

# Vayam Solve Kurmaha at Touché: Power Identification in Parliamentary Speeches Using TF-IDF Vectorizer and SVM Classifier

Working Notes Paper Touché Lab at CLEF 2024

Lakshmi Priya S<sup>1</sup>, Dhannya S M<sup>1</sup>, S. Shwetha<sup>1</sup>, Surabhi Kamath<sup>1</sup>, Shreedevi Seluka Balaji<sup>1</sup>, Sai Nikitha N.S.R<sup>1,\*</sup> and Srinidhi Lakshmi Narayanan<sup>1</sup>

<sup>1</sup>*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai, Tamil Nadu, 603110, India*

## Abstract

Political parties' viewpoints, goals, and policy philosophies are often made clear through parliamentary debates, which have a significant impact on national decision-making processes. Gaining public understanding of these discussions is essential for understanding political efficacy. However, because political statements are inherently ambiguous and strategically indirect, algorithmic analysis of them is challenging. By taking part in the Touché 2024 assignment on Ideology and Power Identification in Parliamentary Debates, this study attempts to address these issues. In this paper we compare traditional classification models namely Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and an ensemble of the three on features extracted using TF-IDF. We found the SVM outperformed the other models and achieved an F1 score of 0.68.

## Keywords

power, parliament, speeches, SVM, TF-IDF, binary classification, political, machine learning

## 1. Introduction

Parliamentary debates are rich repositories for uncovering the outlook of political parties, their motives, and their approach towards the welfare and future of the country. Discussions in parliament have the potential to shape the entire trajectory of a nation, as most decisions of paramount importance originate here. Understanding these discussions is vital, as it allows the public to truly comprehend political parties and evaluate their efficiency in making decisions on their behalf.

Political speeches, however, are elusive to computational analysis of their meanings. A paper published in 1977 [1] theorized that politicians are often strategically indirect to advance their career and gain an edge over their opponents. In another study [2] on vagueness in political language, the author states that political language is kept vague to address different audiences simultaneously and to avoid facing threats. Vagueness and indirectness both make a text challenging to analyze.

The task of power identification given by Touché, Ideology and Power Identification in Parliamentary Debates 2024 [3] aims to identify if a speaker of a given text in a parliamentary debate belongs to the coalition party or the opposition party.

## 2. Background

A paper [4] on political speech analysis, published in 2020, introduced a Graph Political Sentiment analyzer (GPolS), a neural model for speech-level stance analysis of members of parliament (MPs). The model utilizes a fine-tuned BERT for encoding the data and Graph Attention Networks (GAT) for modeling and aggregating contextual relations between transcripts, motions, and speakers. GPolS

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

\*Corresponding author.

 lakshmipriyas@ssn.edu.in (L. P. S); dhannyasm@ssn.edu.in (D. S. M); shwetha2210210@ssn.edu.in (S. Shwetha); surabhi2210196@ssn.edu.in (S. Kamath); shreedevi2210389@ssn.edu.in (S. S. Balaji); sainikitha2210401@ssn.edu.in (S. N. N.S.R); srinidhi2210142@ssn.edu.in (S. L. Narayanan)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

outperforms all baselines significantly under the Wilcoxon signed-rank test, by a large margin greater than 6.5%.

A study [5] aimed to predict the party group from Lithuanian parliamentary speeches found that at the dataset level, removing out-of-domain and irrelevant instances was the best preprocessing technique. Similarly, at the document level, removing digits and using a bag-of-words approach and token bigrams were most effective. The highest accuracy achieved was 0.545 compared to 0.279 and 0.13 on random and majority baselines.

A different study [6], published in 2004, on identifying agreement and disagreement in conversational speech proposed a statistical model that utilizes Bayesian networks to capture pragmatic dependencies and employs maximum entropy ranking to identify adjacency pairs based on lexical, durational, and structural features. The model was shown to achieve high accuracy.

In 2023, Kavallos and Christos-Sotirios conducted a research [7] aimed to assess the feasibility of classifying Greek parliament proceedings for their respective political parties using Multinomial Naïve Bayes classification, Stochastic Gradient Descent classification, Random Forest classification, and Recurrent Neural Network classification. They recorded that Random Forest algorithm performed the best followed by Recurrent Neural Network classification.

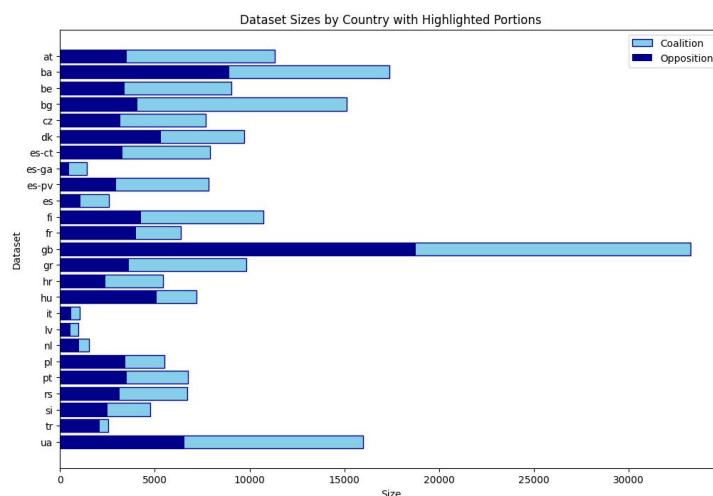
A thesis [8] exploring sentiment analysis of political debates introduces the Debate Graph Extraction (DGE) framework. This framework represents debates as graphs with speakers as nodes and exchanges as links, labeled based on sentiment ("supporting" or "opposing"). It also discusses analyzing these graphs using network mathematics and community detection to understand debate patterns.

### 3. System Overview

In this task, the English translations of the texts for the model that were provided in the dataset were used. The method used involved augmenting the texts using synonym replacement, extracting its features using TF-IDF vectorizer and applying them to classifier models, namely Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and an ensemble of the three.

#### 3.1. Data Augmentation

Data augmentation is the process of artificially increasing the size of the dataset by creating modified versions of existing data. It can also be used to balance datasets by increasing the size of the minority class, thereby addressing class imbalances. Imbalanced datasets can cause the classification model to be biased towards the majority class.



**Figure 1:** Dataset

During the data exploration phase, we found that most datasets were imbalanced and some by large margins. Figure 1 graphically represents the number of entries for labels 'coalition' and 'opposition'. The graph clearly depicts the data imbalances in many datasets. Therefore we used data augmentation to balance the datasets. The particular technique that we employed was synonym replacement. Synonym replacement randomly selects words in a text and replaces them with their synonyms to generate new data.

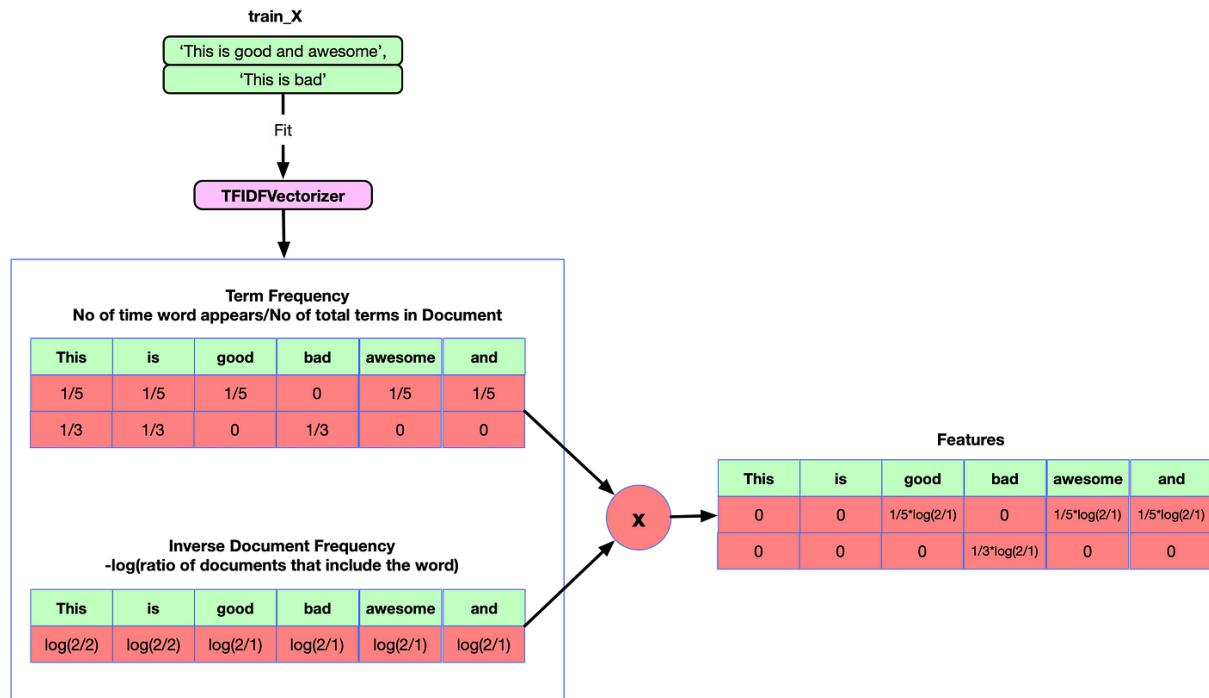
### 3.2. Feature Extraction

Feature extraction is the process of transforming raw data into numerical features that can be processed by machine learning models. The aim of the features is to represent the information contained in the original data in a format that can be efficiently utilised by the machine learning model. We used the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer for feature extraction due to its efficiency when working with large corpora.

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that indicates how important a word is to a document within a collection or corpus of unstructured text data. It scores a word by multiplying the word's Term Frequency (TF) by the Inverse Document Frequency (IDF). The higher the TF-IDF score of a term, the more important that term is to the document. This helps identify words that are informative within a document while not being overly common across all documents.

A comparative study [9] conducted in 2020 examined three vectorizers: Count Vectorizer, TF-IDF Vectorizer, and Hashing Vectorizer. These features were applied to classifiers SVM and KNN for sentiment analysis of YouTube comments on Nokia products. The study found that the TF-IDF vectorizer had the best performance, with nearly no errors in predicting negative values and a higher number of positive predictive values compared to the other vectorizers.

Figure 2, sourced from a blog post by DeepLearning.AI, illustrates this concept effectively.



**Figure 2:** How TF-IDF Vectorizer works [10]

### **3.3. Classification models**

We explored the use of Support Vector Machine, Random Forest, K-Nearest Neighbors and an ensemble of the three for the classification task.

#### **3.3.1. Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification tasks. It works by finding the hyperplane in an N-dimensional space (where N is the number of features) that best separates the data points of different classes in the feature space. SVM performs well with non-linear data by transforming it into a higher-dimensional space where they may be linearly separable.

A study [11] on identifying EHR progress notes pertaining to diabetes employed SVM classifier and achieved a high performance of F1 score 0.93.

The kernel function determines how the data points are mapped to the N-dimensional space. The hyper parameter C controls the error margin of the classification. For our task, we achieved best results with a polynomial kernel function and a 0.01 value of C.

#### **3.3.2. Random Forest (RF)**

Random Forest is an ensemble learning method that constructs multiple decision trees during training, and outputs the class that is the mode of the classes of the individual trees. Each tree is built by the usage of a random subset of features and records factors, leading to diversity inside the ensemble. This approach reduces the risk of over-fitting and improves the model's generalization ability as it is less sensitive to the variability of a single tree.

A paper published in 2019 [12] on the sentiment analysis of data sources from Twitter used Random Forest Classifier for classification on TF-IDF vectors. The model delivered a performance of 75%.

Our best results were obtained using a Random Forest consisting of 350 decision trees with a maximum depth of 5.

#### **3.3.3. K-Nearest Neighbors (KNN)**

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm. For a given test sample, KNN identifies the K training samples closest in feature space and assigns the most common class among those neighbors to the test sample. The simplicity of KNN makes it easy to implement and interpret, though it can be computationally expensive, especially with large datasets.

A study [13] on the classification of news topics in Indonesian language used the KNN classifier model with word2vec for feature extraction. The study yielded an accuracy of 89.2%

We attained optimal performance with a K value of 150.

#### **3.3.4. Ensemble**

Ensemble methods combine multiple machine learning models to improve overall performance. By leveraging the strengths of various models, ensemble methods can achieve higher accuracy and robustness compared to individual models.

A paper [14] published on classification of spam product reviews using an ensemble that combines predictions from Multi-layer perceptron (MLP), K-Nearest Neighbour (KNN), and Random Forest (RF) demonstrated that the ensemble outperformed individual classifiers with an accuracy of 88.13%.

We used a majority voting ensemble that aggregates the classifications made by SVM, RF, and KNN, and takes the majority vote to make the final prediction.

## 4. Results

Table 1 displays the macro average F1 scores of TF-IDF features using Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and an ensemble of these models on datasets from Catalonia, Hungary, and Belgium.

For the Catalonia dataset, RF and the Ensemble performed best with an F1 score of 0.78, closely followed by SVM at 0.77. KNN notably underperformed, trailing by a margin of 0.19 from the next lowest score. In the Hungary dataset, SVM achieved the highest score at 0.86, followed by the Ensemble at 0.78, RF at 0.74, and KNN at 0.72. SVM demonstrated significantly better performance with an F1 score of 0.8 compared to the second-best model, the Ensemble. For the Belgium dataset, SVM and the Ensemble equally attained the highest scores, each achieving an F1 score of 0.66. RF followed at 0.63, while KNN once again had the lowest score at 0.56, showing a notable margin of 0.7 from the next lowest score.

On average, SVM outperformed the other models, followed by the Ensemble, Random Forest, and lastly KNN.

In our submission of classified test data for Touché, Ideology, and Power Identification in Parliamentary Debates 2024, focusing on power identification, our SVM model with TF-IDF achieved an F1 score of 0.68, resulting in a 6th place ranking.

**Table 1**

Selected example datasets of the macro average F1 scores of classifiers with TFIDF Vectors

Datasets	SVM	RF	KNN	Ensemble
es-ct	0.77	0.78	0.58	0.78
hu	0.86	0.74	0.72	0.78
be	0.66	0.63	0.56	0.66

## 5. Conclusion

In our analysis of the different classification models that we tested, i.e., Random Forest, K-Nearest Neighbors, and an Ensemble model, with vectors generated using TF-IDF vectorization, we have concluded that the SVM model performs best on the given dataset. Such a model would prove useful to political analysts and researchers by giving them insights into political dynamics, distribution of power within legislative bodies, and rhetorical strategies. Media and journalists could use this model to determine key power strategies employed by the government and opposition. Our model would also help the education sector, as students will be able to use it to analyze speeches from both the government's and opposition's perspectives, helping them learn about power strategies and understand language dynamics in politics.

## References

- [1] S. G. Obeng, Language and politics: Indirectness in political discourse, *Discourse & Society* 8 (1997) 49–83. URL: <https://doi.org/10.1177/0957926597008001004>. doi:10.1177/0957926597008001004. arXiv:<https://doi.org/10.1177/0957926597008001004>.
- [2] H. Gruber, Political language and textual vagueness, *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* 3 (1993) 1–28.
- [3] J. Kiesel, Ç. Cöltekin, 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. Quénnot, 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.

- [4] R. Sawhney, A. Wadhwa, S. Agarwal, R. Shah, Gpols: A contextual graph-based language model for analyzing parliamentary debates and political cohesion, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4847–4859.
- [5] J. Kapočiūtė-Dzikiénė, A. Krupavičius, Predicting party group from the lithuanian parliamentary speeches, *Information Technology and Control* 43 (2014) 321–332.
- [6] M. Galley, K. McKeown, J. B. Hirschberg, E. Shriberg, Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies (2004).
- [7] C.-S. Kavallos, Parliament proceeding classification via machine learning algorithms: A case of greek parliament proceedings, 2023.
- [8] Z. Salah, Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates, Ph.D. thesis, University of Liverpool, 2014.
- [9] I. Irawaty, R. Andreswari, D. Pramesti, Vectorizer comparison for sentiment analysis on social media youtube: A case study, in: 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), 2020, pp. 69–74. doi:10.1109/IC2IE50715.2020.9274650.
- [10] DeepLearning.AI, Tokenizers and tf-idf, <https://www.deeplearning.ai/resources/natural-language-processing/>, 2022. Accessed: 2024-05-31.
- [11] A. Wright, A. B. McCoy, S. Henkin, A. Kale, D. F. Sittig, Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions, *Journal of the American Medical Informatics Association* 20 (2013) 887–890.
- [12] N. Bahrawi, Sentiment analysis using random forest algorithm-online social media based, *Journal of Information Technology and Its Utilization* 2 (2019) 29–33.
- [13] N. G. Ramadhan, et al., Indonesian online news topics classification using word2vec and k-nearest neighbor, *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5 (2021) 1083–1089.
- [14] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, B. Alouffi, Ensemble machine learning model for classification of spam product reviews, *Complexity* 2020 (2020) 8857570.