001 and implements early stopping to prevent overfitting.



```
CNN Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.85      0.81      1561
           1       0.83      0.75      0.79       799
           2       0.80      0.75      0.78      1113

    accuracy                           0.79      3473
   macro avg       0.80      0.78      0.79      3473
weighted avg       0.79      0.79      0.79      3473
```

Ensemble Integration

The integration of these three models creates a robust ensemble system through a majority voting mechanism. Each model contributes equally to the final classification decision, leveraging their complementary strengths: the SVM's effectiveness with high-dimensional data, Naive Bayes's probabilistic insights, and the CNN's ability to capture sequential patterns. The voting system aggregates predictions from all models, with ties resolved based on confidence scores derived from each model's prediction probabilities. This ensemble approach has proven more robust than any individual model, particularly when dealing with ambiguous cases or content that exhibits characteristics of multiple political orientations.



```
Ensemble Model Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.92      0.82      1561
           1       0.89      0.71      0.79       799
           2       0.87      0.68      0.77      1113

    accuracy                           0.80      3473
   macro avg       0.83      0.77      0.79      3473
weighted avg       0.81      0.80      0.79      3473
```

Topic Modeling on Reddit posts using LDA

The Latent Dirichlet Allocation (LDA) analysis revealed 15 distinct topics in Reddit political discussions, each representing different aspects of political discourse. Here's a comprehensive breakdown of the thematic clusters:

Policy and Economic Issues (Topic 1) focuses on economic policy discussions, with key terms like 'policy', 'inflation', 'economy', and 'tariff', indicating significant discourse around economic concerns and their policy implications.

Social and Identity Politics (Topics 2 & 3) emerge as major themes, with Topic 2 addressing broader identity issues ('woman', 'black', 'president') and Topic 3 specifically focusing on reproductive rights and gender issues ('abortion', 'woman', 'gender', 'ban').

Electoral Politics spans multiple topics: Topic 7 covers electoral mechanics ('electoral', 'college', 'swing', 'blue', 'red'), while Topic 8 focuses on campaign dynamics ('democratic', 'biden', 'nominee', 'debate'). Topic 10 delves into polling and endorsements ('poll', 'pollster', 'endorsement', 'rally').

Legal and Security Issues (Topic 4) encompasses discussions about threats, war, legal matters, and speech rights, suggesting dialogue about national security and constitutional freedoms.

Governance and Policy (Topics 12, 13, & 15) cover different aspects of governance:

Topic 12 addresses institutional politics ('government', 'power', 'corruption')
Topic 13 focuses on economic and demographic factors ('economic', 'business', 'population')
Topic 15 deals with specific policy areas ('legislation', 'gun', 'control', 'spending')

Immigration and Border Policy (Topic 14) emerges as a distinct theme with terms like 'border', 'immigration', 'national', and 'administration'.

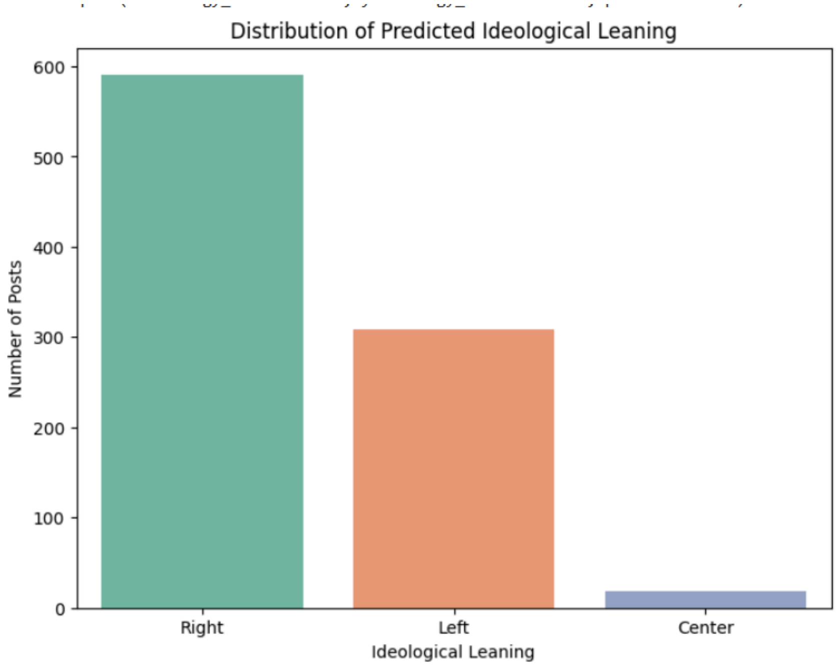
The model effectively captures the multifaceted nature of political discussions on Reddit, from high-level policy debates to specific issue-based dialogues. The distribution of topics suggests a balanced discourse covering both domestic and international issues, social and economic policies, and electoral politics.

Discussion

Our ensemble model's analysis of political discourse on Reddit's r/PoliticalDiscussion presents interesting patterns, though these findings should be interpreted with appropriate caution given the inherent limitations of automated political bias detection. The results suggest certain trends in content distribution and user engagement, but these should be considered preliminary insights rather than definitive conclusions.

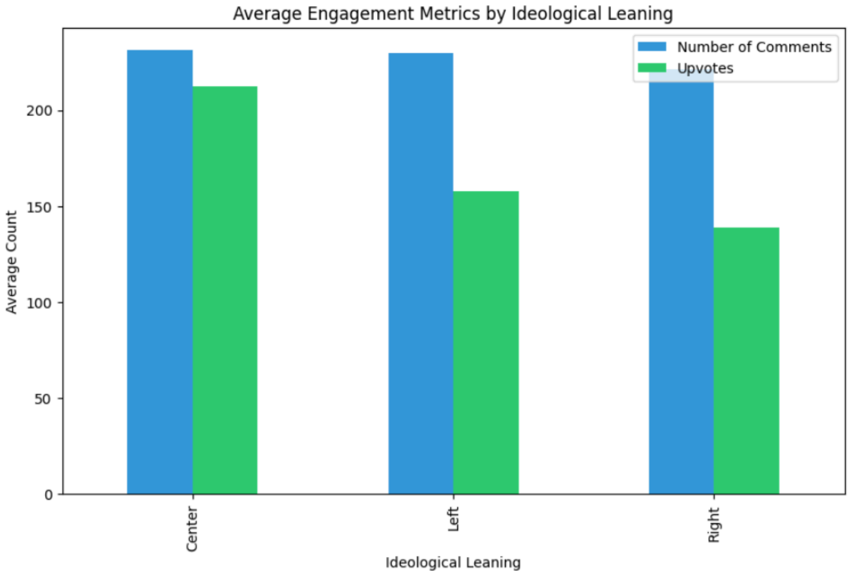
Model Predictions and Distribution

The predicted distribution shows a substantial right-leaning bias in the forum's content, with approximately 600 posts classified as right-leaning compared to 300 left-leaning and fewer than 50 center-aligned posts. However, this stark imbalance, particularly the notably low number of centrist posts, may partially reflect limitations in our classification system. The model might be oversensitive to partisan language markers, potentially misclassifying moderate content as either left or right-leaning. This potential bias in our classification system warrants further investigation and refinement.



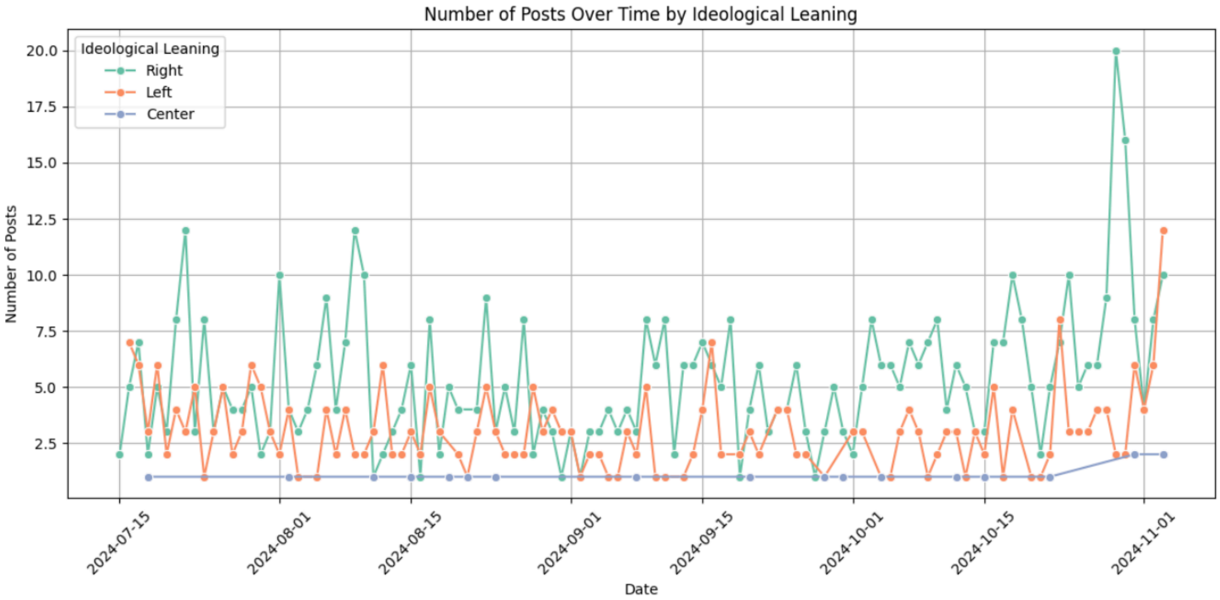
Engagement Analysis

The engagement metrics across different ideological categories present an interesting inverse relationship between post frequency and user interaction. While these patterns are intriguing, they should be interpreted within the context of potential classification errors. The higher engagement with supposedly centrist content, despite its low representation, might indicate either a genuine preference for moderate viewpoints or potential misclassification of highly engaging content. The reliability of these engagement patterns needs further validation through manual content analysis.



Temporal Patterns

The time series analysis reveals varying posting frequencies across ideological categories, with notable volatility in right-leaning content. However, these temporal patterns might reflect both genuine trends and potential systematic biases in our classification system. The observed spikes in activity could represent actual shifts in political discourse or artifacts of our model's sensitivity to certain types of language and topics.



Limitations and Critical Considerations

Several important limitations must be acknowledged:

- The model's apparent bias against identifying centrist content suggests potential oversimplification of political ideology.
- Engagement metrics might be influenced by factors beyond political orientation.
- The classification system may be sensitive to writing style rather than actual political ideology.
- Temporal patterns could be affected by evolving political language and model limitations.

Future Work

Future development should focus on key enhancements including the integration of transformer-based architectures like BERT and GPT to improve contextual understanding. Feature engineering could be enhanced through semantic role labeling and advanced sentiment analysis, enabling better comprehension of political arguments and their emotional undertones. Practical applications could include real-time bias detection systems and browser extensions for immediate bias feedback.

The research could extend into cross-linguistic political bias detection and temporal analysis of bias evolution. Developing more nuanced bias strength metrics and incorporating multi-modal analysis would provide a more comprehensive understanding of political communication. These advancements would contribute to both the technical capabilities of automated bias detection and its practical applications in promoting media literacy.

References

- “A. Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period.” 2023. https://doi.org/10.1007/s13278-023-01048-1

- Chaudhry, Hassan N. 2021. “Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020.” MDPI. https://www.mdpi.com/2079-9292/10/17/2082.
- “Fine-grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning.” 2022. Journal of Artificial Intelligence Research. https://www.jair.org/index.php/jair/article/view/13112

- “Political Bias and Factualness in News Sharing across more than 100,000 Online Communities.” 2021. View of Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. https://ojs.aaai.org/index.php/ICWSM/article/view/18104/17907

- “Temporal Characteristics of Reddit Discussions Across Three Political Events.” n.d. Carnegie Mellon University. Accessed November 6, 2024. https://www.cmu.edu/ideas-social-cybersecurity/events/murdock_temporal_characteristics.pdf