TüNLP at Touché: Finetuning Multilingual Models for Ideology Detection

Notebook for the Touché Lab at CLEF 2025

Aydemir Shamsutdinov[1,†], Joaquin Cherta-Rodríguez[1,†]

[1]University of Tübingen, 72070 Tübingen, Germany

Abstract

The TOUCHÉ 2025 shared task on ideology and power identification in parliamentary debates challenges participants to classify political speeches into left- or right-wing ideologies. Team TüNLP presents an approach using XLM-RoBERTa-large, a multilingual transformer model, fine-tuned on parliamentary data from multiple European countries. We address data imbalance with focal loss and layer-wise learning rate decay, achieving robust performance on validation sets. This paper details our methodology, experimental setup, and discusses the implications of our findings for cross-lingual ideology detection.

Keywords

Ideology Detection, Parliamentary debates, XLM-RoBERTa, TOUCHÉ 2025, Cross-lingual Classification

1. Introduction

Being a key element of the political ecosystem of modern states and countries, parliaments are designed to play a central role in the daily political decisions and represent the ideology and political views of the citizens. These spaces are often subject to political analysis to explore the political and social dynamics within countries, trying to understand general ideological shifts, communication strategies and patterns, and linguistic characteristics associated with different political categories, such as orientation, power, and populism. As such, it is of general interest to be able to perform a thorough analysis of these political spaces to better understand how they function. With this goal in mind, computation can be a valuable tool to perform these analyses on large-scale datasets, or increase the capabilities of the data being processed for further, more detailed analysis. However, the computational analysis of parliamentary speeches presents unique challenges when dealing with the specific characteristics found in political speeches, which are often indirect. This task is critical for understanding political discourse across European parliaments, where linguistic and cultural diversity pose significant challenges for automated analysis systems.

As part of the TOUCHÉ 2025 shared tasks [8], Ideology and Power Identification in Parliamentary Debates 2025 shared task challenges participants to find computational solutions to the difficulties found in the processing of parliamentary speeches and debates. The shared task consists of three Sub-Tasks (TOUCHË, 2025):

Sub-Task 1: Given a parliamentary speech in one of several languages, identify the ideology of the speaker's party.

Sub-Task 2: Given a parliamentary speech in one of several languages, identify whether the speaker's party is currently governing or in opposition.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[†]These authors contributed equally.

aydemir.shams@gmail.com (A. Shamsutdinov); jchertar96@gmail.com (J. Cherta-Rodríguez)

Sub-Task 3: Given a parliamentary speech, identify the position of the speaker's party in populist - pluralist scale.

Our team focuses on Sub-Task 1, with the goal of identifying the ideology of the speaker given a parliamentary speech. Instead of dealing with each parliament and language separately, we propose a solution utilizing a multilingual transformer model, XLM-RoBERTa-large [1], to address the cross-lingual nature of the dataset. Our approach focuses on handling the inherent class imbalance in political speech data and adapting pre-trained multilingual models for ideological classification tasks. In this paper, we aim to outline our approach, the experimental results obtained, and what we learned after participating in this challenging task.

2. Related work

Ideology detection in political texts has been explored using various natural language processing techniques across different domains and languages. The shared tasks of PoliticIT at EVALITA (2023) [7] and PoliticES at IberLEF(2022) [6] demonstrated the application of binary and multiclass classification to extract ideology from a set of tweets in Italian and Spanish respectively, establishing approaches for automated analysis of political text. The participants proposed different methodologies, from transformer-based approaches to traditional machine learning algorithms to combinations of both. Although these shared tasks were based on the detection of monolingual political ideologies, it gives us a solid base to understand the complexity and challenges presented by the TOUCHÉ 2025 shared task, since we can analyze related approaches that have been used previously and evaluate their usefulness and adaptability regarding the detection of multilingual political ideologies.

The development of transformer-based models revolutionized cross-lingual text classification. Conneau et al. (2020) [1] introduced XLM-RoBERTa, which showed significant improvements in multilingual comprehension tasks, making it particularly suitable for parliamentary data covering multiple European languages. The authors demonstrated this by pretraining a Transformer-based masked language model in one hundred languages. The XLM-RoBERTa model outperformed multilingual BERT (mBERT) [3] on various cross-lingual benchmarks. The model is also very capable when dealing with monolingual-based tasks, and can be competitive with strong monolingual models on benchmarks like GLUE [10] and XNLI [2]. This makes the XLM-RoBERTa model a strong option to tackle our shared task on parliamentary data given the lower computational power required in comparison with larger generative models.

Previous iterations of the TOUCHÉ workshop [9] have also highlighted the effectiveness of transformer models in parliamentary data analysis, with participants successfully applying various fine-tuning strategies and data preprocessing techniques. Although there are many options to explore and different possible approaches to tackle this task, the studies performed in the previous iterations consistently demonstrated that multilingual pre-trained models outperform monolingual approaches when dealing with diverse parliamentary datasets.

Our work builds on these foundations by adapting XLM-RoBERTa specifically for ideology classification on a multilingual set of parliamentary data, incorporating advanced techniques such as focal loss to handle class imbalance and layer-wise learning rate decay to improve model convergence in the parliamentary debate domain.

3. Methodology

3.1. Data Preprocessing

The TOUCHÉ 2025 dataset includes parliamentary speeches from 29 European countries, including Austria, Belgium, the Czech Republic, and others. Every speech is labeled as belonging to either the left or the right. Our preprocessing pipeline included several important steps to ensure data quality and model compatibility.

All text with a maximum sequence length of 512 tokens was tokenized using the XLM-RoBERTa tokenizer. Longer speeches were truncated and shorter speeches were batch padded to preserve uniform input dimensions. Missing values and incomplete entries were identified and removed from the training process in order to prevent noise in the model learning process.

Because the dataset was multilingual, we used XLM-RoBERTa's multilingual capabilities to learn cross-lingual representations while maintaining the original language content without translation. This approach maintains the authentic linguistic characteristics of each parliament's discourse patterns.

3.2. Model Architecture

Our base model, XLM-RoBERTa-large, has 24 transformer layers that have been pre-trained on multilingual data in more than 100 languages [1].

A task-specific head comprising a linear layer mapping the model's hidden representations to two output classes (left and right ideology) was added for the classification task. To avoid overfitting on the comparatively small parliamentary dataset, the classification head incorporates dropout regularization with a probability of 0.1.

3.3. Training Strategy

We used focal loss [11] with a gamma parameter of 2.0 to address the intrinsic class imbalance in political speech datasets. When one ideological orientation is noticeably more common than the other, this loss function—which downweights straightforward examples and concentrates learning on cases that are difficult to categorize—is especially beneficial.

To make sure the under-represented class gets the proper attention during training, we created class weights based on the inverse frequency of labels in the training set. Furthermore, we employed layer-wise learning rate decay, which allows higher layers to respond to task demands more rapidly while allowing lower layers to update more slowly. This decay factor is 0.95.

The base learning rate was set to 1×10^{-5}, and training was conducted over a maximum of 10 epochs with early stopping implemented based on validation F1-score with a patience of 3 epochs. This strategy prevents overfitting while ensuring optimal model performance.

4. Experiments

4.1. Dataset

The complete TOUCHÉ 2025 dataset containing training and test sets from various European parliaments was incorporated in the experiments. We performed an initial model training with the training data at hand, for example, `at-train.tsv`. The validation sets allowed for hyperparameter tuning, model selection, and other optimizations.

This data illustrates how distinct parliaments experience varying levels of class imbalance, with some countries having stronger bias toward certain ideologies than others. This variation was particularly useful for benchmarking our approaches to class imbalance mitigation.

For the final evaluation, we applied our trained model in the official test set for the submissions for the evaluation of shared tasks.

4.2. Evaluation Metrics

The primary evaluation metrics were accuracy and macro F1 score, which is the official evaluation protocol defined by the task organizers. F1-score macro requires that both class labels are considered which is helpful in assessing the dual classifications relevant in political text analysis which usually suffers from class imbalance.

For per class precision and recall, we monitored these metrics to understand how the model was classifying each group that was described in the text in terms of bias as certain groups may not be equally represented.

4.3. Experimental Setup

The system specifications required to run the model were a NVIDIA Quadro RTX 5000 GPU, with available memory resources of 16GB. Concerning the training and evaluation, we set a training batch size to 12 and evaluation batch size to 24 to achieve balance in utilizing memory resources without impacting speed of training.

Final predictions were made after model checkpoints on validation F1-scores were set to best for the model state in use. Around three to four epochs were sufficient for the training to reach convergence, and in combination with early stopping, saved computations, reduced resources and avoided overfitting.

5. Results

In the validation datasets from various Parliaments across Europe, our approach proved successful. Table 1 presents the performance statistics from our exploration during training.

Table 1
Performance results on validation sets during training

Metric	Score
Accuracy	0.73
Precision (Left)	0.63
Precision (Right)	0.84
Recall (Left)	0.80
Recall (Right)	0.69

With regards to cross-parliamentary contexts, the model was able to cross-detect ideologies in different languages with relatively constant performance. This suggests multilingual pre-training performed by XLM-RoBERTa captures fundamental patterns associated with ideologies along languages and dialects.

The implementation of focal loss appears to have mitigated the class imbalance problem, potentially improving recall for minority ideological groups compared to cross-entropy loss; however, this requires confirmation with future experiments comparing results with and without focal loss. A per-parliament analysis revealed performance differences, likely due to varying styles of political discourse and class distributions across countries. Overall, the consistency in results suggests that our approach provides a stable foundation for cross-lingual parliamentary analysis.

TIRA's evaluation results for our submission on 2025-05-27 are listed in Table 2. Overall, their score for Orientation F1 metric stands at 0.647 which is somewhat disheartening. For individual reviews, F1 score per example for Es-Ga was 0.844 while Ba sat at 0.485.

Task organizers have yet to provide results from the official test set, which will decide the value of our approach within the shared task setting.

Table 2
Evaluation results on the test set for Orientation (TIRA, 2025-05-27).

Parliament	Precision	Recall	F1
Overall	-	-	0.647

6. Discussion

We employ XLM-RoBERTa's cross-lingual features and address class imbalance with focal loss. Evaluation on the test set shows an overall F1-Score of 0.647, with Es-Ga achieving 0.844 and Ba 0.485. Spain Galicia's higher performance may be due to alignment with the training data distribution; Bosnia's lower F1 score, on the other hand, suggests there may be some linguistic or contextual differences posing greater difficulty. Unlike validation benchmarks (0.73 F1), the test results in comparison to validation metrics indicate overfitting or domain shifts, which we intend to resolve in subsequent iterations. Enhancements in future work could be applying ensemble strategies or directed data augmentation for underperforming parliaments.

7. Conclusion

Parlimentary discussions by TüNLP Team's TOUCHÉ 2025 Sub-Task 1 are ideologically aligned using XLM-RoBERTa. We also conducted a preliminary evaluation on the test set and achieved moderate results, but the official outcomes are still undisclosed. These findings validate the application of more sophisticated transformer architectures for document analysis in political science and computing while noting that additional research in the field of automated political science offers diverse pathways. These and other results encourage further research focusing on ensemble methods as well as advanced techniques for dealing with class imbalance.

Acknowledgments

We thank the TOUCHÉ 2025 organizers for their support during the submission process and for providing an excellent evaluation framework for cross-lingual ideology detection research. We also acknowledge the computational resources provided by our institution that made this research possible.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Citation Management. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[4] Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., et al. ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe. In *Language Resources and Evaluation*, Springer, 2024, pp. 1–32.

[5] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., & Potthast, M. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, edited by J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, pp. 236–241. Lecture Notes in Computer Science. Springer, Berlin Heidelberg New York, 2023. DOI: 10.1007/978-3-031-28241-6_20.

[6] García-Díaz, J. A., Jiménez Zafra, S. M., Martín Valdivia, M. T., García-Sánchez, F., Ureña López, L. A., and Valencia-García, R. Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology. In *Procesamiento del Lenguaje Natural*, Sociedad Española para el Procesamiento del Lenguaje Natural, 2022, pp. 265–272.

[7] Russo, D., Jiménez-Zafra, S.M., García-Díaz, J.A., Caselli, T., Guerini, M., Ureña-López, L.A., and Valencia-García, R. PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, CEUR Workshop Proceedings, vol. 3473, 2023, pp. 1–8.

[8] Kiesel, J., Çöltekin, Ç., Gohsen, M., Heineking, S., Heinrich, M., Fröbe, M., Hagen, T., Aliannejadi, M., Erjavec, T., Hagen, M., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Scells, H., Zelch, I., Potthast, M., & Stein, B. Overview of Touché 2025: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025)*, edited by J. Carrillo-de-Albornoz et al. Lecture Notes in Computer Science. Springer, Berlin Heidelberg New York, 2025.

[9] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., De Longueville, B., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., and Stein, B. Overview of Touché 2024: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14959 of *Lecture Notes in Computer Science*, pages 308–332, September 2024. Springer. researchgate.net+4

[10] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S.R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[11] Lin, T.-Y., et al., 2017. Focal Loss for Dense Object Detection. In *Proceedings of ICCV*, 2980–2988.