JU_NLP_DID at Touché: An Attempt to Identify Aspects of Power from Parliamentary Debates

Notebook for the Touché Lab at CLEF 2024

Adnan Khurshid, Dipankar Das, Rajdeep Khaskel and Suchanda Datta

[1]*Department of Computer Science and Engineering, Jadavpur University, Kolkata, 700032, India*

Abstract

Parliamentary debates shape critical aspects of citizens' lives and often influence global policies. Analyzing these debates computationally poses unique challenges due to the indirect and complex nature of political discourse. This paper addresses two key variables in parliamentary speeches: the political ideology of the speaker and their affiliation with either the governing party or the opposition. We approach these subtasks as binary classification problems, employing a combination of Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and Support Vector Machines (SVM) for our analysis. Our methodology is designed to capture the nuanced language of parliamentary debates and effectively classify speakers based on their political stance and party alignment. The results demonstrate the efficacy of TF-IDF with SVM in handling the intricacies of political speech, providing a robust framework for further research in computational political analysis.

Keywords

TF-IDF, SVM, Binary Classification

1. Introduction

Parliamentary debates are pivotal in shaping not only the national policies of a country but also influencing global political landscapes. These debates, characterized by their indirect and nuanced discourse, pose significant challenges for computational analysis. Understanding the ideological stance and power alignment of speakers within these debates can provide valuable insights into political dynamics and decision-making processes.

1.1. Objective

This paper addresses the task[1] of classifying two critical variables associated with speakers in parlia- mentary debates: the political ideology of the speaker's party and whether the speaker's party is in the governing coalition or in opposition. These tasks are formulated as binary classification problems, necessitating sophisticated natural language processing (NLP) techniques to handle the complexity and variability of political speech.

The data for this study is derived from the ParlaMint corpus, a multilingual and comparable dataset of parliamentary debates across various countries. The corpus has been curated to minimize confounding variables, ensuring that the analysis focuses on the content and context of the speeches rather than extraneous factors such as speaker identity. The provided data includes both the original speeches and their English translations, facilitating the development of multilingual models and cross-linguistic analyses.

To tackle the classification tasks, we employ a combination of Term Frequency-Inverse Document Fre- quency (TF-IDF) vectorization and Support Vector Machines (SVM). TF-IDF is utilized to convert textual data into numerical representations that capture the importance of words within the speeches, while SVM is used for its effectiveness in handling high-dimensional feature spaces and binary classification problems.

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

adnankhurshid251@gmail.com (A. Khurshid); dipankar.dipnil2005@gmail.com (D. Das); deepkhaskel@gmail.com (R. Khaskel); sumidatta769@gmail.com (S. Datta)

1.2. Contribution

This study contributes to the field of computational political analysis by offering a robust framework for identifying political ideology and power alignment in parliamentary debates. By leveraging advanced NLP techniques, we aim to enhance the understanding of political discourse and provide a foundation for further research in this area. The results of our analysis demonstrate the potential of TF-IDF and SVM in addressing the challenges posed by the indirect nature of parliamentary speech, paving the way for more accurate and insightful political analysis tools.

2. Background

In today's digital era, parliamentary debates have transcended the confines of legislative chambers to include online platforms, fundamentally reshaping discourse. Within these digital spaces, social networks wield significant influence over opinions and provide valuable data for sentiment analysis and power identification. Understanding the intricate power dynamics at play is essential for decoding how influence is disseminated and policies are formulated. By harnessing the capabilities of sentiment analysis and computational techniques, we can shed light on the underlying power structures and sentiment trends, ultimately enhancing decision-making processes. This multifaceted approach involves analyzing various factors such as speaking patterns, party contributions, responses to arguments, social network connections, and media coverage to unveil influential actors and dominant dynamics within parliamentary debates.

3. System Overview

The system developed for parliamentary power identification involves several steps, including data preprocessing, feature extraction, and classification. The primary goal is to classify parliamentary text data using machine learning techniques. Below is a detailed overview of the system components and processes.

3.1. Data Preprocessing

We employed automated English translations for our experiments. The raw textual data underwent rigorous preprocessing to facilitate feature extraction and classification. This preprocessing pipeline encompassed the following steps:

Lowercasing: All text is converted to lowercase to ensure uniformity. HTML Tag Removal: HTML tags are removed using regular expressions to clean the text. The English translated text provided in the dataset contained HTML commands, necessitating their removal to ensure accurate preprocessing and avoid interference with subsequent analyses. Punctuation Removal: All punctuation marks are removed to reduce noise. Stopword Removal: Common stop words are removed using NLTK's stopword list, which helps in focusing on the meaningful words in the text. Lemmatization: Words are lemmatized to their base or dictionary form using the WordNet lemmatizer. This involves: Tokenizing the text. Tagging each word with its part of speech. Mapping the POS tag to WordNet's POS tag format. Lemmatizing each word based on its POS tag. This preprocessing ensures that the text data is clean, normalized, and stripped of irrelevant parts, making it suitable for feature extraction.

3.2. Feature Extraction

After preprocessing, the text data is transformed into numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique. This method converts the text into a matrix of TF-IDF features, which reflects the importance of words in the corpus:

TF-IDF Vectorization: This technique is used to convert the preprocessed text data into numerical vectors. It captures the importance of a word in a document relative to the entire corpus. The Tfidf Vectorizer from scikit-learn is used with default parameters.

3.3. Model Building and Training

For the classification task, a Support Vector Machine (SVM) model with a linear kernel is initially employed. The SVM classifier is chosen for its effectiveness in high-dimensional spaces and its capability to handle large feature sets resulting from TF-IDF vectorization. The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing. The SVM model is trained on the TF-IDF vectors of the training set.

Initially, we had tried using bi-grams and n-grams with SVM but did not observe relevant improvements in performance, hence we focused solely on uni-gram TF-IDF representations. To optimize the SVM model, hyper-parameter tuning is performed using RandomizedSearchCV. This approach is selected over GridSearchCV due to its ability to efficiently explore a wide range of parameter combinations with fewer iterations, thus reducing computational burden while still providing robust parameter estimates. Given our system's limited computational power, RandomizedSearchCV is configured with 5 iterations and 2-fold cross-validation.

Hyper-parameter tuning is performed using RandomizedSearchCV with the following parameter distribution:

- **C**: [0.1, 1, 10, 100, 1000]
- **kernel**: ['linear', 'rbf', 'sigmoid']
- **probability**: [True]
- **gamma**: ['scale', 'auto']
- **coef0**: [0.0, 0.1, 0.5, 1.0]

After tuning, the best parameters for the SVM model are found to be:

- **C**: 10
- **kernel**: 'rbf'
- **probability**: True
- **gamma**: 'scale'
- **coef0**: 0.1

These parameters enhance the SVM model's performance significantly for our classification task.

3.4. Model Evaluation

The trained SVM model is evaluated using the test set. Several evaluation metrics are computed to assess the model's performance:

Classification Report: This includes precision, recall, and F1-score for each class, providing a detailed performance analysis. Confusion Matrix: The confusion matrix shows the true positive, true negative, false positive, and false negative counts, offering insights into the model's prediction errors. ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) score are computed to evaluate the model's ability to discriminate between classes. The AUC score provides a single metric to summarize the model's performance.

4. Results

	Power	Orientation
Average F1	**0.629126846**	**0.570052236**
Max Precision	**0.832305837**	**0.876411242**
Max Recall	**0.824644263**	**0.763864404**
Max F1	**0.82763257**	**0.765286062**

Table 1
Average F1, Max Precision, Max Recall, Max F1 Scores of Ideology and Power Identification Shared Task.

The accuracy of the SVM classifier was in the range of 60-80 percent for different languages.

4. Conclusion

This study demonstrates the application of advanced natural language processing techniques to the analysis of parliamentary debates, focusing on identifying the political ideology of speakers and their party's power status. By leveraging the ParlaMint corpus, which provides a rich and multilingual dataset of parliamentary speeches, we have developed a robust framework for addressing these binary classification tasks.

Our approach involved experimenting with various methodologies. We initially utilized Term Frequency-Inverse Document Frequency (TF-IDF) vectorization combined with Support Vector Machines (SVM), which proved effective in handling the complexity and nuance of political discourse. The results highlight the capability of TF-IDF and SVM to capture significant features of parliamentary speeches. Each method demonstrated unique strengths, contributing to a comprehensive understanding of the political dynamics within parliamentary debates. This work not only provides valuable insights into the political dynamics within parliamentary debates but also sets the stage for further research in computational political analysis. Future studies can build on this foundation by refining these models, exploring ensemble methods, and expanding the scope to include additional political variables and more diverse datasets.

In conclusion, our study underscores the importance of computational approaches in understanding political discourse and offers a promising methodology for analyzing parliamentary debates. The techniques and findings presented here contribute to the broader field of political text analysis, enhancing our ability to decipher and interpret the intricate language of politics.

References

1. Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf et al. "The ParlaMint corpora of parliamentary proceedings." *Language resources and evaluation* 57, no. 1 (2023): 415-448.

2. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

3. Çöltekin, Çağrı, Matyáš Kopp, Katja Meden, Vaidas Morkevicius, Nikola Ljubešić, and Tomaž Erjavec. "Multilingual Power and Ideology Identification in the Parliament: a Reference Dataset and Simple Baselines." *arXiv preprint arXiv:2405.07363* (2024).

4. Russo, Daniel, Salud María Jiménez-Zafra, José Antonio García-Díaz, Tommaso Caselli, Marco Guerini, L. Alfonso Ureña-López, and Rafael Valencia-García. "PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task." (2023).

5. Tarkka, Otto, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo, and Veronika Laippala. "Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4." In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*, pp. 70-76. 2024.

6. Mochtak, Michal, Peter Rupnik, and Nikola Ljubešić. "The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings." *arXiv preprint arXiv:2309.09783* (2023).

7. Eskişar, Gül M. Kurtoğlu, and Çağrı Çöltekin. "Emotions running high? a synopsis of the state of turkish politics through the parlamint corpus." In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pp. 61-70. 2022.

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 Confer- ence of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.