Pixel Phantoms at Touché: Ideology and Power Identification in Parliamentary Debates using Linear SVC

Notebook for the Touché Lab at CLEF 2024

Janani Hariharakrishnan, Jithu Morrison S and P Mirunalini

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract

The research addresses two fundamental tasks in parliamentary discourse analysis: the first one is identifying the political ideology of speakers, and the other one is categorizing their party affiliation as either governing or opposition. Analyzing parliamentary speeches involves discerning the ideological leanings of speakers' parties based on their articulated beliefs and determining whether a party is in a governing or opposition role within legislative contexts, which is crucial for understanding their political dynamics. These tasks are formulated as binary classification problems. Pixel Phantoms at Touché presents a framework for classifying speakers' political ideology and party affiliation in parliamentary debates by using a diverse set of models including Linear SVC, Logistic Regression, and DistilBERT. We proposed three different models, In the first model, DistilBERT was used for extracting the embeddings, and then Logistic Regression was applied for classification. In the other two models, Linear SVC and Logistic Regression models were built and trained for classification based on the extracted TF-IDF vectorization features. The study utilizes a multilingual corpus derived from ParlaMint, encompassing parliamentary speeches across various languages and jurisdictions. The proposed system was evaluated using F1 score and found Linear SVC emerging as the top-performing model. It achieved F1 scores of 0.5921 and 0.66 in the orientation and power test data, respectively, showcasing its robustness in handling complex political discourse data.

Keywords

Multilingual Text Classification, Grid Search, Neural Network Architectures, Natural Language Processing (NLP), TF-IDF (Term Frequency-Inverse Document Frequency)

1. Introduction

Understanding the intricate landscape of political discourse within national parliaments necessitates a thorough analysis of two critical variables: political ideology and party alignment. This research aims to elucidate the ideological leanings of parliamentary speakers and their affiliations with either governing or opposition parties by systematically analyzing parliamentary speeches. Through this study, the multifaceted nature of political communication within legislative bodies is explored, contributing to a deeper comprehension of the forces shaping public policy and governance.

Traditionally, the task of categorizing parliamentary speeches based on political ideology and party alignment has been approached through manual methods, relying on human expertise and subjective judgment. However, this manual process is not without its drawbacks. It is time-consuming, labor-intensive, and susceptible to biases inherent in human interpretation. Moreover, the sheer volume of parliamentary data poses significant challenges in maintaining consistency and accuracy across analyses. These limitations underscore the need for innovative automated solutions capable of handling the complexity and scale of political discourse data.

In recent years, advancements in machine learning and deep learning have revolutionized the field of natural language processing, offering promising avenues for automated analysis of textual

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

janani2210181@ssn.edu.in (J. Hariharakrishnan); jithumorrison2210564@ssn.edu.in (J. M. S); miruna@ssn.edu.in (P. Mirunalini)

data. Techniques such as neural networks, recurrent neural networks (RNNs), and transformers have demonstrated remarkable capabilities in tasks ranging from text classification to sentiment analysis. Leveraging these techniques, researchers have begun exploring automated approaches to political discourse analysis, aiming to overcome the limitations of traditional manual methods.

In this context, the research work represents a significant innovation in the field of political discourse analysis. By harnessing state-of-the-art deep learning models and leveraging advanced computational techniques, the team aims to develop a robust framework for classifying parliamentary speeches based on political ideology and party alignment. The approach seeks to address the shortcomings of traditional methods by providing a scalable, efficient, and objective means of analyzing political discourse. Through the innovative solution, the team endeavors to contribute to the advancement of computational methods in political science and facilitate deeper insights into the dynamics of parliamentary debates.

2. Background

The dataset [1] derived from ParlaMint, which comprises corpora of transcribed parliamentary speeches from 29 national and regional parliaments. Statistics on the dataset are presented, along with baseline results obtained using a simple classifier. These results pertain to predicting political orientation on the left-to-right axis and distinguishing between speeches delivered by governing coalition party members and those from opposition party members. In [2], authors analysis neoliberal, traditionalist, and social justice perspectives, provided crucial insights for the power identification task. The research work in [3] focuses on linguistic hierarchies and the marginalization of non-official languages informed the approach to detecting power dynamics in parliamentary debates. The study also aided in developing a more nuanced understanding of how linguistic diversity impacts political discourse.

In [4], provides crucial insights which enhances feature engineering and data annotation for accurate classification of political ideology and party affiliation. In [5], Wang and Banko describe practical methods for implementing and optimizing transformer models to handle text classification tasks across multiple languages. In [6], authors reviewed text vectorization methods for sentiment classification, covering Tf-Idf, Word2vec, and Doc2vec and the study emphasizes the significance of selecting suitable vectorization techniques tailored to the characteristics and objectives of sentiment analysis tasks

The research work [7] provides a review of political sentiment analysis techniques for tweets, covering lexicon-based and machine learning methods. In [8], the research offers insights of innovative methodologies in enhancing information access and retrieval across a range of linguistic and contextual settings, including experimental approaches that integrate multilinguality, multimodality, and interaction within information retrieval. In [9], Owais Ahmad et al. hybridize Sentence Transformer embeddings with Multi KNN to enhance the accuracy and efficiency of multi-label text classification tasks, specifically in biomedical literature analysis. In [10], Varun Dogra et al. analyze DistilBERT for sentiment classification of banking financial news. The research provides insights into leveraging modern NLP models techniques for accurate sentiment analysis.

In [11], Hardik Jadia conducts a comparative analysis of sentiment analysis techniques, including Support Vector Machines (SVM), Logistic Regression, and TF-IDF feature extraction. This research provides insights into the effectiveness of these methods for text classification tasks. The findings offer guidance for optimizing models to accurately identify political ideologies and party affiliations in parliamentary speeches, contributing to more reliable computational political analysis. In [12], the comparison of supervised classification models on textual data highlights the effectiveness of different algorithms for text classification tasks. This research work underscores the importance of selecting and optimizing classification models for accurately identifying political ideologies and party affiliations in parliamentary speeches, thereby improving the overall performance and reliability of the analysis.

The literature review emphasizes the application of advanced NLP techniques and computational methods to enhance the accuracy of identifying political ideologies and party affiliations in parliamentary discourse. Insights from referenced studies inform the integration of transformer models like DistilBERT, logistic regression, and LinearSVC, offering a structured approach for computational political analysis.

3. System Overview

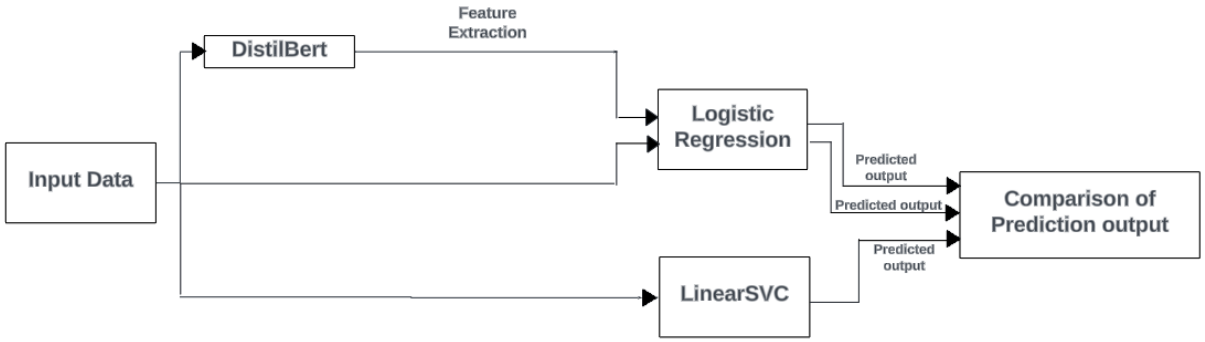

Figure 1: Overall Flow Diagram of the Proposed Model

This research work is structured around two primary objectives within parliamentary discourse analysis: distinguishing speakers' political ideologies and classifying their party affiliations into governing or opposition categories. To achieve these goals, separate binary classification models were employed for both orientation (political ideology) and power classification. Models such as Linear SVC and Logistic Regression were utilized for both tasks, alongside a DistilBERT-based model for generating text embeddings, followed by Logistic Regression for prediction.

In this task, all training was conducted on the English translations of the parliamentary speeches. The dataset includes both the original texts and their English translations (where available), and we focused on the English texts to maintain consistency and leverage the capabilities of pre-trained language models like DistilBERT, which are primarily trained on English data. By doing so, it is ensured that the feature extraction using TF-IDF and sentence embeddings was uniform across the dataset, allowing for effective training and evaluation of our models.

Separate models were developed for both distinguishing political orientation and categorizing power dynamics within parliamentary discourse using approaches like Linear SVC, Logistic Regression, and DistilBERT combined with Logistic Regression.

3.1. Dataset

The dataset for this task is derived from ParlaMint, a multilingual comparable corpus of parliamentary debates. The data is provided in tab-separated text files, containing the following fields: id, speaker, sex, text, text_en, and label. Each entry represents a speech, with information on the speaker and automatic translations where available. The dataset covers parliamentary speeches from various national and regional parliaments, and is designed to minimize confounding variables like speaker identity.

3.2. Data Preprocessing

During the data preprocessing phase, missing values were removed and the columns with null values were replaced with empty strings. TF-IDF features were extracted from the

English text using TfidfVectorizer with specified parameters, and were used in Linear SVC and Logistic Regression. The text data (X_train and X_val) was transformed into sentence embeddings using the pre-trained DistilBERT model from the Sentence Transformers library (`SentenceTransformer('distilbert-base-nli-mean-tokens')`). This step converted the text data into numerical vectors suitable for machine learning algorithms, which were used in DistilBERT.

3.3. Feature Extraction using DistilBERT for Text Classification

DistilBERT excels in capturing semantic meaning and relationships within sentences, making it highly versatile for various natural language processing tasks like sentiment analysis, question answering, document classification, and named entity recognition. Its embedding process converts text into numerical representations that encode deep contextual information, enabling precise model predictions across nuanced linguistic contexts. This capability to extract meaningful embeddings from input text is essential for a wide array of NLP applications.

After preprocessing, the model initiates by utilizing the "distilbert-base-nli-mean-tokens" pre-trained transformer from the Sentence Transformers library to convert parliamentary speeches into dense semantic embeddings. These embeddings serve as high-dimensional features fed directly into a Logistic Regression classifier.

Following embedding generation, the Logistic Regression classifier is directly trained on these embeddings. This training phase involves optimizing the Logistic Regression's parameters, including regularization strength ('C') and penalties ('l1' and 'l2'), using GridSearchCV. This hyperparameter tuning ensures that the model maximizes its predictive accuracy based on the embeddings' semantic content.This methodology guarantees that the model effectively leverages the semantic information distilled by DistilBERT to make precise predictions regarding the political ideologies and party affiliations of speakers based solely on their parliamentary speeches.

The adoption of an alternative model from considerations of computational complexity and resource requirements, particularly given the size of the dataset and the robust nature of the classification task. Moreover, the interpretability offered by Logistic Regression's feature coefficients seemed more congruent.

3.4. Text Classification Using Logistic Regression

Logistic Regression is fundamental in supervised learning for its interpretability and efficacy across various domains, prominently in natural language processing (NLP). In tasks like sentiment analysis and political discourse classification, Logistic Regression harnesses TF-IDF vectorization to convert text into numerical features. This process optimizes parameters such as the maximum feature count (10,000), inclusion of both unigrams and bigrams (n-grams of length 1 and 2), and thresholds for document frequencies (min_df=2 and max_df=0.9). Additionally, common English stop words are removed during the vectorization process.

GridSearchCV refines these parameters to maximize predictive accuracy, ensuring the model performs optimally. Post-vectorization, Logistic Regression incorporates these optimized features, adjusting class weights to manage dataset imbalances and ensure balanced learning outcomes. Utilizing a fixed random state guarantees consistent results across different runs. Leveraging the Scikit-learn library, these methodologies provide robust and scalable solutions for diverse NLP applications. This integrated approach underscores Logistic Regression's versatility and pivotal role in converting raw textual data into actionable insights through parameter optimization and effective model training.

3.5. Text Classification Using LinearSVC

LinearSVC is renowned for its efficiency and effectiveness in text classification tasks, particularly suited for high-dimensional data and widely applied in areas such as spam detection and sentiment analysis. Like Logistic Regression, LinearSVC starts with TF-IDF vectorization to convert textual data into numerical features. It employs similar TF-IDF settings, including the limitation of features, exclusion of rare terms, and removal of common terms and stop words to enhance the quality of the feature set.

Despite using similar TF-IDF settings, there is a distinction between Logistic Regression, which predicts class probabilities using a logistic function, and LinearSVC, which separates classes with a hyperplane based on maximum margin criteria, emphasizing clear class distinction rather than probability estimation.

During model training, LinearSVC adjusts class weights to manage imbalanced data distributions effectively. Hyperparameter optimization via GridSearchCV focuses on parameters like the regularization parameter C (tuned across values of [0.1, 1]) and maximum iterations (explored with [1000, 1700, 2500]), aiming to enhance model performance without overstating capabilities. Evaluation metrics, including the F1 score, provide rigorous assessment of the model's classification accuracy, ensuring reliable performance validation in practical applications.

4. Results

We have experimented with three different models and performance of each model was analysed using different measures such as F1 score, Recall and Precision. The results were tabulated in the Table 1 and Table 2.

Table 1
Performance Comparison - Orientation

Model	F1 Score	Recall	Precision
Linear SVC	0.5921	0.599	0.606
Logistic Regression	0.5926	0.6	0.592
DistilBERT	0.563	0.535	0.504
Baseline	0.560	0.592	0.571

Table 2
Performance Comparison - Power

Model	F1 Score	Recall	Precision
Linear SVC	0.66	0.658	0.666
Logistic Regression	0.657	0.658	0.66
DistilBERT	0.453	0.477	0.394
Baseline	0.640	0.650	0.632

Comparing the performance of three distinct models—Linear SVC, Logistic Regression, and DistilBERT approach—reveals their respective strengths and nuances in handling the tasks of identifying political ideology and categorizing party affiliation in parliamentary debates. These models were evaluated for both tasks, assessing their capabilities and limitations in computational political analysis.

Table 3
Top 10 prediction outputs based on F1 Score for Orientation

Country	Orientation	Baseline
Turkey (tr)	0.783	0.841
Galicia (es-ga)	0.754	0.667
Spain (es)	0.723	0.717
Great Britain (gb)	0.719	0.745
Greece (gr)	0.713	0.742
Hungary (hu)	0.684	0.570
Catalonia (es-ct)	0.643	0.655
Sweden (se)	0.608	0.550
The Netherlands (nl)	0.599	0.513
Poland (pl)	0.588	0.460

Table 4
Top 10 prediction outputs based on F1 Score for Power

Country	Power	Baseline
Turkey (tr)	0.801	0.831
Hungary (hu)	0.796	0.858
Galicia (es-ga)	0.768	0.829
Great Britain (gb)	0.726	0.711
Greece (gr)	0.724	0.627
Austria (at)	0.698	0.660
Poland (pl)	0.697	0.758
Basque Country (es-pv)	0.692	0.715
Denmark (dk)	0.692	0.561
Serbia (rs)	0.688	0.647

Table 3 shows scores for the linear SVC model and Baseline across ten countries for Orientation. Table 4 focuses on power performance, presenting linear SVC and Baseline scores for ten countries. These tables highlight each country's relative performance in specific dimensions, offering a clear comparative overview.

The Linear SVC model achieves the highest training and testing F1 scores of approximately 0.74 and 0.66, respectively, for categorizing party affiliation (Power) and 0.77 and 0.5921 for identifying political ideology (Orientation). This showcases its robust capability in accurately classifying whether speakers' parties are governing or in opposition and their political beliefs. Known for its effectiveness in handling linearly separable data, Linear SVC constructs hyperplanes to delineate data points, ensuring high precision and recall in this context.

The Logistic Regression model achieves the training and testing F1 scores of approximately 0.73 and 0.657, respectively, for categorizing party affiliation (Power) and 0.77 and 0.5926 for identifying political ideology (Orientation). Enhanced by TfidfVectorizer for feature extraction and optimized hyperparameters, Logistic Regression effectively manages collinear features and provides precise class probability estimates. This probabilistic framework makes it adept at scenarios where the relationship between features and target variables is probabilistic rather than strictly linear.

The DistilBERT model, achieves the training and testing F1 scores of approximately 0.74 and 0.453, respectively, for categorizing party affiliation (Power) and 0.74 and 0.563 for identifying political ideology (Orientation). DistilBERT's strength lies in its ability to encode complex, non-linear relationships within the feature space, leveraging its capability as a powerful feature extraction tool.

DistilBERT is employed primarily for feature extraction, harnessing its capacity to uncover intricate data patterns relevant to political discourse analysis.

Overall, while Linear SVC demonstrates superior performance in this study for the task of categorizing party affiliation and Logistic Regression for identifying political ideology, each model presents trade-offs between complexity, interpretability, and performance. These insights underscore the nuanced considerations necessary when applying machine learning to computational political analysis.

5. Conclusion

In conclusion, the Linear SVC model exhibits the highest validation F1 score, showcasing its robustness in handling linearly separable data. Logistic Regression closely follows, demonstrating competitive performance by efficiently managing collinear features and providing precise class probability estimates. Meanwhile, the DistilBERT model offers flexibility in capturing complex relationships but may be prone to overfitting.

References

[1] Ç. Çöltekin, M. Kopp, K. Meden, V. Morkevicius, N. Ljubešić, T. Erjavec, Multilingual power and ideology identification in the parliament: a reference dataset and simple baselines, arXiv preprint arXiv:2405.07363 (2024).

[2] A. Becker, Identity, power, and prestige in Switzerland's multilingual education, transcript Verlag, 2023.

[3] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, A. Pančur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrímsson, Çöltekin, J. de Does, K. Depuydt, T. Agnoloni, G. Venturi, M. Calzada Pérez, L. D. de Macedo, C. Navarretta, G. Luxardo, M. Coole, P. Rayson, V. Morkevičius, T. Krilavičius, R. Darǵis, O. Ring, R. van Heusden, M. Marx, D. Fišer, The parlamint corpora of parliamentary proceedings, Language resources and evaluation 57 (2022) 415–448. doi:10.1007/s10579-021-09574-0.

[4] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, et al., Refining the Theory of Basic Individual Values, Journal of personality and social psychology 103 (2012). doi:10.1037/a0029393.

[5] C. Wang, M. Banko, Practical transformer-based multilingual text classification, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021, pp. 121–129.

[6] H. D. Abubakar, M. Umar, M. A. Bakale, Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec, SLU Journal of Science and Technology 4 (2022) 27–33.

[7] A. Britzolakis, H. Kondylakis, N. Papadakis, A review on lexicon-based and machine learning political sentiment analysis using tweets, International Journal of Semantic Computing 14 (2020) 517–563.

[8] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Munstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[9] O. Ahmad, S. Verma, S. Azim, A. Sharan, Hybridizing sentence transformer embeddings with multi knn for improved multi-label text classification for pubmed (2023).

[10] V. Dogra, A. Singh, S. Verma, Kavita, N. Jhanjhi, M. Talib, Analyzing distilbert for sentiment classification of banking financial news, in: Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021, Springer, 2021, pp. 501–510.
[11] H. Jadia, Comparative analysis of sentiment analysis techniques: Svm, logistic regression, and tf-idf feature extraction (2023).
[12] B.-M. Hsu, Comparison of supervised classification models on textual data, Mathematics 8 (2020) 851.