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, P Mirunalini

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

CLEF 2024, Grenoble, France

Introduction

The problem statement addresses the need for scalable, unbiased analysis of political ideology and party affiliation in parliamentary speeches, which traditional methods fail to achieve efficiently.
Traditional methods are manual, biased, and time-consuming.
The increasing volume of speeches necessitates consistent and accurate analysis.
Automated analysis using machine learning and NLP can:

- Enhance scalability and efficiency in processing large datasets.
- Reduce human bias by providing objective, data-driven insights.
- Provides clearer insights into parliamentary dynamics.

This innovation addresses the limitations of manual methods, enabling more reliable and comprehensive political discourse analysis.

Literature Survey

Key references that shaped our research and methodology:

- **Coltekin et al. (2024)**: Developed the multilingual ParlaMint corpus for parliamentary debates, crucial for cross-linguistic political discourse analysis.
- **Becker et al. (2023)**: Explored power dynamics in political debates, identifying linguistic markers that differentiate between government and opposition speakers. This influenced our approach to speaker classification.
- **Schwartz et al. (2012)**: Pioneered feature engineering for political ideology classification using lexical and syntactic features, guiding our feature extraction techniques like TF-IDF.
- **Jadia (2023)**: Compared classification models for text data, influencing our use of Linear SVC and transformer-based embeddings such as DistilBERT for ideology and sentiment classification.

Proposed Solution

Our solution for analyzing parliamentary discourse focuses on two tasks: identifying the speaker's political ideology (Left vs Right) and classifying them as governing or opposition. We implemented three models:

- DistilBERT + Logistic Regression: Extracts contextual embeddings from speeches and uses Logistic Regression for classification.
- Linear SVC: A powerful text classifier that directly predicts the categories, showing strong performance in high-dimensional data.
- Logistic Regression: A baseline model to compare against more complex approaches.

Using the ParlaMint corpus, the Linear SVC model achieved the highest F1 scores of 0.5921 for ideology and 0.66 for power classification. This project demonstrates the potential of machine learning to streamline and enhance the analysis of political discourse in multilingual parliamentary data.

Proposed System - Overview

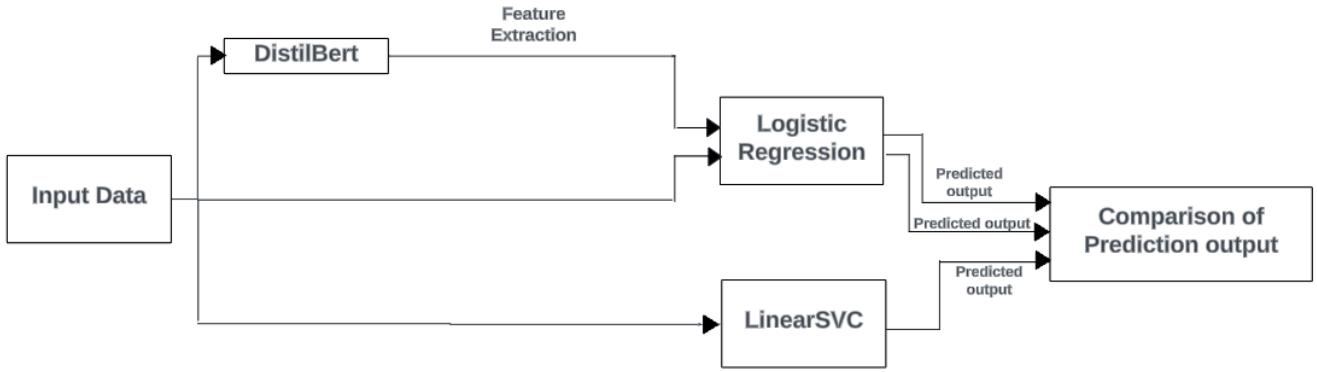

Figure: Flow Diagram of the Proposed Model

The workflow for classifying political discourse processes parliamentary speeches through three paths: DistilBERT with Logistic Regression, Linear SVC, and Logistic Regression on raw data. The outputs from these methods are then compared to determine the best classification approach.

Experimental Setup

Dataset: ParlaMint - Multilingual parliamentary debates corpus.
Data Preprocessing:

- Missing values handled
- TF-IDF features used for Linear SVC and Logistic Regression
- DistilBERT embeddings for Logistic Regression

Model Training:

- Linear SVC and Logistic Regression trained on TF-IDF features
- DistilBERT embeddings used in the Logistic Regression model

Results - Ideology

The Linear SVC and Logistic Regression models performed similarly, with both achieving an F1 score around 0.59, indicating a good balance between precision and recall. DistilBERT underperformed with a lower F1 score of 0.563, which suggests that traditional machine learning models may be better suited for this specific task.

Table: Performance Comparison - Ideology

Model	**F1 Score**	**Recall**	**Precision**
Linear SVC	0.5921	0.599	0.606
Logistic Regression	0.5926	0.600	0.592
DistilBERT	0.563	0.535	0.504

Results - Power

Linear SVC outperformed the other models with an F1 score of 0.66, showing that it is well-suited for identifying power dynamics in political speeches. Logistic Regression yielded similar results, making it a reliable alternative.

Table: 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

Conclusion

Linear SVC emerged as the top-performing model for both tasks:

- Ideology classification (F1 Score: 0.5921)
- Power classification (F1 Score: 0.66)

This system provides a scalable, efficient solution for classifying political discourse, offering insights into the ideological and power dynamics within parliamentary debates.