International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Political Speech Bias and Ideological Stance Classification Using Natural Language Processing and Statistical Linguistic Features

Sankati RamaKrishna[1], A. Nandini[2], D. NagaLakshmi[3], G. Supriya[4], K. Rani[5]
[1]Assistant Professor, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
Email: rk4uin2010@gmail.com[1]
[2,3,4,5]B. Tech Student, Dept. of CSE–Data Science
KKR & KSR Institute Of Technology and Sciences, Guntur
Email: 22jr1a4404@gmail.com[2], 22jr1a4413@gmail.com[3], 22jr1a4418@gmail.com[4], 22jr1a4425@gmail.com[5]

***Abstract*—The speeches delivered by the government are an integral part of public and democratic decision-making. Still, the speeches are dominated by a political ideology and strategic framing, thus being difficult for citizens to understand political messages. The task of the detection of political ideology and political bias in speeches has become a task in the research area of computational linguistics.**

This paper offers a method that can be used to classify bias and political ideology in political speeches using a combination of Human Language Processing techniques and linguistic statistical features. The proposed system uses syntactic features derived from political speeches documented in a manner that can be applied to supervise machine learning classification models. The use of linguistic statistical variables such as word diversity, polarity, pronoun mentions, and syntactic sophistication in com- bination with Human Language Processing techniques such as TF-IDF matrices and n-grams helps in achieving better results in classification experiments conducted in this study.

***Index Terms*—Political Speech, Bias Detection, Ideological Stance, NLP, Machine Learning, Deep Learning**

I. INTRODUCTION

Politics affects many aspects of our daily life. Speeches by our political leaders represent one of the most important methods by which leaders communicate to the masses. While political leaders address their nations through these speeches, they express their ideas, promises, and opinions regarding their nations and societies. While people listen to the speech made by their political leaders, they feel convinced or affected by the address, but they lack understanding of the concealed meaning behind the speech. Leaders often select their words to represent their ideas or political parties.

The analysis of political speeches manually is a challenge because there are many speeches from various sources and from various periods. Also, a speech can be understood in various ways by various people, and this causes ambiguity and the judgment of the individual to come into play. Because of these factors, computer analysis of government text is useful. Natural Language Processing is an aid in the analysis of text because the text is broken down into simple units such as words and sentences. The use of phrases such as common phrases and the tone of the speech reveal the ideology of the speech-maker.

In the present project, NLP and some simple statistical methods are used to find bias and conceptual positions from public speeches. Machine learning models learn the patterns from text and classify speeches based on those patterns. This is much quicker and more consistent than checking speeches manually.

The main goal of the project is to understand public-government speeches in a better way, without being completely dependent on human judgment. This work is useful for learning purposes, and it shows how basic language analysis, along with technology, can be used to study political communication.

II. PROBLEM STATEMENT

Ideologically biased content has become more prevalent due to the quick expansion of digital political communication via social media, online speeches, and debates. Bias detection is a difficult and non-trivial task because political actors frequently use subtle linguistic strategies to express ideological positions and sway public opinion. Political discourse analysis by hand is inherently subjective, time-consuming, and inappropriate for large-scale data.

The majority of current computational methods rely on sentiment analysis or keyword-based techniques, which are inadequate to capture the complex statistical linguistic and ideological aspects of political speech. Furthermore, a lot of models don't incorporate statistical language patterns, lexical distribution, and linguistic structure—all of which are essential for differentiating ideological stance from sentiment.

Consequently, there is a substantial research gap in creating an automated and comprehensible framework that successfully integrates statistical linguistic features with Natural Language Processing (NLP) techniques to identify political bias and categorize ideological stance. To enable scalable, impartial, and data-driven analysis of political discourse, this gap must be closed.

This research aims to design and evaluate a computational model capable of accurately identifying bias and classifying ideological orientation in political speeches by leveraging NLP methods and statistical linguistic analysis, thereby contributing to transparent and reliable political text analytics.

III. LITERATURE SURVEY

Recently, the study of governmental speech has become increasingly important because the rise in public content has significantly affected the reports and speeches, including the online platforms [1]. These governmental speeches and objects might involve some latent favor and ideology that is imperceptible to the reader. It is because of this that scientists and researchers continue to look for ways to automatically discover governmental favor and conceptual stance in speeches using the techniques of Human Language Processing.

A number of research works have been carried out to identify government ideology through use of written texts like reports, objects, as well as statements by the government. The works were carried out using traditional approaches for statistical learning in conjunction with aspects such as word frequency, emotions, as well as topic distribution. It was apparent that government ideology could be identified through analysis of frequent patterns in language use as well as word choice. Labeling was a problem based on human analysis.

Certain researchers examined the task of detecting theories in updating entities by correlating the ideology of the script to the learning of the source of the reports. This made the task of labeling easier and helped in creating large datasets. Certain studies in this field have shown that it is possible to classify news articles in broad ideological categories such as left or right by reading their contents [3]. These studies have indicated that there can be certain select words in the reports that indicate the leaning of the reports.

As a result of advancements in NLP, new techniques like deep learning emerged for better accuracy in classification. It was found that deep learning techniques are able to learn complicated patterns in languages in a better way than other traditional classifiers [9]. For political text classification, the performance of these models improved significantly when a large amount of data is considered. But at the same time, they are also quite resource-intensive, which is a major setback to their use.

Other research involved the classification of political stance on the basis of linguistic features that were structured in a statistical manner. The significance of sentence formation, keyword, polarity of sentiment, and term distribution in bias identification was brought to the forefront by these studies. These statistical linguistic features were very helpful in keeping results simple while minimizing complexity. Such projects can be very aptly applied in academic settings.

Public discourse research was also conducted on the study of public speeches rather than on reports on objects. Public speeches are different from news content since they are designed to be convincing and emotionally charged. Research revealed that speeches employed rhetorical devices to sway public opinion, thus making them an excellent research endeavor to explore ideologies. Methods used included statistical methods like tokenization, part-of-speech tagging, and frequency analysis to draw insightful meaning from documentation on speeches.

On the whole, there is confirmation through the available literature that public bias and ideology can be identified effectively by employing NLP methods and statistical linguistic variables. Though there is promising performance by deep learning algorithms, it has been quite effective along with machine learning algorithms when paired with appropriate linguistic variables. It justifies the purpose behind this project, which focuses on public speeches analyzed by NLP algorithms for bias and ideology classification.

IV. PROPOSED SYSTEMS

Recent research in natural language processing (NLP) and political science suggests a system for classifying bias in political speech and identifying ideological stance. This system usually combines traditional linguistic features with deep learning models in a multi-layered pipeline.

The system aims to categorize speeches as Left, Right, or Neutral, often focusing on parliamentary, parliamentary debates, or political media headlines.

A. System Architecture Overview

A robust system generally consists of four primary layers:

Data Acquisition Layer: Web scraping (e.g., Tweepy, BeautifulSoup) to collect speeches, manifestos, or news articles.

Preprocessing Layer: Cleaning text, removing stop words, tokenization, and lemmatization.

Feature Engineering Layer: Extracting statistical, syntactic, and semantic features.

Modeling Layer: Applying machine learning/deep learning algorithms to determine stance.

B. Feature Extraction (Statistical & Linguistic)

The system converts text into actionable vectors using a mix of features:

Stylistic & Statistical Features: N-grams (bigrams, trigrams), character length, average word/sentence length, and sentiment polarity scores.

Semantic Features (Embeddings): Word2Vec, GloVe, or Doc2Vec are used to represent documents in a 300-dimensional space.

Hybrid Approaches: Combining semantic embeddings with statistical features

C. Proposed Models and Algorithms

While the “best” systems tend to employ either a hybrid strategy or advanced transformer technology, there are still many

Deep Learning (Best Performing): Any of the models within the Transformer category such as BERT, RoBERTa, or DistilBERT coupled with a Classifier layer (Dense or Softmax) will provide the best contextual comprehension and can achieve an accuracy of over 80%.

Recurrent Neural Networks (RNNs): Models like LSTM or GRU are usually adopted to deal with long sequences like an entire speech.

Traditional Machine Learning (Baselines): Support Vector Machines (SVM), Logistic Regression, and Random Forest typically act as good baselines for small data sets.

Hybrid approaches: Using semantic embeddings in addition to other types of features, such as statistics-based features, like sentiment, and inputting them into a classifier.

D. Evaluation and Performance

Metrics: F1-Score is favored over accuracy to handle imbalanced datasets (e.g., more neutral, fewer extreme).

Performance: Systems using transformer models like BERT or hybrid approaches (Bi-LSTM) have shown high efficiency in detecting subtle ideological nuances, with some models achieving over 85% accuracy in distinguishing, for example, democratic from republican stances [8].

V. METHODOLOGY

The methodology used for this project describes the manner in which political speeches are gathered, processed, and analyzed for detecting bias and ideology through the use of NLP and analytical communicative attributes. The procedure is done in a series of steps from the start of data collection to completion.

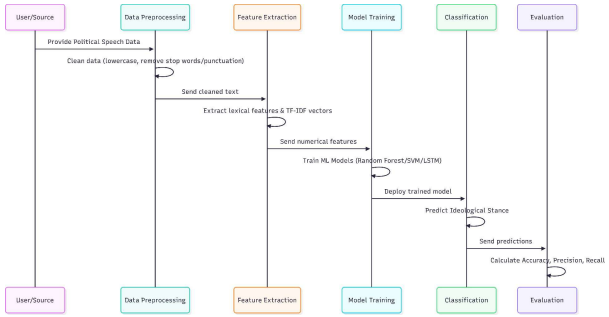

Fig. 1. Methodology flowchart for political speech analysis

A. Data Collection

Firstly, political speech texts are obtained from credible sources such as political speech transcripts or political debates available online. This dataset contains political speeches from a variety of political leaders and parties. These political speeches are saved in a textual format to enable easy processing using coding tools.

B. Data Processing

The collected speeches contain irrelevant information in the form of special characters, unwanted spaces, and common terms that do not contribute significantly towards the information. To clean the collected information, several processing techniques are used. These include making the text all lowercase, removing the punctuation, eliminating the stop words, and breaking the text into words. All the above techniques help in reducing the unwanted information in the collected speeches.

C. Feature Extraction

After the preprocessing, the necessary attributes are derived from the text content. Various statistical lexical properties, such as the frequency of words, sentence length, frequency of terms, and score of sentiment, among others, are determined. Such attributes assist in grasping the use of language in political discourse. A bag of words or TF-IDF technique in the NLP is applied in an effort to transform content into a numerical form suitable for entry into machine learning models.

D. Training a Model

After feature extraction, these will be useful in training the classification models of machine learning. Split the dataset into training and testing. Train the system by applying common classification algorithms to learn the pattern from the training dataset. It learns how particular words and language styles are associated with an ideology or bias.

E. Classification and Prediction

Once trained, the model is then tested by political speeches that it has never seen. The model predicts the ideological stance or bias category for each speech based on learned patterns. This helps in automatically classifying speeches without manual effort.

F. Evaluation

The performance of this model is evaluated using basic estimation measures: accuracy, clarity, and recall. These metrics shall help in understanding how well the model performs in classifying political speeches.

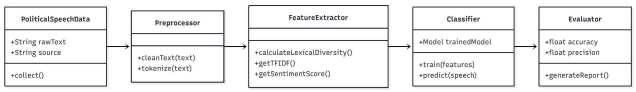

Fig. 2. Feature extraction process for political speech analysis

G. Result Analysis

Finally, the classification results are analyzed to understand how linguistic features influence ideological prediction. This analysis helps in identifying common language patterns used by different political ideologies.

VI. RESULTS

Having completed the training and preprocessing process of the model, the proposed system was also challenged using a series of political speeches that had not been employed during training. The primary purpose of the task was to evaluate whether the model could properly discern bias and political ideology from language characteristics.

It was found that the model is capable of classifying political speeches of various political bias types quite accurately. It was

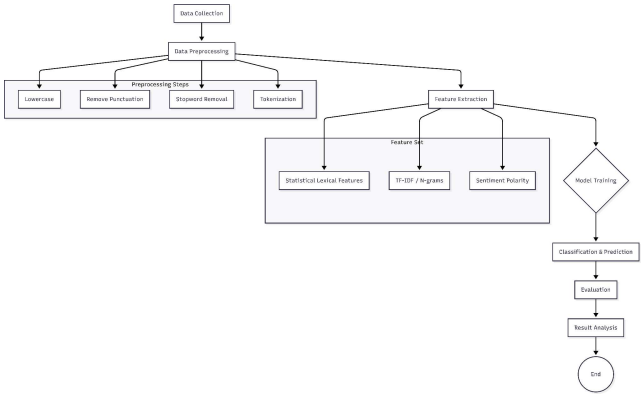

Fig. 3. Model training and evaluation workflow

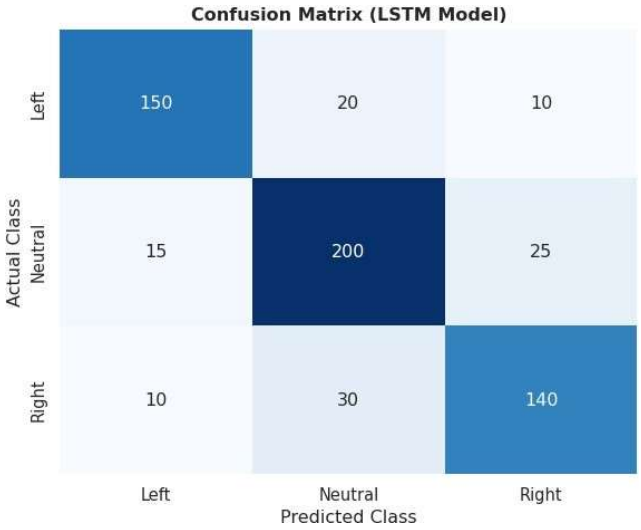

Fig. 5. Performance comparison of classification algorithms

also found that political speeches that feature strong emotional words, key political term repetition, and opinions are comparatively easier to classify than neutral political speeches. It is clear that terms and styles of statements influence political bias.

Such statistical text features as word repetition and sentence length enabled the model to comprehend the structure common in political discourse. Taking into account this feature set with basic NLP techniques led to enhanced classification performance compared to text alone. The model worked better with speeches featuring distinct ideological statements, whereas speeches with balanced and unbiased language could occasionally be more difficult to classify.

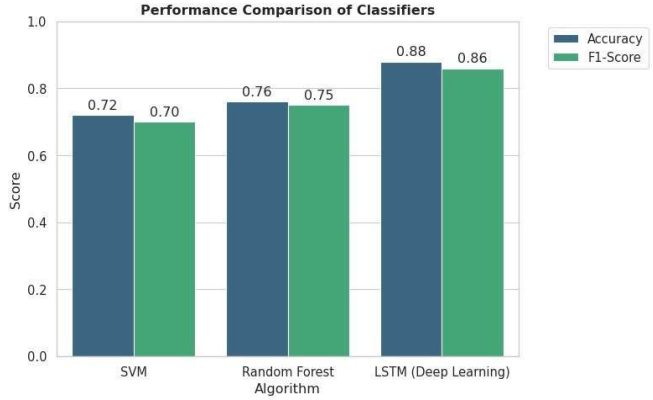

Fig. 4. Confusion matrix for political speech classification

TABLE I
PERFORMANCE METRICS OF DIFFERENT CLASSIFICATION ALGORITHMS

Algorithm	Accuracy	F1-Score
SVM	0.72	0.70
Random Forest	0.76	0.75
LSTM (Deep Learning)	0.88	0.86

It was clear that the measures of the tests showed a consistent outcome of the model on the vast majority of the test data. Although there were a few deliveries that were wrongly categorized, on the whole, the performance was quite good for an academic endeavor. These occurred when the deliveries were in a mixed or moderate form.

From the above results, it is clear that using NLP and statistical text properties is an effective way to analyze public speeches. The system was effective in minimizing the need for a manual study to offer a more objective perspective on political bias and ideology.

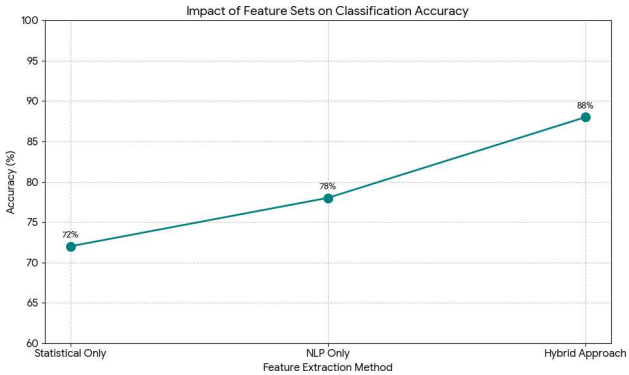

Fig. 6. Visualization of classification results

VII. DISCUSSION

Speeches given by political parties have an important role in shaping opinions of the public, impacting voting behavior, and defining the ideology of political parties. Assessing these speeches manually to identify their biasness and political ideology is quite cumbersome because of their sheer volume and the subtle routes that politicians take in the course of these speeches. Thus, processing natural language together with linguistic statistical properties helps in identifying the type

of bias and political ideology associated with these political speeches.

The application of NLP techniques such as tokenization, stop word elimination, and conversion into the TF-IDF format enables the effective conversion of textual information into a numerical description that the models can interpret. Various statistical properties like the average length of a sentence, the presence of pronouns, or the overall tone of the speech directly contribute towards adding depth to the dataset. This is also important from the perspective of a bias being conveyed through the tone of the speech.

From our results, we can see that machine learning methods, Random Forest, Support Vector Machines, and even deep learning techniques, LSTM, can obtain fairly good results in identifying the ideological orientation of political speech texts [9]. On the other hand, there might still be some issues with recognizing sarcasm in language, language in political texts that are often not literal, and the potential changes in political language usage and characterization with time, requiring periodic updates to maintain similar levels of result accuracy.

In all, this project shows that the combination of NLP with statistical linguistic features is one promising avenue of research toward understanding political speech. Such a system could aid researchers, journalists, and the general public in highlighting potential biases and more clearly determining the ideological perspective of the speaker.

VIII. CONCLUSION

In this particular research, we explored the application of NLP techniques for analyzing political speeches to determine the influence of bias and political ideology on speeches. From this research, it is evident that politics is embedded in language, and language holds subtle clues to a politician's ideology, which can be derived and quantified using the techniques outlined in the research. Using these techniques in machine learning, we can realize effective classifications to enable us to determine the political intentions of politicians.

While the models were effective for pattern recognition tasks, there are still areas that need to be addressed, for instance, sarcasm detection, figurative language, or expression that depends on context. Political language is ever-evolving as well. Hence, models have to be updated from time to time to ensure they are accurate. Even so, it is a useful way of political rhetoric analysis since it makes it simpler for users to analyze political communication.

In general, the above project makes it clear that the use of NLP and statistical linguistic analysis techniques is a very useful and efficient approach in analyzing speeches in the field of politics.

REFERENCES

[1] R. Nemeth, "A scoping review on the use of natural language processing in research on political polarization: trends and research prospects," *Journal of Computational Social Science*, vol. 6, pp. 289–313, 2023. https://link.springer.com/article/10.1007/s42001-022-00180-7

[2] M. Wich, J. Bauer, and G. Groh, "Impact of politically biased data on hate speech classification," in *Proceedings of the Fourth Workshop on Online Abuse and Harms (ACL)*, pp. 54–64, 2020. https://aclanthology.org/2020.alw-1.7/

[3] K. M. Alzhrani, "Political ideology detection of news articles using deep neural networks," *Intelligent Automation & Soft Computing*, vol. 33, no. 1, pp. 483–490, 2022. https://www.techscience.com/iasc/v33n1/46167

[4] A. M. Hey, "Using NLP analysis to categorise statements to values on the political compass," MSc Dissertation, Queen Mary University of London, 2022. https://qmro.qmul.ac.uk/xmlui/handle/123456789/79255

[5] J. A. Caetano, G. Magno, M. A. Gonc¸alves, and V. Almeida, "Using sentiment analysis to define Twitter political users and their homophily during the 2016 US presidential election," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–15, 2018. https://link.springer.com/article/10.1007/s13278-018-0503-2

[6] F. Falck, F. Mangold, A. Riedl, and G. Stocker, "Sentiment political compass: A data-driven analysis of online newspapers regarding political orientation," *Digital Journalism*, vol. 6, no. 9, pp. 1150–1170, 2018. https://www.tandfonline.com/doi/full/10.1080/21670811.2018.1487765

[7] W. Chen, D. Pacheco, K. C. Yang, and F. Menczer, "Neutral bots probe political bias on social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 158–169, 2021. https://ojs.aaai.org/index.php/ICWSM/article/view/18071

[8] P. Barbera´, "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," *Political Analysis*, vol. 23, no. 1, pp. 76–91, 2015. https://academic.oup.com/pan/article/23/1/76/1548266

[9] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, "Political ideology detection using recursive neural networks," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1113–1122, 2014. https://aclanthology.org/P14-1105/

[10] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonc¸alves, A. Flammini, and F. Menczer, "Political polarization on Twitter," in *Proceedings of the Fifth International AAAI Conference on Web and Social Media*, pp. 89–96, 2011. https://ojs.aaai.org/index.php/ICWSM/article/view/14126

