Detecting Political Affiliation and Linguistic Patterns in Reddit Discourse

Faizan Waheed
fawaheed@iu.edu
April 20, 2025

Abstract

This paper investigates how political ideology is reflected in Reddit discourse by analyzing text from two ideologically distinct communities: r/Democrats and r/Republicans. We apply a combination of traditional machine learning classifiers (Logistic Regression, SVM, Random Forest) and state-of-the-art transformer-based models (DistilBERT, BART, and RoBERTa) to predict political affiliation from post content. Text data is preprocessed using standard natural language processing techniques, followed by TF-IDF vectorization for classical models and tokenization for fine-tuned transformers. In addition to classification, we extract linguistic features such as lexical richness, pronoun usage, readability, and sentiment, and generate word clouds to visualize thematic patterns. Our results show that transformer models outperform traditional models in precision and accuracy, while classical models achieve higher recall and F1-scores. Linguistic analysis further reveals that Democrats tend to use more personal language and express more positive sentiment, whereas Republicans demonstrate higher lexical diversity and readability. These findings underscore the potential of NLP techniques for detecting ideological signals in social media and contribute to a deeper understanding of how partisanship is manifested linguistically in online political communities.

1. Introduction

In an era of deepening political polarization, understanding how ideological divisions are expressed and reinforced through online discourse has become increasingly important. Social media platforms now play a central role in shaping public opinion and facilitating political engagement, but they also contribute to the fragmentation of political communities into ideologically homogeneous spaces. Reddit, one of the most influential user-driven forums, offers a unique environment to examine this dynamic. With thousands of thematic subreddits and semi-anonymous participation, it fosters open political discussions while also reflecting clear partisan boundaries—particularly in politically-oriented communities such as r/Democrats and r/Republicans.

This study is motivated by the need to explore how political affiliation is encoded in language and whether it can be reliably inferred from user-generated content. Uncovering such linguistic markers not only enhances our understanding of online ideological alignment but also has broader implications for content moderation, misinformation detection, political campaign strategy and computational political science. Reddit's longer-form and often more nuanced political discussions make it an especially valuable platform for applying Natural Language Processing (NLP) tools that go beyond surface-level text features.

While prior research has made significant strides in ideology detection on platforms like Twitter, where content is brief and often reactive, a critical research gap lies in the application and comparative evaluation of advanced transformer-based language models—particularly on Reddit's extended discourse. Although models like BERT and its variants have been successfully applied to social media data (Preoţiuc-Pietro et al. 2017; Jiang et al. 2021), existing work has not yet fully explored how models such as DistilBERT, BART, and RoBERTa perform on Reddit data in the context of political affiliation classification.

This paper addresses this gap by leveraging a Reddit dataset collected from r/Democrats and r/Republicans to conduct both classification and topic modeling tasks. We compare the effectiveness of traditional machine learning models (logistic regression, SVM and Random Forest) with state-of-the-art transformers (DistilBERT, BART, RoBERTa) in predicting political affiliation based on post content. Additionally, we apply BERTopic to uncover dominant themes within each subreddit. Through this dual approach, the study contributes new empirical insights into how political identity is expressed in Reddit communities and evaluates the capacity of modern NLP techniques to capture ideological signals in longer-form discourse.

2. Related work

The intersection of Natural Language Processing (NLP) and political communication has garnered significant attention in recent years, particularly as researchers seek to understand how language use on social platforms reflects underlying ideological divisions. A growing body of work has demonstrated that linguistic features can serve as strong indicators of political affiliation, while also shedding light on the structure of online discourse, the formation of echo chambers, and the thematic focus of partisan communities. The following studies represent foundational contributions that inform the design and motivation of our project.

Preoţiuc-Pietro et al. (2017) conducted a large-scale study to predict political ideology on Twitter by analyzing linguistic features such as lexical choices, hashtags, and topical patterns. Using traditional machine learning models, they demonstrated that political orientation could be reliably inferred from textual cues, even in short-form content like tweets. This work is directly relevant to our project, as it establishes the feasibility of using textual content alone for ideological classification. We build on their framework by applying similar content-based methods to Reddit, which differs from Twitter in its allowance for longer, more complex discourse. Our findings reinforce their conclusion that linguistic patterns serve as strong indicators of political alignment, with transformer-based models yielding high accuracy even without metadata or user history.

Jiang et al. (2021) examined ideological polarization during the COVID-19 pandemic using Retweet-BERT, a model that combines textual representations with network-based features to detect political leanings on Twitter. Their results highlighted the prevalence of ideological homophily, showing that conservative users tend to cluster more tightly within echo chambers. While our study does not incorporate social network structures, this work validates the use of transformer architectures—such as RoBERTa and BART—for ideological classification based purely on text. Moreover, their findings on digital polarization contextualize our linguistic and

sentiment analysis, which shows similar divides in Reddit communities, thus connecting both methodological and thematic elements of their research to our own.

Iyyer et al. (2014) contributed to political text classification by developing a recursive neural network (RNN) trained on congressional floor speeches. Unlike traditional models focused on surface-level word frequencies, their approach captured hierarchical syntactic relationships and rhetorical framing to detect ideological bias. This deep learning framework demonstrates that subtle linguistic structures can encode political stance. Their work supports our rationale for using transformer models like DistilBERT and BART, which also capture contextual and semantic nuances in text. Our results align with theirs by showing that deep contextual models outperform traditional classifiers, particularly in capturing the tone and structure of politically motivated discourse on Reddit.

Gerrish and Blei (2011) introduced a supervised topic modeling approach to predict legislative roll call votes from textual descriptions of bills. Their model not only classified political alignment but also revealed issue-specific themes underlying voting behavior. This dual capacity—prediction and interpretation—mirrors the goals of our project, which combines classification with topic modeling using LDA and BERTopic. Their work is relevant as it highlights the value of uncovering latent thematic structures that correspond with ideology, a principle we applied to Reddit posts to identify dominant topics across partisan communities. The interpretability of their method informed our own use of topic modeling to complement the quantitative results from classifiers.

Baly et al. (2018) explored how linguistic features could be used to predict the political bias and factual accuracy of news media sources. By integrating content-based analysis with shallow metadata, they classified sources along ideological lines with high precision. Their findings underscore that textual features—such as sentiment, style, and topicality—are sufficient to infer political orientation, even without user interaction data. This directly aligns with our project, which also relies on textual features to classify Reddit posts. Our use of sentiment analysis and lexical richness mirrors their emphasis on stylistic indicators of bias, and our findings confirm that political identity can be effectively inferred from the way language is structured and expressed in online political discourse.

Together, these studies establish a robust foundation for examining political discourse through the lens of natural language processing. They demonstrate that linguistic patterns can effectively reveal ideological alignment, whether in social media posts, legislative texts, or news content. By leveraging insights from these works, our study advances this line of research by applying both classical and transformer-based models to Reddit—a platform that remains relatively underexplored despite its rich, community-driven political discussions. Furthermore, by integrating classification and topic modeling, our approach not only predicts political affiliation but also reveals the thematic structures that differentiate partisan communities, contributing both methodological innovation and empirical depth to the ongoing study of political polarization online.

3. Data Collection

To develop a dataset suitable for political ideology classification, we collected textual content from two ideologically polarized Reddit communities: r/Democrats and r/Republicans. These subreddits were deliberately chosen due to their clear partisan alignment, high levels of user engagement, and active discussion of contemporary political issues. As self-organized political communities, they provide organic, user-generated content that reflects grassroots ideological expression, making them valuable sources for studying partisan discourse at scale.

Data collection was conducted using the Python Reddit API Wrapper (PRAW), which provides authenticated access to Reddit's public API and allows developers to interact with subreddit content in a structured and programmatic way. We authenticated via Reddit's developer credentials and queried the top 1,000 posts from each subreddit. The "top" sorting parameter used Reddit's native ranking algorithm, which prioritizes posts with high engagement over time, ensuring that we sampled high-quality, community-relevant content rather than transient or low-visibility posts.

For each retrieved post, we extracted two key fields: the title and the selftext. The title captures the post's headline or summary, while the selftext contains the main body of the user's submission. These two components were concatenated into a single string for analysis, allowing the model to interpret both the framing and the narrative of each post. This combined text was stored in a column labeled text. To ensure data integrity, we filtered out posts that lacked a selftext or contained empty values after concatenation.

The final dataset consisted of 1,989 posts, comprising 997 from r/Democrats and 992 from r/Republicans. Each post was labeled according to its originating subreddit using a binary label: Democrat or Republican. The complete dataset was stored in a pandas DataFrame with two primary columns: text, containing the full textual content of the post, and label, denoting the subreddit affiliation. Both columns were of type object, enabling compatibility with downstream natural language processing workflows.

Importantly, the data collection process strictly adhered to Reddit's public API terms of service. Only publicly visible content was retrieved, and no personally identifiable information (PII), user handles, or metadata beyond the textual content of the post was collected, stored, or analyzed. This ensured that the study remained ethically compliant while enabling the extraction of meaningful textual data for political discourse analysis.

4. Methods

This study employed a multi-stage methodology that combined supervised machine learning with modern transformer-based architectures to classify political affiliation based on Reddit discourse. The pipeline consisted of data preprocessing, model training and optimization, and linguistic analysis to uncover deeper stylistic patterns between ideological groups.

In the preprocessing phase, all Reddit post content was standardized by converting to lowercase and removing noise such as URLs, punctuation, HTML tags, and bracketed expressions. Stopwords were removed using NLTK's English stopword list, and each token was lemmatized using the WordNet Lemmatizer to preserve semantic root forms. The cleaned output was stored in a new column (clean_text) and used as input to both classical and deep learning models.

The first classical model implemented was Logistic Regression, a linear classifier well-suited for high-dimensional text data. It was embedded in a pipeline with TF-IDF vectorization, enabling transformation of the textual input into weighted numerical features. Hyperparameters such as the inverse regularization strength C, the n-gram range, and document frequency thresholds (min_df, max_df) were tuned using grid search with five-fold stratified cross-validation. To address potential class imbalance, we evaluated both balanced and unbalanced class weights. The model was trained using the saga solver and L2 penalty to enforce generalizability. Performance was assessed using precision, recall, F1-score, and log loss.

The Support Vector Machine (SVM) model employed a linear kernel and was wrapped within a CalibratedClassifierCV to generate probability estimates required for probabilistic evaluation. The TF-IDF parameters were tuned in a similar manner, focusing on unigrams and pruning rare or overly common features. The SVM classifier was regularized via the C parameter and trained with class-weight adjustments to account for distributional skew. This model demonstrated strong performance in accuracy and precision while offering robustness to overfitting due to the simplicity of its kernel.

Next, we trained a Random Forest classifier, an ensemble model built on decision trees. To enhance interpretability and reduce overfitting, we constrained the model with a limited number of estimators (trees), shallow tree depth (max_depth), and stricter thresholds for splits (min_samples_split). TF-IDF features were limited to unigrams with high-frequency cutoff thresholds to focus the model on more semantically stable terms. Class weights were balanced to ensure fairness across political categories. The model was evaluated on the same stratified splits using all standard classification metrics.

In the transformer-based portion of our pipeline, we first fine-tuned DistilBERT, a compact version of BERT optimized for speed and efficiency. The model was trained using HuggingFace's Trainer API, and tokenization was handled using the DistilBERT tokenizer with input sequences truncated to 128 tokens. We introduced dropout regularization and weight decay to prevent overfitting and configured early stopping to terminate training upon performance plateau. The learning rate and batch size were chosen to balance convergence and generalization. This model served as a computationally efficient baseline among transformers, with significantly fewer parameters than BART and RoBERTa.

We then trained BART model (facebook/bart-base), a denoising autoencoder transformer with a bidirectional encoder and autoregressive decoder. Fine-tuning was again handled using the HuggingFace trainer, with additional techniques such as cosine learning rate decay and gradient accumulation used to stabilize training. Dropout layers were intensified to enhance generalization,

and early stopping was applied to reduce overfitting. The model was trained for six epochs, and evaluation metrics were recorded at each epoch to track convergence. BART's architecture enabled it to capture complex sentence structure, but it was comparatively slower to train due to its size.

Finally, we fine-tuned the Twitter RoBERTa model (cardiffnlp/twitter-roberta-base), which was pre-trained specifically on social media data, making it well-suited for Reddit's informal and varied discourse. Similar to other transformer setups, we applied an 80/20 stratified split, used a maximum sequence length of 128, and tuned dropout and weight decay to improve generalization. The training regimen included cosine learning rate scheduling, warmup steps, and early stopping. This model demonstrated strong performance in precision and overall accuracy, benefiting from its social media-focused pretraining.

To complement our classification efforts, we conducted a linguistic analysis to examine stylistic differences across subreddits. Using spaCy, we extracted metrics such as lexical richness, pronoun usage, and Flesch Reading Ease to quantify complexity and tone. Sentiment analysis was conducted using VADER to assign compound sentiment scores to each post. These features were averaged by label and used to compare discourse patterns. Additionally, we generated word clouds using the WordCloud library to visualize dominant themes in each subreddit, providing a qualitative lens into partisan language and focus areas.

5. Findings

The results of our experiments revealed distinct trade-offs between model architectures in terms of classification performance. Among the classical models, Support Vector Machine (SVM) achieved the highest overall accuracy at 0.648, slightly outperforming Random Forest (0.633) and Logistic Regression (0.623). However, when considering transformer-based models, DistilBERT emerged as the top performer with an accuracy of 0.661, followed closely by Twitter RoBERTa (0.651) and BART (0.646), suggesting that modern contextual embeddings can indeed provide a measurable improvement over traditional feature-based approaches.

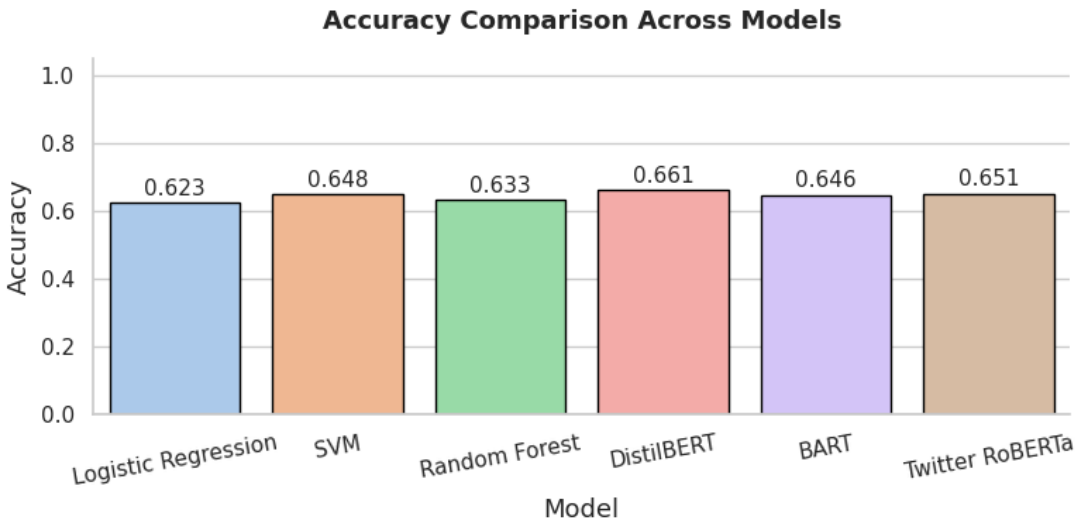


Figure 1: Accuracy comparison across all classification models, illustrating differences in performance between traditional machine learning and transformer-based approaches

In terms of precision, BART achieved the best result with a score of 0.681, indicating it was most conservative in labeling posts as belonging to a specific political class—focusing more on correctness than coverage. DistilBERT and Twitter RoBERTa also demonstrated strong precision (0.647 and 0.645, respectively), outperforming all classical models, with Logistic Regression lagging behind at 0.582.

However, a different trend emerged in the recall metric. Classical models such as Logistic Regression and Random Forest yielded the highest recall scores (0.885 and 0.860, respectively), indicating that they were more liberal in capturing posts across both classes, at the expense of precision. In contrast, transformer-based models such as BART and Twitter RoBERTa had notably lower recall (0.555 and 0.680, respectively), highlighting a tendency to miss positive cases while minimizing false positives. Precision and recall comparisons are shown in Figure 4 and Figure 5, respectively, with detailed metric values provided in the Appendix.

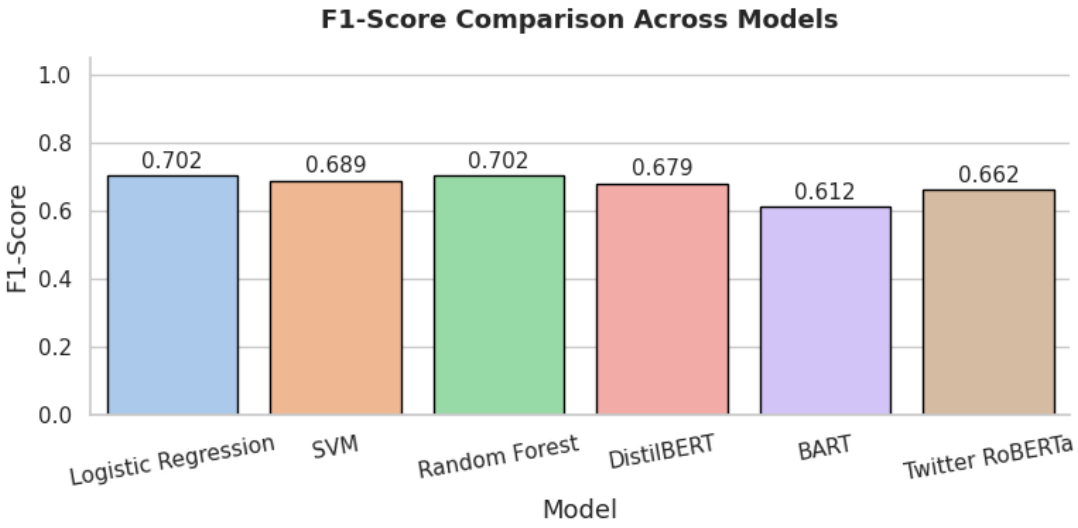


Figure 2: F1-Score comparison across all classification models, illustrating differences in performance between traditional machine learning and transformer-based approaches

When balanced through the F1-score, which combines precision and recall, both Logistic Regression and Random Forest shared the highest performance at 0.702, with SVM closely following at 0.689. Among the transformers, Twitter RoBERTa scored 0.662, followed by DistilBERT (0.679) and BART (0.612). These results suggest that while transformer models achieved marginal improvements in precision and overall accuracy, classical models—especially Random Forest and Logistic Regression—remained competitive due to their strong recall and balanced performance.

These findings indicate that no single model dominated across all evaluation metrics. Transformer-based models provided better precision and interpretive power through deep contextual representations, but classical models offered better recall and more robust performance on this moderately-sized Reddit dataset. The trade-off between precision and recall is particularly important in ideological text classification, as it reflects whether models prioritize correctness over coverage or vice versa.

Quantitative linguistic feature analysis further revealed systematic stylistic differences. Republican posts exhibited slightly higher lexical richness (0.8062) than Democratic posts (0.7823), indicating a broader and more varied vocabulary. In contrast, Democratic posts used more personal pronouns on average (pronoun ratio = 0.0815) compared to Republican posts (0.0708), suggesting a more personalized or individual-centered style of communication. Readability scores were moderately high for both groups, with Republican posts scoring 66.4 and Democratic posts 65.1 on the Flesch Reading Ease scale—both indicating standard to easy readability. Finally, sentiment analysis using VADER showed that Democratic posts expressed a more positive average sentiment (compound score = 0.0301) than Republican posts (0.0021), though both scores hovered near neutrality.

Table 1: Comparison of average linguistic features between Democratic and Republican subreddit posts.

Subreddit	Lexical Richness	Pronoun Ratio	Flesch Reading Ease	Sentiment Score
Democrat	0.7823	0.0815	65.14	0.0301
Republican	0.8062	0.0708	66.40	0.0021

In addition to classification results, our linguistic and lexical analysis provided valuable insight into how language differs between the two ideological communities. The word cloud visualizations highlighted distinctive thematic emphases within each group. Posts from r/Democrats prominently featured terms such as *vote*, *Trump*, *Republican*, *Biden*, and *president*, suggesting an ongoing focus on electoral processes, criticism of Republican figures, and internal discussion on Democratic leadership. Conversely, posts from r/Republicans were dominated by terms like *Trump*, *Democrat*, *people*, *Biden*, and *election*, reflecting a similar attention to political opponents, as well as national identity and institutional critique. Notably, terms such as *freedom*, *truth*, *media*, and *BLM* appeared more frequently in Republican discourse, signaling a concern with cultural and ideological narratives.

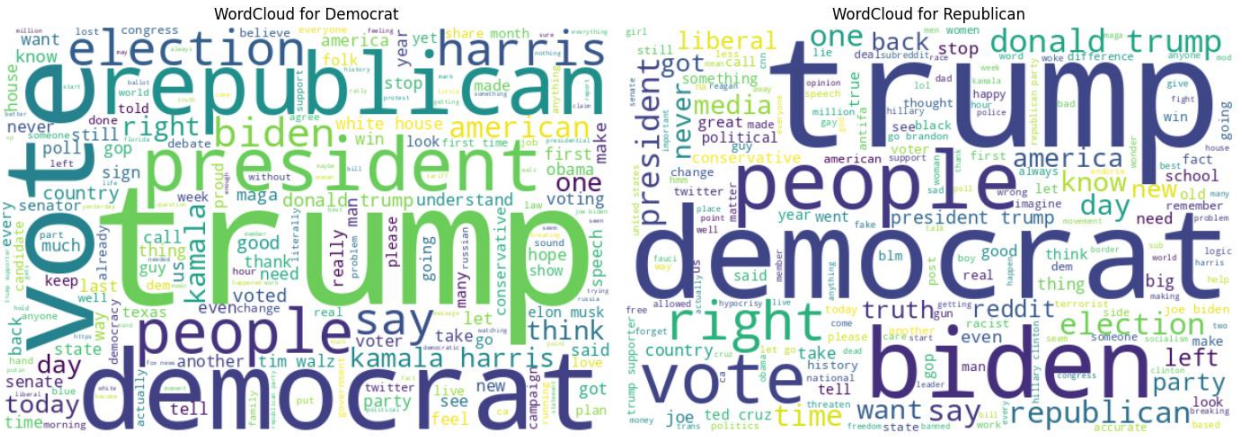


Figure 3: Word cloud visualization of frequently used terms in r/Democrats and r/Republicans, highlighting thematic and lexical differences between the two subreddits.

Together, these findings point to subtle but meaningful differences in how each community articulates political ideas. While both subreddits engage in discussions centered on similar political entities and figures, their lexical styles, use of personal language, and emotional tone distinguish the ideological framing and affective strategies at play.

6. Conclusion

This study investigated how political affiliation is expressed through language on Reddit by analyzing user-generated content from r/Democrats and r/Republicans. Using a combination of classical machine learning models and transformer-based architectures, we demonstrated that political ideology can be inferred with moderate accuracy based solely on textual features. While transformer models such as DistilBERT and RoBERTa offered marginal improvements in precision and interpretive depth, classical models like Logistic Regression and Random Forest remained highly competitive, particularly in recall and overall F1 performance. These results highlight the trade-offs between complex language models and simpler, more interpretable algorithms, especially when applied to medium-scale, noisy social media data.

Beyond classification, our linguistic analysis revealed important differences in discourse patterns between the two partisan communities. Posts from Republican users exhibited greater lexical diversity and higher readability scores, whereas Democratic posts showed a higher rate of pronoun usage and slightly more positive sentiment. Word cloud visualizations confirmed a shared focus on political figures and elections, but also revealed diverging topical emphases that reflect each group's ideological priorities. These stylistic and thematic distinctions support prior findings in computational political science that online communities not only differ in what they say, but also in how they say it.

Overall, our findings contribute to the growing body of work on political text classification and polarization in online platforms. By applying both traditional and state-of-the-art NLP methods to Reddit discourse, we validate the feasibility of inferring ideological alignment from language and

uncover the communicative patterns that shape partisan identity in digital environments. These insights may inform future research in political communication, computational social science, and content moderation by offering a linguistic lens into the mechanisms of ideological division. Future extensions of this work may benefit from incorporating network-based features (e.g., upvotes, user interactions) or expanding the dataset temporally to observe how discourse evolves in response to political events.

Reference

[1] Baly, Ramy, et al. “Predicting Factuality and Bias of News Media Sources.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3528–3539. https://aclanthology.org/D18-1389/

[2] Gerrish, Sean, and David M. Blei. “Predicting Legislative Roll Calls from Text.” *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 489–496. http://www.icml-2011.org/papers/333_icmlpaper.pdf

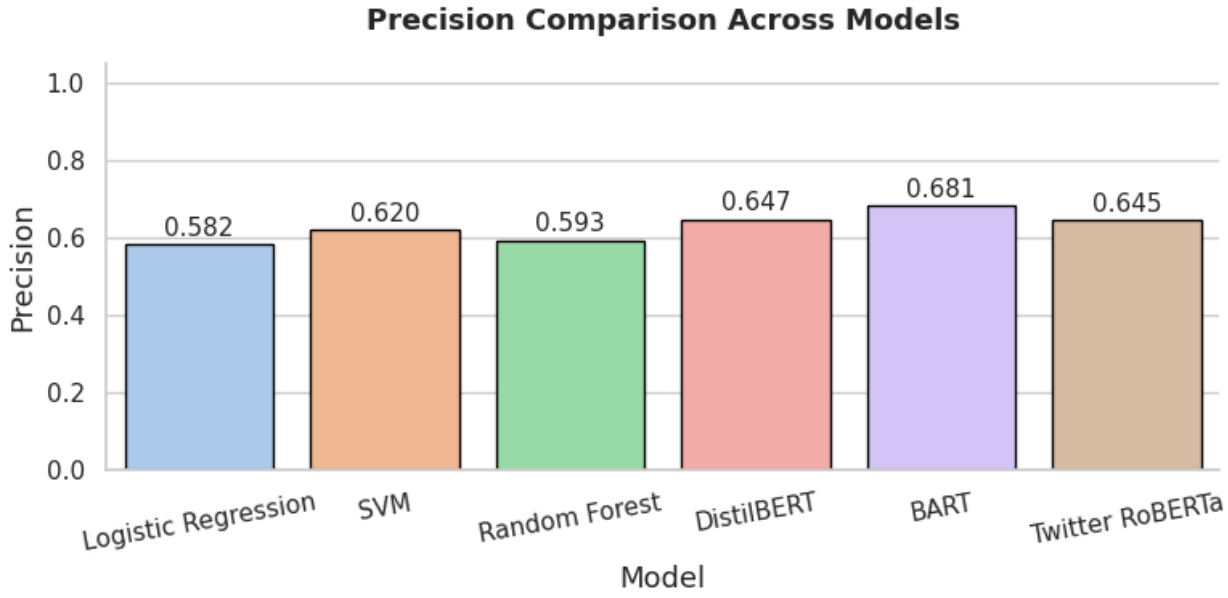
[3] Iyyer, Mohit, et al. “Political Ideology Detection Using Recursive Neural Networks.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1113–1122. https://aclanthology.org/P14-1105.pdf

[4] Jiang, Julie, et al. “Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study.” *JMIRx Med*, vol. 2, no. 3, 2021, p. e29570. https://arxiv.org/abs/2103.10979

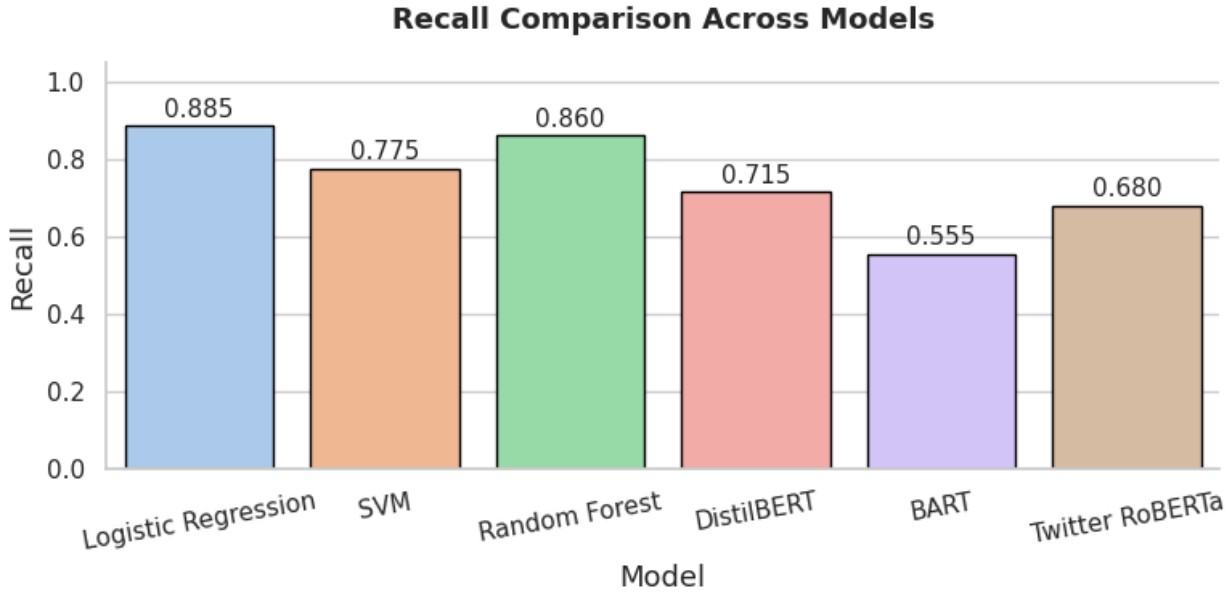
[5] Preoţiuc-Pietro, Daniel, et al. “Beyond Binary Labels: Political Ideology Prediction of Twitter Users.” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 729–740. https://aclanthology.org/P17-1068.pdf

Appendix

[1] Figure 4: Precision comparison across all classification models, illustrating differences in performance between traditional machine learning and transformer-based approaches



[2] Figure 5: Recall comparison across all classification models, illustrating differences in performance between traditional machine learning and transformer-based approaches



[3] The complete implementation, including the detailed code and all associated CSV files, is available on GitHub at: https://github.com/faizan9536/Assignment-3-SMM.