Trojan Horses at Touché: Logistic Regression for Classification of Political Debates

Notebook for the Touché Lab at CLEF 2024

Deepak Chandar S, Diya Seshan, Avaneesh Koushik and P Mirunalini

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

Abstract

This study focuses on multilingual parliamentary speech analysis, specifically identification and classification of the ideology of the speaker's party and their governing status. The approach used here provides valuable insights into the political dynamics of parliamentary debates, enhancing the understanding of legislative discourse. A Logistic Regression model (combined with Count Vectorizer) is employed, trained on a dataset comprising of diverse multilingual parliamentary speech. The model achieves an F1-score of 0.59 for ideology classification and 0.69 for determining governing status. The effectiveness of the model is demonstrated in the context of evaluating parliamentary speeches from multiple countries.

Keywords

Multilingual Speech Analysis, Binary Classification, Logistic Regression, Count Vectorizer

1. Introduction

Understanding the political environment is crucial in parliamentary discussions in order to appreciate the intricacies of legislative language. The philosophy of the speaker's party and whether the party is in power or not are two crucial factors that greatly impact the substance and tone of speeches. Using the speech content, which may be in many languages, Sub-Task 1 aims to determine the speaker's party's ideological stance (left-wing or right-wing political orientation). Sub-Task 2 is determining where the party is in the present political structure: whether they are in power as the ruling party or in opposition [1]. Determining these components with accuracy improves one's comprehension of the speaker's viewpoint and the larger political forces at work.

In today's computing environment, the capability to perform data-intensive natural language processing tasks has expanded significantly. In the context of identifying key aspects of parliamentary speakers, this paper explores the use of the Logistic Regression model (combined with Count Vectorizer) for binary classification of speeches.

2. Background

Analysing political ideologies has traditionally been a challenging task due to the lack of a detailed dataset representing individual views. A commonly employed approach which has shown remarkable prowess in capturing nuanced linguistic patterns by utilizing advanced language models (LLMs) like BERT and GPT-4, as outlined in this study which analyzes parliamentary representatives' ideological positions [2]. Previous studies have also explored the efficacy of integrating natural language processing (NLP) methods into political science research [3]. Some such studies made use of advanced NLP methods to perform sentiment analysis of parliamentary debate transcripts from European parliaments, assessing if the age, gender, and political orientation of speakers could be detected from their speeches [4]. Furthermore, a study on Indian parliamentary debates introduces structured datasets and demonstrates

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

deepakchandar2210436@ssn.edu.in (D. C. S); diya2210208@ssn.edu.in (D. Seshan); avaneesh2210179@ssn.edu.in (A. Koushik); miruna@ssn.edu.in (P. Mirunalini)

promising results in stance classification and pragmatic analysis [5]. These methodologies have provided valuable insights into the political dynamics of legislative discourse across various linguistic contexts.

Inspired by the success of these approaches in various previous studies, the work done here aims to extend and refine existing methodologies for multilingual parliamentary speech analysis. Building upon the foundation laid by these prior investigations, the accuracy and robustness of classification models can be enhanced by incorporating additional linguistic features and optimizing model parameters. By leveraging logistic regression as a reliable framework for binary classification, the aim is to deepen the understanding of the intricate interplay between political ideology, governing status, and parliamentary discourse, thereby contributing to the broader discourse on computational approaches to political analysis.

3. System Overview

3.1. Dataset Overview

The dataset comprises a collection of speeches along with metadata that includes the speaker's gender and a classification label. The following are the dataset's attributes:

- **id**: A unique identifier for each speech record. This attribute helps in referencing and tracking specific speeches within the dataset.
- **text**: The original speech text in 28 different European languages. This attribute is crucial for analyses that require the original language.
- **text_en**: The translated speech text in English. This attribute is useful for analyses where English is the preferred language for processing or interpretation. This attribute has been used as the primary source for analysis and was fed into the machine learning model for further processing.
- **sex**: The gender of the speaker, indicated by 'M' for male and 'F' for female. This attribute allows for gender-based analysis and comparisons.
- **label**: A classification label for the speech, with possible values '1' and '0'. In the context of Sub-Task 1, this label indicates the the speaker's party's ideological stance: left-wing (0) or right-wing (1). In the context of Sub-Task 2, this label indicates where the party is in the present political structure: ruling party (0) or opposition (1).

Furthermore, on analyzing the dataset for the two sub-tasks, for sub-task 1 (orientation), an average of 10422 speeches per dataset were present, with an average of 5784 instances as left-wing and 4638 as right-wing. For sub-task 2 (power), an average of 8370 speeches were present, with 4445 instances as ruling party as 3925 for opposition party. This illustrates that the dataset for both tasks was well-distributed and balanced, which is crucial for the model to effectively understand the characteristics of the data.

3.2. Data Preprocessing

Preprocessing involves extracting the relevant fields (text_en and label) from the dataset, and converting the text data into a suitable format for the model by employing vectorization using CountVectorizer. CountVectorizer is a class in scikit-learn [6] that transforms a collection of text documents into a numerical matrix of word or token counts. This class has a number of parameters that can also assist in text preprocessing tasks, such as stop word removal, word count thresholds (i.e. maximums and minimums), vocab limits, n-gram creation and more.

The parameters used for the CountVectorizer tool are as follows:

- **lowercase**: Convert all characters to lowercase before tokenizing. Set to True.
- **ngram_range**: Range of n-values for different n-grams to be extracted. Set to (1, 1): only unigrams.
- **analyzer**: Level at which the input text will be tokenized. Set to 'word'.

3.3. Proposed Model

The first model which used for experimentation was Bidirectional Encoder Representations from Transformers (BERT) uncased classifier [7]. The embeddings obtained from the CountVectorizer tool were used to train the BERT model. The model was trained separately for each language, and while some languages yielded training F1-scores of around 0.60, others yielded significantly lower F1-scores. The dataset contains a wide range of text lengths- this could have been a reason why the model exhibited low F1-scores. Another reason for the unpredictable results of the BERT model could be batch size. Batch size is the number of samples processed together in each training step. The batch size used in the model turned out to be sub-optimal for the dataset and the hardware, and due to low computational efficiency, experimentation with different batch sizes was not possible.

Hence, a different approach was chosen. The second model that was used for experimentation was logistic regression. Logistic regression is widely applied across various domains, often demonstrating superior accuracy compared to classifiers such as random forest and K-nearest neighbor in numerous empirical studies [8]. A common application of logistic regression is in sentiment analysis tasks, where it effectively categorizes text data into sentiment classes to classify emotions or opinions [9].

The initial approach involved combining the training datasets of all the languages and using this aggregated dataset to train the logistic regression model. However, this method resulted in a notably low average training F1-score. Subsequently, the model was trained separately for each language, and upon analyzing the outcomes, it was found that this method provided a higher average training F1-score. Hence, the latter method was used for subsequent analysis and evaluation.

3.4. Methodology

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome- something that can take two values such as true/false, yes/no, and so on. Logistic regression is a useful analysis method for classification problems, where the goal is to determine which category a new sample is most likely to belong to.

The logistic function is represented by the following formula:

$$Logit(\pi) = \frac{1}{1 + e^{-\pi}}$$

The embedding obtained from the CountVectorizer tool were used to train the logistic regression model for both the sub-tasks. The parameters used for the logistic regression model are as follows:

- **class_weight**: Weights associated with classes. Set to None.
- **max_iter**: Maximum number of iterations taken for the solvers to converge. Set to 300.
- **penalty**: Penalized logistic regression imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contributive variables toward zero. This is also known as regularization. Set to 'l2'.
- **random_state**: Controls the randomness of the estimator. Set to 42.
- **tol**: Tolerance for stopping criteria. Set to e-4.
- **fit_intercept**: Allows the model to make predictions more accurately by shifting the decision boundary. Set to true.
- **regularization strength 'C'**: Used to penalize large coefficients. Set to 0.3.

4. Results

The model was trained on speeches in around 25 different languages, which were translated into English for the purpose of evaluation. In the training phase, the model achieved an average F1-score of 0.99 for sub-task 1 (orientation), 0.98 for sub-task 2 (power).

In the testing phase, the model achieved its highest F1-score of 0.83 for the Power task for the Greek language, and 0.72 for the Orientation task for the Italian language. Additionally, the model's F1-score

surpassed that of the baseline for several languages. On average, the metrics measured were 3 to 5 percent higher than those of the baseline model [10]. The results obtained were analyzed based on the average performance metrics of precision, recall, and F1-scores, as illustrated in 1.

Table 1
Average Results for each Sub-Task

	Task	Precision	Recall	F1-score
	Orientation	0.62	0.60	0.59
	Power	0.67	0.70	0.69

One possible reason for the improvement in results over the baseline model could be the incorporation of the hyperparameter C into the model. The value of C was set to 0.3, determined through random search, where the model was trained and evaluated across various C values. This value of 0.3 yielded the best performance. Therefore, proper regularization helped improve the model's generalizability to new, unseen data by reducing overfitting.

The model was also analysed based on the parliamentary languages, and the top 4 highest performing languages of the two sub-tasks have been listed in tables 2 and 3.

Table 2
Top F1-scores for the Power Task

	Parliament	F1-score	Baseline F1-score
	Greece	0.83	0.79
	Austria	0.72	0.67
	Italy	0.71	0.65
	Bosnia and Herzegovina	0.57	0.41

Table 3
Top F1-scores for the Orientation Task

	Parliament	F1-score	Baseline F1-score
	Spain	0.72	0.72
	Italy	0.66	0.65
	Denmark	0.60	0.56
	Netherlands	0.59	0.58

It was found that the model performed better for the power sub-task, benefiting from the presence of well-defined discriminating features in the parliamentary speeches of the dataset. However, in case of classification based on ideology, the model struggled due to the lack of clear discriminating features for the model to capture.

5. Conclusion

In conclusion, the research demonstrates the efficacy of logistic regression as a reliable technique for binary classification in the nuanced domain of multilingual parliamentary speech analysis. Through meticulous analysis of the dataset and model training, the proposed model demonstrated the utility of the approach in interpreting key attributes of political discourse, namely party ideology and governing status.

The findings revealed compelling F1-scores, averaging at around 0.59 and 0.69 respectively for the two tasks of identifying party ideology and governing status. This highlights the reliability of Logistic Regression in capturing the inherent complexities of parliamentary debates, even in diverse linguistic contexts.

Looking ahead, further refinements and extensions of the approach hold promises in enhancing the predictive capabilities and applicability across a broader spectrum of parliamentary contexts. This includes exploring the integration of advanced language models (LLMs) such as BERT or GPT. By leveraging LLMs, it is possible to delve deeper into the complexities of parliamentary discourse, uncovering subtle semantic nuances and contextual cues that traditional methods may overlook. Additionally, the plan is to investigate novel techniques for fine-tuning LLMs on parliamentary speech data, as well as exploring ensemble methods that combine the strengths of multiple models. Through these endeavors, the aim is to develop a more comprehensive understanding of the intricate dynamics of legislative language and its implications for governance and policy making.

Due to the lack of computational resources and time, the model was trained with same features for both the sub-tasks. This work can be improved and extended by using different features for both the sub-tasks.

References

[1] 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.

[2] K. Kato, A. Purnomo, C. Cochrane, R. Saqur, L(u)pin: LLM-based Political Ideology Nowcasting, 2024. URL: https://arxiv.org/abs/2405.07320. arXiv:2405.07320.

[3] G. Glavaš, F. Nanni, S. P. Ponzetto, Computational Analysis of Political Texts: Bridging Research Efforts Across Communities, in: P. Nakov, A. Palmer (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Florence, Italy, 2019, pp. 18–23. URL: https://aclanthology.org/P19-4004. doi:10.18653/v1/P19-4004.

[4] K. Miok, E. Hidalgo-Tenorio, P. Osenova, M.-A. Benitez-Castro, M. Robnik-Sikonja, Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis, 2023. URL: https://arxiv.org/abs/2207.01054. arXiv:2207.01054.

[5] S. V. K. Rohit, N. Singh, Analysis of Speeches in Indian Parliamentary Debates, CoRR abs/1808.06834 (2018). URL: http://arxiv.org/abs/1808.06834. arXiv:1808.06834.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[7] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[8] K. Shah, H. Patel, D. Sanghvi, M. Shah, A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification, Augmented Human Research 5 (2020) 1–16. URL: https://doi.org/10.1007/s41133-020-00032-0. doi:10.1007/s41133-020-00032-0.

[9] A. Kumar, A. Mangotra, A. Ailawadi, R. Jain, M. Arora, Sentiment analysis on multilingual data: Hinglish, in: A. Swaroop, Z. Polkowski, S. D. Correia, B. Virdee (Eds.), Proceedings of Data Analytics and Management, Springer Nature Singapore, 2024, pp. 607–620.

[10] Çağrı 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, 2024. arXiv:2405.07363.