9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

When the sum is more than its parts: Comparing traditional and transformer-based classifiers for cross-country and cross-domain policy topic classification

Keywords: transformer-based multilingual models, supervised machine learning, comparative politics, policy topic classification

Extended Abstract

A key challenge of comparative politics is to accurately identify, track, and compare attention to policy topics across different *content* and *linguistic* domains–such as for example attention for environmental issues in parliament and the media in different countries. Yet, conducting such analyses manually is challenging and costly, requiring access to expert coders who speak the target language and significant project management resources. To overcome these challenges, researchers have—with varying success—explored automated techniques to aid the coding of predefined topics [3, 1, 4, 5, 8], particularly supervised machine learning.

Yet, the measurement of topics using traditional supervised techniques is limited in two important ways. Firstly, these models rely on bag-of-words representations of texts, making them prone to challenges related to language transferability [6]. Without an explicit translation step, the semantic meaning of the text can be lost, which is especially problematic when dealing with research questions that span geographic borders. Secondly, traditional classifiers have limited ability to understand the contextual meaning of languages [3, 7]. This poses a significant challenge when applying the models outside the specific context on which they are trained, such as using models trained on parliamentary speeches to classify media texts. Together, traditional supervised techniques are expected to prioritize domain-specific performance, sacrificing generalizability to new domains and contexts beyond their original training data.

The current study investigates the advantages of employing transformer-based multilingual models for policy topic classification across countries and domains. These models enable researchers to exploit the potential of available multilingual datasets for classification tasks within and across countries while enhancing the classifiers' contextual and semantic comprehension. The study aims to determine the extent to which the performance of cross-domain classification can be improved by transformer-based models using training data in multiple languages, in comparison to traditional supervised machine learning.

Approach

To systematically investigate the classification performance of traditional and transformer-based models across *linguistic* and *content* domains, we draw on existing annotated datasets from the *Comparative Agendas Projects* (CAP), the *Comparative Manifesto Project* (CMP) as well as annotated datasets part of published studies [3, 5]. We selected datasets from the following content domains: *executive speeches*, *parliamentary questions*, *news media*, and *party manifestos* from the following linguistic domains: *Netherlands*, *Spain*, and *UK*

($N_{totalsample} = 512503$). Our analyses center on the multi-class classification problem of five major topics—*civil rights*, *education*, *environment*, and *immigration*–along with a residual *other* category. All these topics appear in the master codebooks of both CAP and CMP. To ensure a level playing field, we have taken a stratified sample of annotated training data for each combination of linguistic and content domains ($N_{language*content} = 4996$, $N_{Final} = 59952$). Ultimately, the data represents a highly imbalanced multiclass classification problem.

We evaluated the performance of supervised machine learning models using various combinations of classifiers (*logistic regression*, *linear SVC*, *multinomial NB*, and *RandomForest*) and vectorizers (*count* and *tfdf*). The data was split into separate training and test sets for each linguistic * content domain combination. To estimate the performance per content domain, we compared these results with the effectiveness of a multilingual transformer-based classifier that was fine-tuned on the *bert-base-multilingual-uncased model* using the combined multilingual data. The classifiers were optimized towards Macro F1, in order to give relatively more weight to the minority classes. We used a learning rate of 3e-5 and a maximum of 5 epochs.

Selected Results

Figure 1 displays the in-domain performance of traditional and transformer-based classifiers. The results indicate that, across a range of content domains, transformer-based classifiers consistently outperformed the best-performing traditional classifiers, with only minor variations observed. Furthermore, except for parliamentary questions, transformer-based classifiers demonstrated comparable performance across linguistic domains, indicating that similar concepts can be measured with consistent accuracy across country-specific data.

Turning to Figure 2, we can see that performance drops for both traditional and transformer-based classifiers when attempting to predict topics in content domains on which they were not trained. Nevertheless, transformer-based approaches continued to outperform traditional approaches even in out-of-domain performance.

Conclusions

The findings presented in this study suggest that transformer-based classifiers offer significant advantages over traditional approaches and may be an effective solution for various linguistic domains. Although transformer-based models show potential for out-of-domain classification when compared to traditional models, researchers must exercise caution when using them beyond their fine-tuned context.

Overall, this study sheds light on the potential of transformer-based models trained on multilingual data to enhance cross-domain classification performance compared to traditional supervised machine learning methods. The field CSS has primarily focused on the English language, and there is a lack of adequate tools to validly measure concepts in other languages [2]. The current study helps demonstrate how state-of-the-art transformer models can enable researchers to make the most of available multilingual datasets for classification tasks within and across content and linguistic domains and improve the contextual and semantic understanding of the classifiers. By leveraging available annotated datasets for cross-comparative research, the study's insight accelerate progress in this field.

References

[1] Quinn Albaugh et al. "Comparing and combining machine learning and dictionary-based approaches to topic coding". In: *th Annual Comparative Agendas Project (CAP) Conference, Konstanz, Germany*. 2014.

[2] Christian Baden et al. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda". en. In: *Communication Methods and Measures* 16.1 (Jan. 2022), pp. 1–18.

[3] Bjorn Burscher, Rens Vliegenthart, and Claes H. De Vreese. "Using supervised machine learning to code policy issues: Can classifiers generalize across contexts?" en. In: *The ANNALS of the American Academy of Political and Social Science* 659.1 (May 2015), pp. 122–131.

[4] Mladen Karan et al. "Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts". en. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 12–21.

[5] Anne C. Kroon, Toni van der Meer, and Rens Vliegenthart. "Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification". en. In: *Computational Communication Research* 4.2 (Oct. 2022), pp. 528–570.

[6] Hauke Licht. "Cross-lingual classification of political texts using multilingual sentence embeddings". en. In: *Political Analysis* (Jan. 2023), pp. 1–14.

[7] Moritz Osnabrügge, Elliott Ash, and Massimo Morelli. "Cross-domain topic classification for political texts". en. In: *Political Analysis* 31.1 (Jan. 2023), pp. 59–80.

[8] Miklós Sebők and Zoltán Kacsuk. "The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach". en. In: *Political Analysis* 29.2 (Apr. 2021), pp. 236–249.

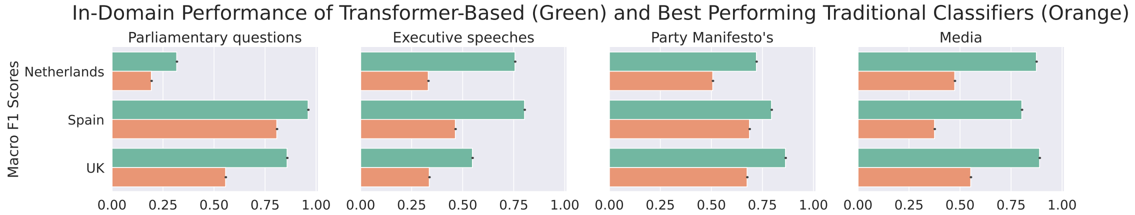

Figure 1: Macro F1 in-domain performance of Transformer-based classifiers trained on a combined multilingual corpus compared to traditional bag-of-word based classifiers.

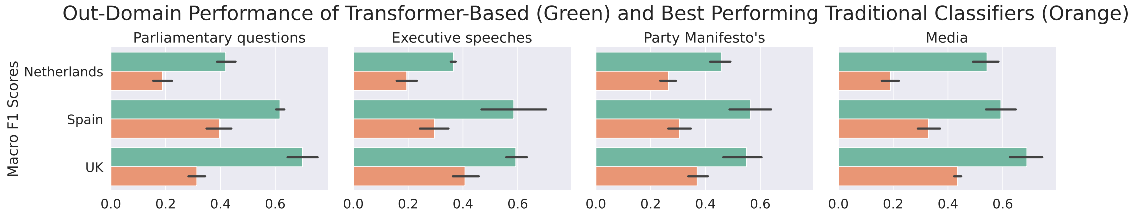

Figure 2: Macro F1 out-of-domain performance of Transformer-based classifiers trained on a combined multilingual corpus compared to traditional bag-of-word based classifiers.