Ssnites at Touché: Ideology and Power Identification in Parliamentary Debates using BERT Model

Notebook for the Touché Lab at CLEF 2024

Keerthick V, Kaushik Ananth Kumar S, Kathir Kaman A, P Mirunalini and Sripriya N

Sri Sivasubramaniya Nadar College of Engineering College of Engineering, Chennai, TamilNadu, India

Abstract

Debates in national parliaments do not only affect the fundamental aspects of citizens' life, but often a broader area, or even the whole world. As a form of political debate, however, parliamentary speeches are often indirect and present a number of challenges to computational analyses. In this task, we focus on identifying two views associated with speakers in a parliamentary debate: their political ideology and whether they belong to a governing party or a party in opposition. Both subtasks are formulated as binary classification tasks. We have implemented a BERT model which have effectively discern underlying political orientations and affiliations in parliamentary discourse, despite its inherent indirectness and complexity. The model's performance was evaluated using the F1 score in which our model achieved an average F1 score of 0.5894 for the Orientation Task (Ideology Identification) and for the Power Task (Governing vs. Opposition), the model achieved an average F1 score of 0.6026

Keywords

Ideology, Speech, Power, BERT

1. Introduction

In the modern era, social media platforms like Twitter, Facebook, and Reddit have become central spaces for discussing various issues, including politics. When people share their ideologies or opinions on these platforms, the debates often become controversial. Gaining insight into the political climate is vital for engaging effectively in parliamentary debates and understanding the nuances of legislative terminology. This study presents the development and evaluation of machine learning models for the shared task "Ideology and Power Identification in Parliamentary Debates" as part of CLEF 2024. Utilizing the dataset derived from the ParlaMint corpora [1] [2], the proposed system aims to classify parliamentary speeches based on two distinct variables: the political ideology of the speaker (left or right) and their party affiliation (governing party or opposition).

In the current computing era, the capability to perform data-intensive natural language processing tasks has significantly advanced. In order to perform the tasks it necessary to perform political text analysis by providing robust methodologies for understanding the underlying political contexts of parliamentary discourse, ultimately aiding in the development of tools for more transparent and informed governance. So the research work explores extracting the relevant linguistic features from the dataset and these features were used to train the BERT model.

2. Background

In the task, the advancements in sentiment analysis, particularly the work of [3], who demonstrated the effectiveness of BERT in aspect-level sentiment classification was utilized. Their research highlights the challenges of accurately interpreting context-dependent information, a challenge shared by our political text classification task. By incorporating target information into BERT, they achieved state-of-the-art performance, underscoring BERT's powerful contextual representation capabilities. In the task the insights from sentiment analysis research from the work of [4] was utilized. Their study addresses the complexities of emotion detection in code-switching text, utilizing BERT for multi-label sentiment analysis. They highlight challenges such as mixed-language text, co-existing emotions, and unbalanced data distribution, which are also pertinent to political text classification. By fine-tuning BERT and employing strategies like data augmentation and ensemble learning, their approach demonstrates improved performance. This aligns with the methodology of adapting BERT to capture the nuanced political sentiments in parliamentary discourse.

Iyyer et al. (2014) [5] explored the recognition of different political ideologies using a Recurrent Neural Network (RNN). They applied tree-structured RNNs to capture the systematic composition of text in a document, achieving better accuracy in this task. Hojoon and Minbyul (2017) [6] found that RNNs suffered from gradient vanishing problems with long sentences. To address this issue, the proposed study used LSTMs. Devlin et al. (2018) [7] assert that pre-training deep bidirectional representations from unlabeled text using Bidirectional Encoder Representations from Transformers (BERT), which jointly conditions on both left and right context in all layers, surpasses previous Natural Language Processing methods in performance across various tasks, including text classification.

3. Dataset

To perform the tasks, two distinct training datasets were provided, one for the orientation task and another for the power task. Both datasets share the same structure, containing the columns: **id, speaker, sex, text, text translated in english, and label**. The dataset comprises speeches from multiple national and regional parliaments, providing a diverse linguistic and political landscape. The orientation task dataset includes a total of 148,000 records, while the power task dataset comprises 200,000 records. Each dataset is further divided into separate CSV files for different languages, with each set including English translations. Additionally, similar attributes as test datasets were also be provided. This structure and distribution ensure that we have a comprehensive and diverse collection of data to support the training and evaluation of models for both orientation and power tasks.

4. System Overview

The system developed for this study integrates multiple components designed to process, analyze, and classify parliamentary speeches. The flow diagram of the proposed system is depicted in the figure 1.

The system architecture of the have been divided into several key modules:

4.1. Data Ingestion and Preprocessing

Data collection involves acquiring parliamentary debate transcripts from given sources. The text cleaning process removes non-essential elements such as procedural annotations and speaker identifiers. Tokenization splits the text into individual words or tokens.
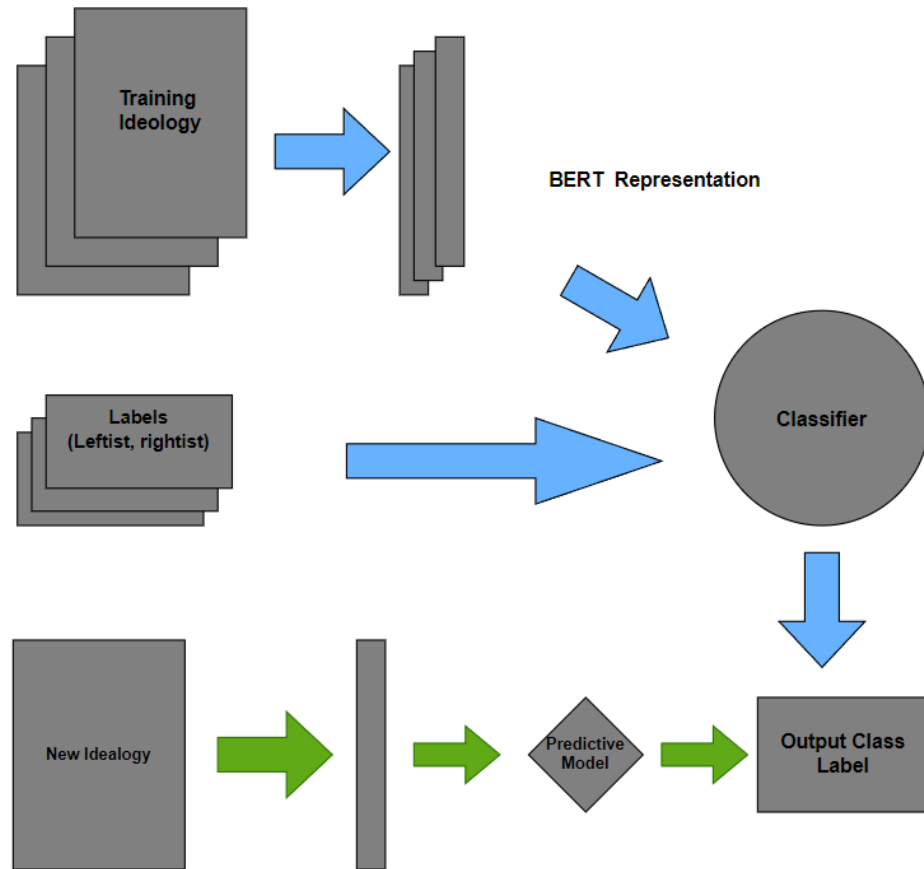
Figure 1: Flow Diagram of the proposed System

4.2. Feature Extraction

The Bidirectional Encoder Representations from Transformers (BERT) model [8] was proposed to capture contextual meanings and dependencies within the text for semantic features. The "bert-base-uncased" variant of the BERT model was used, which was accessible through the Hugging Face library. This model is specifically designed for natural language processing tasks and features 12 layers with 768 hidden units per layer, employing 12 attention heads. The "uncased" attribute indicates that the model operates in a case-insensitive manner, which is suitable for various text processing applications.

4.3. Proposed Model

In this proposed work, the pre-trained BERT model is fine-tuned along with an additional untrained classification layer for our specific tasks which is depicted in Figure 2. BERT is a contextual model that captures relationships in a deeply bidirectional manner, allowing it to represent tokens based on their dependencies with other tokens in the text. Its pre-training tasks and architecture enable BERT to detect semantic relationships between words and sentences, making it versatile for various tasks. Pre-trained on large corpora for next sentence prediction and language modeling, BERT uses an innovative method for language modeling, considering both the left and right contexts of the masked token for predictions.

4.4. Model Training

We have implemented a transformer based model BERT for our study. The training process uses stratified train-test splits and cross-validation to ensure balanced and robust training. The dataset is
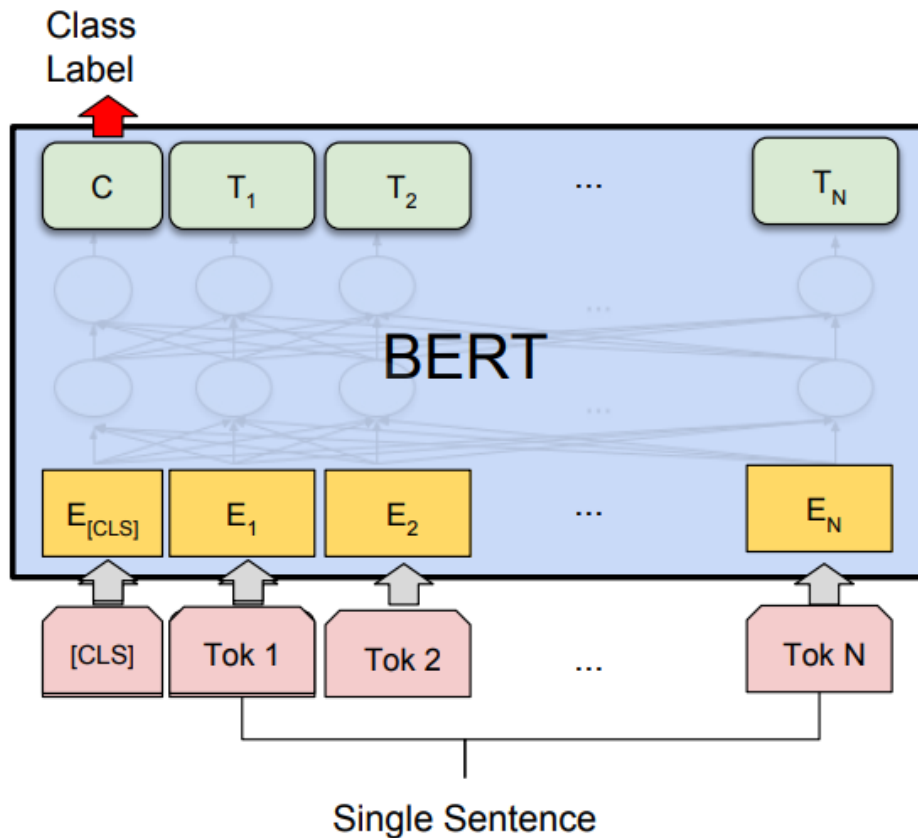
Figure 2: BERT Diagram

loaded from a CSV file using pandas, and a subset of the data is selected based on specified indices. Each text entry in the dataset is tokenized using the BERT tokenizer, which converts text into token IDs that the BERT model can process. The tokenized sequences are then padded to ensure uniform input lengths.The special tokens (such as [CLS] and [SEP]) required by the BERT model were added, ensured that sequences longer than the max length are truncated, and set the maximum sequence length for tokenized texts to 256 tokens. The batch size is set to 16, determining the number of samples processed together in one forward/backward pass.

Next, the input IDs and labels are combined into a TensorData. This data was loaded into a DataLoader for batching and shuffling during training. The pre-trained BERT model (bert-base-uncased) is initialized with a sequence classification head. If training is continuing from a checkpoint, the model's state dictionary is optionally loaded from a file.

The optimization process involves using the AdamW optimizer for updating the model parameters, with a specified learning rate.The learning rate for the AdamW optimizer is set to 2e-5, controlling the adjustment of model weights with respect to the loss gradient. We have set the number of epochs to 10, which defines the number of complete passes through the training dataset. For each batch, the model performs forward and backward passes to calculate loss and update weights. The loss is computed, accumulated, and averaged for each epoch. The optimizer updates the model's parameters based on the computed gradients. The final state of the trained model is saved to a file.To optimize performance, the CUDA [9] was employed to accelerate the training of BERT-based models [10], improving efficiency and scalability.

4.5. Classification

The BERT model trained for both ideology and party status classification of parliamentary speeches.

In this research work, a BertForSequenceClassification model [7] which specifically tailored for sequence classification tasks was used. This model variant is part of the Hugging Face Transformers library, renowned for its effectiveness in natural language processing applications.The BertForSequenceClassification model is initialized with parameters suitable for binary classification tasks, accommodating two distinct labels. It integrates BERT's pre-trained layers, which capture contextual information from text inputs through attention mechanisms across multiple layers and attention heads. This architecture enables the model to encode and understand complex relationships within sequences, essential for tasks such as sentiment analysis, text categorization, and other forms of sequence-based classification. This dataset is then used to train the BERT model for sequence classification. After training, the model is capable of classifying new speeches to determine the speaker's political ideology and their party's status (governing or opposition) based on the speech content.

5. Results

In our study, transformer model was employed to tackle two specific tasks using parliamentary speeches in multiple languages. The first task involves identifying the ideology of a speaker's party based on their speech content. The second task focuses on determining whether the speaker's party is currently governing or in opposition.

The model's performance was evaluated using the F1 score, which balances precision and recall. The model achieved an average F1 score of 0.5894, 0.6026 for the Orientation Task (Ideology Identification) and the Power Task (Governing vs. Opposition), and also achieved an average F1 score of 0.6026. These scores indicate moderate performance, suggesting the model performs better than random guessing but still has significant room for improvement.The vector representation of datasets includes masks and labels. For example, masks are tensors indicating which tokens are actual words and which are padding, while labels are tensors indicating the class of each input.

The team has achieved the 5th position in the orientation and the 10th position in the power tasks with respect to average F1_Score. The results of the task have been displayed in the Table ??

To further enhance our results, several strategies can be implemented. Firstly, expanding our training across a broader range of datasets can enrich our model's understanding and performance. Additionally, fine-tuning tokenization parameters and adjusting maximum sequence lengths to better suit the input data characteristics can optimize text processing. Introducing regularization techniques such as dropout and layer normalization can stabilize training, mitigate overfitting, and enhance generalization. Furthermore, exploring advanced BERT model architectures, including larger variants or domain-specific pre-trained models, will enable capturing more nuanced linguistic features pertinent to specific tasks, thereby boosting classification accuracy.Developing language-specific models or using techniques like transfer learning to adapt the model to different languages can help manage the diversity in political discourse. Utilizing multilingual training approaches can also be beneficial. Implementing these strategies is crucial for advancing our accuracy and performance metrics.

Table 1
Result

Task	F1 Score	Precision	Recall
Orientation	0.589	0.589	0.600
Power	0.602	0.622	0.626

Table 2
Performance Comparison- Orientation

Orientation	F1 Score
TR (Turkey)	0.790
SE (Sweden)	0.747
PL (Poland)	0.712
NO (Norway)	0.659
PT (Portugal)	0.644
RS (Serbia)	0.641
GB (United Kingdom)	0.637
ES-GA (Galicia, Spain)	0.637
ES (Spain)	0.608
DK (Denmark)	0.608

Table 3
Performance Comparison- Power

Power	F1 Score
HU (Hungary)	0.785
GR (Greece)	0.690
AT (Austria)	0.659
FR (France)	0.651
HR (Croatia) (Hrvatska)	0.650
CZ (Czech Republic)	0.610
DK (Denmark)	0.612
TR (Turkey)	0.613
ES (Spain)	0.621
IT (Italy)	0.621

6. Conclusion

This study demonstrates the viability of binary classification for political ideology and party status in parliamentary debates, contributing valuable insights into political discourse analysis. Our findings underscore the potential of computational techniques in political science, paving the way for more sophisticated and scalable analysis methods.

References

[1] 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`.

[2] Ç. Çöltekin, M. Kopp, M. Katja, V. Morkevicius, N. Ljubešić, T. Erjavec, Multilingual power and ideology identification in the parliament: a reference dataset and simple baselines, in: D. Fiser, M. Eskevich, D. Bordon (Eds.), Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 94–100. URL: https://aclanthology.org/2024.parlaclarin-1.14.

[3] Z. Gao, A. Feng, X. Song, X. Wu, Target-dependent sentiment classification with bert, Ieee Access 7 (2019) 154290–154299.

[4] T. Tang, X. Tang, T. Yuan, Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text, IEEE Access 8 (2020) 193248–193256.

[5] M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers), 2014, pp. 1113–1122.

[6] L. Hojoon, J. Minbyul, Ideology detection with using deep neural network (2017).

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[8] S.-H. Chiu, B. Chen, Innovative BERT-based reranking language models for speech recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 266–271.

[9] K. Ho, H. Zhao, A. Jog, S. Mohanty, Improving gpu throughput through parallel execution using tensor cores and cuda cores, in: 2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), IEEE, 2022, pp. 223–228.

[10] M. Koroteev, BERT: a review of applications in natural language processing and understanding, arXiv preprint arXiv:2103.11943 (2021).

[11] 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-Münstermann, 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.

[12] X. Zhao, J. Greenberg, Y. An, X. T. Hu, Fine-tuning BERT model for materials named entity recognition, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 3717–3720.

[13] E. Demir, M. Bilgin, Sentiment analysis from Turkish news texts with BERT-based language models and machine learning algorithms, in: 2023 8th International Conference on Computer Science and Engineering (UBMK), IEEE, 2023, pp. 01–04.

[14] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).