OPTED

Observatory for Political Texts in European Democracies:
A European research infrastructure

Is the Sum More Than Its Parts? Multilingual and Cross-Domain Topic Classification

Authors: Anne Kroon[1], Jeroen Jonkman[1], Rens Vliegenthart[2], Christoffer Green-Pedersen[3], Shaun Bevan[4]

[1] *Department of Communication, University of Amsterdam*

[2] *Strategic Communication, University of Wageningen*

[3] *Department of Political Science, Aarhus University*

[4] *School of Social and Political Science, University of Edinburgh*

Disclaimer

This project has received funding from the European Union’s Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors’ views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public/ Confidential (Please select based on Table 3.1d in the Proposal: PU = Public, CO = Confidential)

Type

Report/Website/Demonstrator (Please select based on Table 3.1d in the Proposal: R = Report, DEC = Websites, Patent, filling, DEM = Demonstrator)

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Is the Sum More Than Its Parts? Multilingual and Cross-Domain Topic Classification

Deliverable D8.5

Authors: Anne Kroon[1], Jeroen Jonkman[1], Rens Vliegenthart[2], Christoffer Green-Pedersen[3], Shaun Bevan[4]

[1]***Department of Communication, University of Amsterdam***
[2]***Strategic Communication, University of Wageningen***
[3]***Department of Political Science, Aarhus University***
[4]***School of Social and Political Science, University of Edinburgh***

Due date: September 2023

Executive Summary

This study explores different automated supervised techniques for *multilingual* and *cross-domain* topic classification in the context of comparative politics research. There is a lack of comparative research evaluating the effectiveness of different techniques for classifying multilingual content originating from diverse text domains. Consequently, it is unclear which combination of supervised techniques can preserve accuracy during transitions between distinct social domains. To address this issue, the current study investigates the effectiveness of two central approaches for the automated classification of multilingual content: *multilingual* and *monolingual* approaches, employing a range of different techniques, namely *bag of words*, *machine translation*, *multilingual sentence embeddings* (MSE) and the *fine-tuning* (FT) framework using pre-trained language models. To evaluate the effectiveness of these techniques across different lingual and social domains, the study conducts experiments using annotated data from the *Comparative Agendas Project* and the *Comparative Manifesto Project* from three domains (parliamentary questions, media content, and party manifestos) and five languages (English, Dutch, German, Spanish, and Hungarian). It furthermore demonstrates the effectiveness of the classifiers to study overtime attention to policy issues in the *Guardian* and the UK Parliament. The findings demonstrate that *multilingual* approaches render the most optimal classification results *within* social domain. However, when moving to *out of domain* predictions, *multilingual* models using fine-tuning techniques outperform the other models. The findings highlight the potential of the fine-tuning framework as a powerful technique for cross-domain topic classification while emphasizing the need for thoughtful consideration when applying the framework in diverse settings.

Full paper

Is the Sum More Than Its Parts? Multilingual and Cross-Domain Topic Classification

Anne Kroon

Department of Communication

University of Amsterdam

Jeroen Jonkman

Department of Communication

University of Amsterdam

Rens Vliegenthart

Strategic Communication

Wageningen University

Christoffer Green-Pedersen

Department of Political Science

Aarhus University

Shaun Bevan

School of Social and Political Science

University of Edinburgh

Author Note

Anne Kroon https://orcid.org/0000-0001-7600-7979

Correspondence concerning this article should be addressed to Anne Kroon, Department of Communication Science, University of Amsterdam, Postbus 15791, 1001 NG Amsterdam, Netherlands. E-mail: a.c.kroon@uva.nl

Earlier versions of this paper were presented at the *15th Annual Comparative Agendas Project Conference*, Antwerp, June 27-29 2023 and at the *9th International Conference on Computational Social Science*, Copenhagen, July 17-20 2023.

Abstract

This study explores different automated supervised techniques for *multilingual* and *cross-domain* topic classification in the context of comparative politics research. There is a lack of comparative research evaluating the effectiveness of different techniques for classifying multilingual content originating from diverse text domains. Consequently, it is unclear which combination of supervised techniques can preserve accuracy during transitions between distinct social domains. To address this issue, the current study investigates the effectiveness of two central approaches for the automated classification of multilingual content: *multilingual* and *monolingual* approaches, employing a range of different techniques, namely *bag of words, machine translation, multilingual sentence embeddings* (MSE) and the *fine-tuning* (FT) framework using pre-trained language models. To evaluate the effectiveness of these techniques across different lingual and social domains, the study conducts experiments using annotated data from the *Comparative Agendas Project* and the *Comparative Manifesto Project* from three domains (parliamentary questions, media content, and party manifestos) and five languages (English, Dutch, German, Spanish, and Hungarian). It furthermore demonstrates the effectiveness of the classifiers to study overtime attention to policy issues in the *Guardian* and the UK Parliament. The findings demonstrate that *multilingual* approaches render the most optimal classification results *within* social domain. However, when moving to *out of domain* predictions, *multilingual* models using fine-tuning techniques outperform the other models. The findings highlight the potential of the fine-tuning framework as a powerful technique for cross-domain topic classification while emphasizing the need for thoughtful consideration when applying the framework in diverse settings.

Keywords: multilingual text classification, comparative politics, transformer-based models, automated text analysis

Is the Sum More Than Its Parts? Multilingual and Cross-Domain Topic Classification

Introduction

A key challenge of comparative politics is to accurately identify, track, and compare attention to policy topics across different *lingual* and *content* domains–such as for example attention to immigration or environmental issues in parliament, party rhetorics, executive communication and media coverage in different countries. Attention is considered a vital aspect of politics - with questions relating to (transfer of) issue salience figuring prominently in research fields such as political communication (Walgrave & Van Aelst, 2006) and public policy (Baumgartner & Jones, 2010). Yet, analyzing attention for policy topics manually is challenging and costly, requiring access to expert coders with domain knowledge who speak the target language, as well as significant project management resources. To overcome these challenges, researchers have—with varying success—explored automated techniques to aid the coding of predefined topics (Albaugh et al., 2014; Burscher et al., 2015; Sebők & Kacsuk, 2021), particularly supervised machine learning.

When dealing with multilingual corpora, researchers typically have two main options. Firstly, they can choose to combine all training data and train classifiers using a *multilingual* corpus. This approach offers the advantage of *cross-lingual* transfer, enabling high-quality predictions in different languages (Chun-ting Ho, Justing & Chan, Chung-hong, 2023). Secondly, researchers may opt to train classifiers separately for the different corpora under study, using *monolingual* corpora (e.g., see Rust et al., 2021). While monolingual models generally do not transfer well to other linguistic contexts, they offer specific benefits, such as optimal performance on particular languages due to optimization and language complexity handling.

However, the field of political communication lacks comprehensive research comparing the benefits and drawbacks of the use of various multilingual and monolingual approaches (see also Chun-ting Ho, Justing & Chan, Chung-hong, 2023; Lind et al., 2021).

Even less is known about which classification technique works best when researchers aim to investigate transitions between different political domains, such as examining inter-agenda dynamics between parliamentary questions, media and party manifestos. This issue is pertinent because the text used to train a model contains both linguistic *and* contextual information (Chun-ting Ho, Justing & Chan, Chung-hong, 2023). While recent developments in large language models (LLMs) show promise for multilingual text analysis, in particular, due to the introduction of Multilingual Sentence Embeddings (MSE) and the fine-tuning framework, it remains unclear whether multilingual text analysis benefits more from these techniques in multilingual or monolingual settings.

The study aims to determine the effectiveness of these techniques when dealing with scenarios involving multiple domains and languages, in comparison to traditional supervised machine learning approaches. By evaluating the performance of multilingual and monolingual approaches, the study seeks to provide insights into the most suitable techniques for effectively tackling the challenges of multilingual and cross-lingual topic classification. In that sense, it offers an extension of the work reported in OPTED deliverable 4.5, where unsupervised, semi-supervised and supervised topic coding applications are compared across different types of documents, but within a single language domain. As deliverable 4.6 demonstrates, logic and patterns of issue salience differ considerably across different domains. Additionally our study serves as an exemplary case as how different content analytical data sources can be linked and combined to provide more extensive analyses and ultimately address substantial research questions that have been unanswered so far.

More specifically, the current study makes several significant contributions. Firstly, it leverages a unique and extensive multilingual dataset with manual annotations. In particular, it explores the untapped benefits of combining the *Comparative Agendas Project* (CAP) and *Comparative Manifesto Project* (CMP) datasets based on overlapping topic categories, creating a more comprehensive training set for machine learning models that spans lingual and content domains. This integration allows for extensive and valid testing of hypotheses in a wide range of contexts. By expanding the usefulness of these datasets for cross-comparative research, the study underscores the importance of enhancing the availability and usability of annotated data for researchers in comparative politics. It furthermore demonstrates the effectiveness of some of the tested classifiers to study overtime attention to policy issues in the *Guardian* and the UK Parliament.

Moreover, the study not only acknowledges the limitations of standalone monolingual and monodomain models but also goes beyond by providing tangible solutions that researchers can implement. Despite the wide use and application of these models in computational social and communication science, their performance may fall short in certain contexts or when tackling novel research questions. The study reveals effective strategies for leveraging these datasets and identifies techniques that can be applied to successfully address and overcome these challenges.

Comparative Politics and Topic Classification

The field of comparative politics pays considerable attention to the flow of attention to policy topics across different domains and countries (e.g., Baumgartner et al., 2006; Eissler et al., 2014; Vliegenthart et al., 2013). Scholars have dedicated significant efforts to systematically monitor and measure this attention across domains. This has led to coordinated initiatives that aim to track policy domains in a comprehensive manner. One notable project in this regard is the Comparative Agendas Project, which involves extensive country-specific annotation of a wide variety of political documents - including but not limited to party manifestoes - for the presence of topics. Within this project, 21 major policy topics and 200 subtopics have been identified and analyzed (Baumgartner et al., 2013; Bevan, 2017). This rich dataset allows researchers to explore the patterns of attention to policy issues across countries and over time.

Another prominent research endeavor is the Comparative Manifesto Project (CMP) (Lehmann et al., 2023). The CMP is a long-standing research initiative that seeks to analyze and compare political party manifestos from countries worldwide. By employing a systematic and quantitative approach, the project sheds light on the ideologies, policy positions, and evolution of political parties across different countries and time periods.

Both the CAP and the CMP have made significant contributions to the field of political science. They have deepened our understanding of various aspects of party politics, including ideological shifts, (changing) policy preferences, and the dynamics of attention to policy topics. These projects provide valuable datasets that researchers can freely access, enabling further analysis and comparative studies in the field.

Given that the coding of policy topics is expensive and time-consuming, efforts have been made in the past decade to automate the coding process (Albaugh et al., 2014; Burscher et al., 2015; Karan et al., 2016; Sebők & Kacsuk, 2021). Scholars have experimented with various techniques, including dictionary-based approaches and supervised methods.

One approach that has been explored is the use of dictionary-based methods, where a predefined list of relevant terms and keywords indicative of specific policy domains is created (Albaugh et al., 2014). This dictionary enables researchers to automate the identification and categorization of policy topics within texts.

Supervised techniques have also been employed in the automated coding of policy topics (Burscher et al., 2015; Karan et al., 2016; Sebők & Kacsuk, 2021). These techniques involve training machine learning algorithms on annotated data, allowing them to learn patterns and make predictions regarding the categorization of policy topics in new unannotated texts.

The advancement of automated coding techniques offers the potential for increased efficiency and cost-effectiveness in analyzing policy topics. However, this progress also brings about unique challenges that need to be addressed. One crucial challenge is ensuring that classifiers can accurately detect policy issues across different linguistic domains, enabling cross-country comparative research.

In addition, it is essential for automated classifiers to achieve comparable levels of accuracy when identifying policy issues across diverse languages. This is particularly important as the Comparative Agendas Project (CAP) and Comparative Manifesto Project (CMP) datasets cover a wide range of content domains. These domains include formal texts such as speeches from political figures, parliamentary questions, media texts, and party's press releases or other types of communication. Given the significant variation in content domains, it raises questions about the extent to which new automated techniques can effectively capture and analyze this diverse range of textual sources (see Kroon et al., 2022). Together, it is crucial to address these questions and ensure that the automated methods can effectively handle the breadth and depth of the content present in datasets like CAP and CMP.

Multilingual and Cross-Domain Challenges

When addressing multilingual classification challenges, researchers must decide how to curate their training dataset. Broadly speaking, they can decide to manage training data either on a per-language basis (i.e., *monolingual*) or by combining data from different languages (i.e., *multilingual*). These multilingual and monolingual training sets subsequently serve as the foundation for a diverse array of classification techniques. We will discuss these techniques next.

Established Approaches

Traditional approaches to automated topic classification have relied on the use of bag-of-words representations of text. As the term suggests, these approaches treat textual data as a collection of individual words, disregarding their order or context. These techniques have been widely used in lexicon-based methods and supervised machine learning algorithms. However, a major drawback of these approaches is their limited consideration of semantic context. Words with multiple meanings, such as "light" or "crane", as well as synonyms, are not adequately captured by these models. This lack of contextual understanding can hinder the accuracy and effectiveness of topic classification systems (Boukes et al., 2020; Guo et al., 2016; Kroon et al., 2022; Rudkowsky et al., 2018)

Conventional supervised techniques for measuring topics using bag-of-words representations (BoW) have inherent limitations, in particular when the research question addresses a cross-national question. First, these models are prone to challenges related to language transferability. To avoid duplicating research efforts, researchers often incorporate a translation step when working with languages under study (Licht, 2023). This process comes with the risk of losing the full subtleties and cultural nuances of language, which poses a significant challenge when addressing research questions that extend across geographic borders.

Second, BoW-based approaches have a limited ability to capture latent constructs when different terminologies, jargon, and contextual understandings are employed across different domains (Burscher et al., 2015; Osnabrügge et al., 2023). For example, such models may find it particularly difficult to classify news articles when trained on parliamentary texts, and visa versa. This poses a significant challenge when applying the models outside the specific context on which they are trained.

Third, challenges related to domain transferability may become amplified in a cross-lingual setting. The complexities of transferring knowledge across different domains are further complicated by the additional layer of language variation (Sánchez et al., 2022).

Advanced Approaches

To explore and overcome these limitations, the current study investigates alternative approaches, specifically *multilingual sentence embedding* (MSE) and *fine-tuning* (FT) of pre-trained multilingual and monolingual language models. These techniques leverage the power of pre-trained large language models (LLMs) to enhance the model's contextual and semantic understanding, bypassing the limitations of traditional bag-of-words (BoW) approaches.

LLMs have transformed the capabilities of computational social scientists by leveraging pretraining on vast multilingual data. Prominent examples of these models include BERT (Devlin et al., 2019) and GPT (Radford et al., 2019), which have recently revolutionized the field of (political) communication science (Bestvater & Monroe, 2022; Laurer et al., 2023; Licht, 2023; Lin et al., 2023; Viehmann et al., 2022; Widmann & Wich, 2022) and extended their impact to other domains (Devlin et al., 2019). By explicitly considering the contextual meaning of language, these models have achieved remarkable advancements in performance, providing a breakthrough for researchers in their understanding and analysis of language (Acheampong et al., 2021).

By leveraging these pre-trained models, MSE techniques enable analyses in multiple languages by encoding sentences from various languages into a shared embedding space. This shared representation facilitates cross-lingual comparisons and classification, as semantic relationships can be explored and used across different languages. Recent work suggests that MSE is a more effective technique compared to BoW-based approaches when modelling multilingual party manifestos (Licht, 2023). MSE models go beyond traditional Bag-of-Words (BoW) approaches by representing words and sentences in a higher-dimensional space, where the distance between vectors reflects their semantic similarity.

The Fine-Tuning (FT) framework involves the process of fine-tuning a pre-trained multilingual language model on annotated data specific to the task at hand (Laurer et al., 2023; Lin et al., 2023; Viehmann et al., 2022; Widmann & Wich, 2022). This approach leverages the power of this type of model by taking a pre-trained language model and adapting it to data that are language or content specific, allowing the model to adjust to the specific linguistic characteristics and nuances of a particular task or domain (Laurer et al., 2023; Lin et al., 2023). This fine-tuning process further enhances the model's ability to capture language semantics and improve performance in various multilingual and monolingual applications.

In the FT framework, one can choose to fine-tune a multilingual or monolingual pretrained model. For example, several multilingual Language Learning Models (LLMs) like mBert, XLM-R, and MT5 (Conneau et al., 2020; Devlin et al., 2019) exist alongside monolingual models (e.g., BERT, Bertje). While monolingual pre-trained models are less suited to handle cross-lingual transferability, as they learn from only a single linguistic domain, they are still often used due to the assumption that multilingual models suffer from the *curse of multilinguality* (Conneau et al., 2020). As multilingual models may not represent all languages equally (Rust et al., 2021) they might perform less well in monolingual settings (Ronnqvist et al., n.d.; Virtanen et al., 2019). The extent to which multilingual or monolingual LLMS work better in the context of topic classification in political texts, remains unclear. Thus, we ask:

> **RQ1**: When performing a multilingual analysis, is it more effective to use monolingual models (models trained on individual languages separately) or to combine all available training data and use a multilingual supervised approach?
>
> **RQ2:** Can context-aware approaches, such as *Fine-Tuning (FT)* and *Multilingual Sentence Embedding (MSE)*, outperform baseline models (monolingual Bag of Words-based models) in terms of classification accuracy for both multilingual and monolingual tasks?

Social domain Shifts

BoW models are often limited by their inability to handle out-of-vocabulary words, and they struggle to capture the contextual nuances and semantic relationships between words (Kroon et al., 2022; Rudkowsky et al., 2018). Consequently, when faced with out-of-domain data or new vocabulary, their performance tends to degrade significantly. In contrast, FT and MSE approaches leverage contextual embeddings and semantic representations, which enable them to handle unseen words or domains more effectively.

By learning contextualized representations and capturing the semantic similarities between sentences, techniques that leverage LLMs – such as FT and MSE models – are arguably more robust to domain shifts, e.g., trying to classify parliamentary questions using a classifier trained on news media articles.

In particular, one the key advantages of the FT framework is its ability to learn from task-specific data while incorporating domain-specific knowledge. By updating the pre-trained model's parameters during fine-tuning, the model can quickly adapt and specialize in domain-specific features and nuances relevant to the task at hand (Widmann & Wich, 2022). This targeted learning enables FT to achieve higher performance with a smaller amount of task-specific data compared to training models from scratch (Laurer et al., 2023).

Moreover, FT enables knowledge transfer to specific tasks. The pre-trained language models have already learned comprehensive language representations and patterns from large-scale monolingual or multilingual data corpora (Devlin et al., 2019). By fine-tuning these models, the domain-specific data can be effectively incorporated, allowing the model to benefit from both the general language knowledge and the specific task information. This transfer of knowledge enhances the efficiency and effectiveness of the fine-tuning process.

We ask:

> **RQ3:** Do context-aware approaches, specifically *Fine-Tuning (FT)* and *Multilingual Sentence Embedding (MSE)*, mitigate performance degradation when transitioning from in-domain to out-domain predictions, in comparison to Bag of Words (BoW) models, in both multilingual and monolingual classification tasks?

Methods

Data

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) and the *Comparative*

Manifesto Project (CMP). We focus our analysis on five languages: English, Dutch, Spanish, Hungarian and German. Specifically, we rely on the *Comparative Agendas Project* and *Comparative Manifesto Project* to select datasets from corresponding countries. See Table 1 for an overview of the origin of the datasets used.

Table 1
Data Sources

Language	Domain	Dataset URL	Source
Dutch	Media	https://surfdrive.surf.nl/files/index.php/s/HIW4nnOJJPBoGZD	Burscher et al., 2015
Dutch	Parliamentary Questions	10._PA_OQ_COMPLETE_1984-2009_CW_UNI.xls	CAP
Dutch	Parliamentary Questions	https://surfdrive.surf.nl/files/index.php/s/cPsqhgtGOy7mIdd	Burscher et al., 2015; Kroon et al., 2022
Dutch	Party Manifesto’s	CMP API MPDS2022a	CMP
Dutch	Party Manifesto’s	CMP API MPDS2022a	CMP
English	Media	uk_media.csv	CAP
English	Parliamentary Questions	uk_pmqs.csv	CAP
English	Party Manifesto’s	CMP API MPDS2022a	CMP
English	Party Manifesto’s	CMP API MPDS2022a	CMP
German	Media	switzerland_media.csv	CAP
German	Parliamentary Questions	anfrage_1976-2005_website-release_2.5.csv	CAP
German	Party Manifesto’s	CMP API MPDS2022a	CMP
German	Party Manifesto’s	CMP API MPDS2022a	CMP
Hungarian	Media	hungary_media_magyarnemzet.csv	CAP
Hungarian	Media	hungary_medianepszab_1990_2014_1.csv	CAP
Hungarian	Parliamentary Questions	CAP_-_Oral_Questions.csv	CAP
Hungarian	Parliamentary Questions	CAP_-_Urgent_Questions_Hungary_3.csv	CAP
Hungarian	Party Manifesto’s	CMP API MPDS2022a	CMP
Hungarian	Party Manifesto’s	CMP API MPDS2022a	CMP
Spanish	Media	Media_El_Mundo_Web_CAP_csv.csv	CAP
Spanish	Media	Media_El_Pas_Web_CAP_csv.csv	CAP
Spanish	Parliamentary Questions	Spain_OralQuestions19772019_19.1.csv	CAP
Spanish	Party Manifesto’s	CMP API MPDS2022a	CMP
Spanish	Party Manifesto’s	CMP API MPDS2022a	CMP

We selected datasets from the following content domains: *parliamentary questions*, *news media*, and *party manifestos* from the following lingual domains: *Netherlands*, *Spain*, *Hungary*, *Germany* and *UK*. Entries with missing labels are excluded from the analysis ($N_{totalsample}$ = 621903).

Combining CAP and CMP

To comprehensively test our hypotheses across various domains, we merged the CAP and CMP datasets. The CMP codebook includes codes for both policy topics and stance, while the focus of CAP is solely on policy topics. To combine the annotations from both domains, we selected policy topics from CMP that aligned with the domains assessed in CAP, disregarding the specific valence assigned to each topic in the CMP coding scheme.

A careful review of the codebooks of CMP and CAP allowed us to identify policy topics that were conceptually aligned between the two projects. Based on this review, we selected the following policy topics: *Environment*, *Culture*, *Civil Rights*, *Education*, and *Immigration*.

From the CMP codebook, we extracted the scores corresponding to the identified policy topic codes, focusing exclusively on these policy topics and disregarding the specific valence assigned to each score in CMP.

By aligning the selected policy topics from CMP with the cultural domains assessed in CAP based on their substantive alignment, we successfully combined the scores from both projects. This approach facilitated a meaningful integration of the datasets, providing a comprehensive assessment of the policy topics within the cultural context. We considered the overall stance of each topic rather than the specific valence, resulting in combined scores that served as valuable indicators for evaluating and comparing entities or regions based on their stance on the aligned policy topics of environment, culture, civil rights, education, and immigration.

Linking procedure

For the linking procedure, we borrow from the Linkage tool as detailed in deliverable 8.4 of the OPTED project. More specifically, this tool facilitated the automated capture of media content, using CMP's API as well as CAP's online registry of datasets. Additionally, it facilitates the integration and linking of textual data from different sources. We used the tool for aggregating the data to different levels of granularity. This facilitated the smooth handling of the datasets.

For the combined analysis, we utilized the following topic codes from the original CMP codebook: *environment* (code: 501), *culture* (code: 502), *civil rights* (codes: 201.1 and 201.2), *education* (codes: 506 and 507), and *immigration* (codes: 601.2 and 602.2). Additionally, we incorporated the corresponding scores from the CAP codebook: *environment* (score: 7.0), *civil rights* (score: 2.0), *culture* (score: 23.0), *education* (score: 6.0), and *immigration* (score: 9.0). All other topics were classified as *other*.

Sampling procedure

We used a multi-step stratified sampling approach to arrive a balanced samples for each language * domain combination. We aim to arrive at representative training and testing samples while mitigating issues that may arise due to the high-class imbalance in the selected datasets. The procedure unfolds as follows:

First, the dataset is categorized based on the attributes *language* and *domain*. From these categories, random samples are drawn, limited to a maximum of 9000 samples per group.

Subsequently, and to deal with challenges posed by the disproportionately represented *other* class, a downsampling procedure is applied–exclusively to the *training* dataset. This equips our training datasets to display a more even distribution of classes, while the test dataset retains the original imbalanced class distribution, thereby offering a realistic representation of the classification challenge as encountered in real-world scenarios. In order to ensure a fair comparison, we opt to randomly select 601 samples from each combination of language and content domain. Table 2 summarizes the final data samples.

Classification

Our analyses center on the multi-class classification problem of six major topics—*civil rights, education, environment, culture* 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_{Final} = 56198). Ultimately, the data represents a highly imbalanced multi-class classification problem.

Train-Val-Test

The data was split into separate training and test sets for each lingual * content domain combination. For the BoW approach–which is considered the baseline–we train monolingual models. For the BoW, BoW MT, and MSE approach, we use 5-fold cross-validation. For the FT approach, 20% of the training set is used for validation. For

Table 2

Descriptive Statistics of Document Sentences

		Total Documents	Average Tokens	
		N	M	SD
Language	Domain			
Dutch	Party Manifestos	601	16.49	7.77
	Media	601	262.03	312.69
	Parliamentary Questions	601	263.37	169.56
English	Party Manifestos	601	17.92	10.50
	Media	601	6.99	2.93
	Parliamentary Questions	601	89.91	41.75
German	Party Manifestos	601	16.69	8.28
	Media	601	4.93	1.45
	Parliamentary Questions	601	8.46	5.37
Hungarian	Party Manifestos	601	13.49	8.13
	Media	601	266.61	348.52
	Parliamentary Questions	601	9.95	5.53
Spanish	Party Manifestos	601	26.89	16.24
	Media	601	10.74	4.73
	Parliamentary Questions	601	27.99	11.39

the MSE and FT approach, we combine training data for all countries, split by domain. Subsequently, we test the performance on fully held-out samples for the separate country and domain combinations.

Multilingual Models

In the multilingual approach, we aggregate all training data from the five languages within each social domain.

Monolingual Models

In the monolingual approach, we exclusively train and evaluate models using corpora from a single language.

Bag of Words (BoW)

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*). For the *monolingual approach*, we trained models separately for each linguistic and content domain. In contrast, for the *multilingual approach*, we aggregated content from different linguistic domains and trained models based on content domains.

Machine Translated Bag of Words (BoW MT)

In this approach, we used machine-translated bag of words. Specifically, we employed an open-source machine translation model developed by the Helsinki-NLP research group (Tiedemann & Thottingal, 2020). We translated all features into English and subsequently evaluated performance using the same set of classifiers used for the BoW approach. We proceeded with the best-performing model.

MSE Approach

We used *distiluse-base-multilingual-cased-v2* (Reimers & Gurevych, 2019) as a Multilingual Sentence Embedding vectorizer. We vectorized our text using this pre-trained model and subsequently used the same set of classifier options as we used for the BoW approach. We continued with the best-performing model. For the *monolingual approach*, we trained models separately for each linguistic and content domain. Similarly, for the *multilingual approach,* we combined content from different linguistic domains and trained models based on content domains.

Fine-Tuning Approach

To estimate performance per content domain, we compared these results with the effectiveness of a transformer-based classifier fine-tuned on open-source large language models. Specifically, for the *monolingual approach*, we trained models separately for each linguistic and content domain. We selected specific monolingual Large Language Models (LLMs) for each linguistic domain: for Hungarian; *hubert-base-cc* (Nemeskey, 2021), for Dutch *robbert-2022-dutch-base* (Delobelle et al., 2020, 2022), for English *bert-base-uncased* (Devlin et al., 2018), for Spanish *bert-base-spanish-wwm-cased* (Cañete et al., 2020), and for German *bert-base-german-cased*. For our *multilingual approach,* we combined all training data and used a multilingual LLM, a smaller version of Roberta (*xlm-roberta-comet-small*).

The classifiers were optimized towards Macro F1, in order to give relatively more weight to the minority classes. We iterated over a learning rate range of 3e-5 to 12-6, warm-up steps ranging from 0 to 100, and a maximum of 20 epochs.

Baseline

To assess the effectiveness of the various approaches, we use the best-performing classifier trained with the BoW approach on monolingual data as our baseline.

Results

Performance Within the Social Domain: Baseline, Monolingual, and Multilingual Models

The first research question (RQ1) explores the effectiveness of using either monolingual or multilingual techniques for classifying policy topics in different languages.

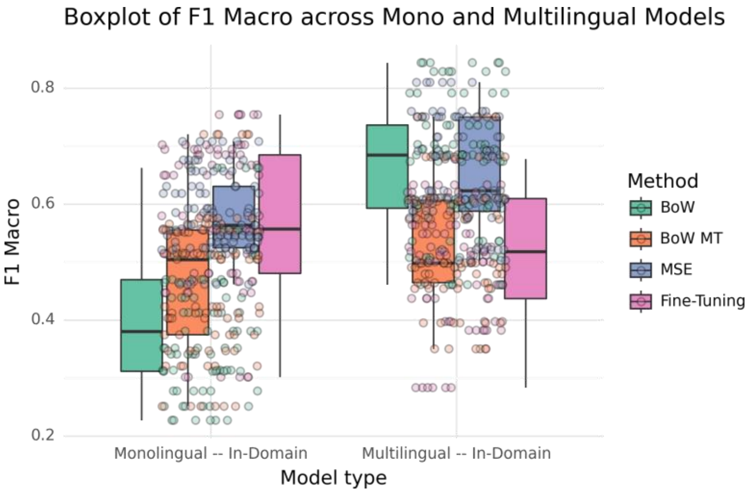


Figure 1

Macro F1 Scores for different Techniques across Multilingual and Monolingual Model

Figure 1 illustrates the performance of various techniques. It is evident that there is variation in performance across techniques in both monolingual and multilingual corpora. Specifically, BoW approaches do not perform well in monolingual settings but exhibit improved performance when combining all training data. Conversely, FT works exceptionally well in monolingual domains.

Figure 2 demonstrates variations in performance across different languages. As observed, slight variations exist in how well different techniques predict performance across countries. Within the monolingual models, it is noticeable that the performance of the German language model increases significantly when comparing the BoW model to the fine-tuned approach.

Table 3 summarizes all comparisons between baseline, monolingual, and multilingual models. In all cases, we find that the combination of monolingual and multilingual techniques outperformed the baseline models.

Table 3

Comparing Baseline, Monolingual, and Multilingual Models – Within Domain Performance

		Baseline		Monolingual		Multilingual	
Test language	Target Domain	Method	F1-Macro	Method	F1-Macro	Method	F1-Macro
Dutch	Media	BoW	0.47	Fine-Tuning	0.65	BoW	**0.68**
Dutch	Parl Questions	BoW	0.66	BoW MT	0.72	BoW	**0.83**
Dutch	Party Manifestos	BoW	0.31	MSE	0.56	MSE	**0.62**
English	Media	BoW	0.25	Fine-Tuning	**0.69**	Fine-Tuning	0.52
English	Parl Questions	BoW	0.51	MSE	0.63	BoW	**0.84**
English	Party Manifestos	BoW	0.56	MSE	0.71	MSE	**0.76**
German	Media	BoW	0.23	Fine-Tuning	**0.75**	Fine-Tuning	0.68
German	Parl Questions	BoW	0.35	Fine-Tuning	0.70	BoW	**0.79**
German	Party Manifestos	BoW	0.28	MSE	0.66	MSE	**0.72**
Hungarian	Media	BoW	0.46	Fine-Tuning	0.55	BoW	**0.69**
Hungarian	Parl Questions	BoW	0.31	Fine-Tuning	0.56	BoW	**0.61**
Hungarian	Party Manifestos	BoW	0.34	Fine-Tuning	0.55	MSE	**0.69**
Spanish	Media	BoW	0.38	Fine-Tuning	**0.69**	Fine-Tuning	0.63
Spanish	Parl Questions	BoW	0.44	Fine-Tuning	0.67	BoW	**0.67**
Spanish	Party Manifestos	BoW	0.41	MSE	0.55	MSE	**0.59**

The results indicate that, in most cases, the multilingual approach (which combines all training data) is the most efficient choice. In a few cases, monolingual approaches proved to be the best-performing option. Therefore, in response to RQ1, we can conclude that classification based on multilingual corpora leads in most cases to the best performance.

Next, and in response to the question (RQ2) of whether context-aware approaches, such as *Fine-Tuning (FT)* and *Multilingual Sentence Embedding (MSE)*, can outperform baseline models (monolingual Bag of Words-based models) in terms of classification accuracy, we further inspect differences in performance across different classification techniques.

As displayed in Table 3, we find that within the selection of multilingual models, traditional bag-of-words models outperformed more advanced techniques, such as fine-tuning with multilingual pre-trained models. Additionally, results indicate that MSE performed well on multilingual corpera, particularly in the classification of party manifestos.

Moving to the selection of models trained on monolingual corpora, we find that the fine-tuning framework works best, and systematically outperformed baseline (BoW) models. More in particular, for news media texts in Spanish, German, and English, the best performance was achieved when fine-tuning monolingual language models.

Together, and in response to RQ2, we find support for higher performance gains when using monolingual corpora, but not per definition when the training sample is large and combines multilingual variation. Hence, when combining training data from different lingual domains, BoW proved a good strategy.

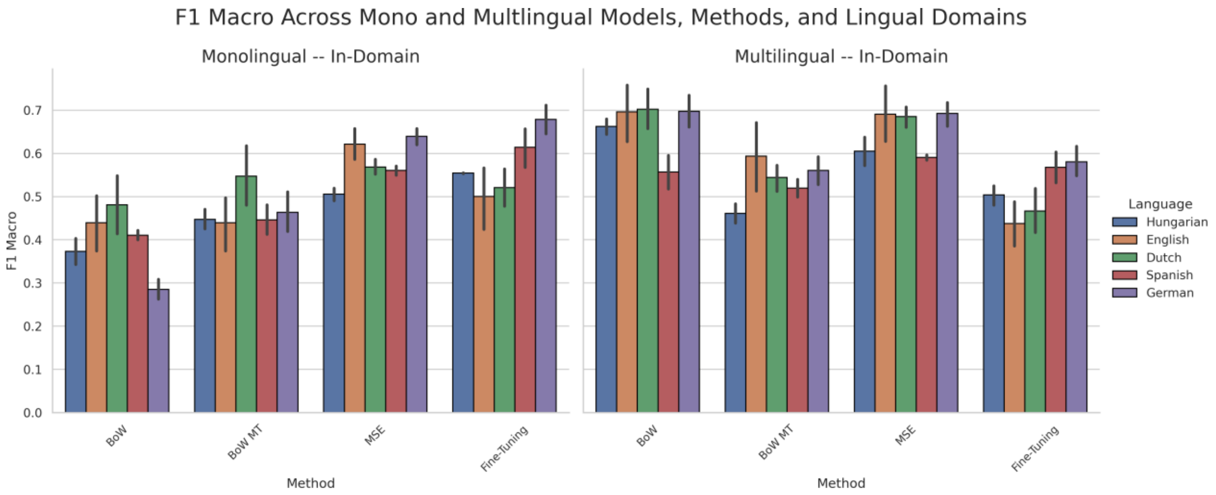


Figure 2
Macro F1 Scores per Lingual Domain across different Techniques and Multilingual and Monolingual Models

Table 4

Comparing Baseline, Monolingual, and Multilingual Models – Out of Domain Performance

			Baseline		Monolingual		Multilingual	
Test language	Train Domain	Target Domain	Method	F1-Macro	Method	F1-Macro	Method	F1-Macro
Dutch	Media	Parl Questions	BoW	0.40	MSE	0.52	MSE	**0.67**
Dutch	Media	Party Manifestos	BoW	0.18	Fine-Tuning	**0.47**	MSE	0.41
Dutch	Parl Questions	Media	BoW	0.36	Fine-Tuning	**0.65**	Fine-Tuning	0.62
Dutch	Parl Questions	Party Manifestos	BoW	0.22	Fine-Tuning	**0.47**	BoW MT	0.46
Dutch	Party Manifestos	Media	BoW	0.23	Fine-Tuning	**0.65**	Fine-Tuning	0.62
Dutch	Party Manifestos	Parl Questions	BoW	0.30	Fine-Tuning	**0.45**	MSE	0.43
English	Media	Parl Questions	BoW	0.22	MSE	0.41	MSE	**0.58**
English	Media	Party Manifestos	BoW	0.28	Fine-Tuning	0.51	MSE	**0.62**
English	Parl Questions	Media	BoW	0.33	Fine-Tuning	**0.69**	Fine-Tuning	0.52
English	Parl Questions	Party Manifestos	BoW	0.48	MSE	0.53	MSE	**0.63**
English	Party Manifestos	Media	BoW	0.23	Fine-Tuning	**0.69**	Fine-Tuning	0.52
English	Party Manifestos	Parl Questions	BoW	0.33	MSE	**0.54**	MSE	0.46
German	Media	Parl Questions	BoW	0.28	Fine-Tuning	**0.70**	MSE	0.61
German	Media	Party Manifestos	BoW	0.14	Fine-Tuning	**0.58**	Fine-Tuning	0.50
German	Parl Questions	Media	BoW	0.15	Fine-Tuning	**0.75**	Fine-Tuning	0.68
German	Parl Questions	Party Manifestos	BoW	0.15	Fine-Tuning	**0.58**	MSE	0.52
German	Party Manifestos	Media	BoW	0.14	Fine-Tuning	**0.75**	Fine-Tuning	0.68
German	Party Manifestos	Parl Questions	BoW	0.23	Fine-Tuning	**0.70**	MSE	0.60
Hungarian	Media	Parl Questions	BoW	0.21	Fine-Tuning	**0.56**	Fine-Tuning	0.55
Hungarian	Media	Party Manifestos	BoW	0.14	Fine-Tuning	**0.55**	MSE	0.48
Hungarian	Parl Questions	Media	BoW	0.23	MSE	**0.56**	Fine-Tuning	0.53
Hungarian	Parl Questions	Party Manifestos	BoW	0.27	Fine-Tuning	**0.55**	MSE	0.49
Hungarian	Party Manifestos	Media	BoW	0.17	Fine-Tuning	**0.55**	Fine-Tuning	0.53
Hungarian	Party Manifestos	Parl Questions	BoW	0.28	Fine-Tuning	0.56	MSE	**0.59**
Spanish	Media	Parl Questions	BoW	0.28	Fine-Tuning	**0.67**	Fine-Tuning	0.61
Spanish	Media	Party Manifestos	BoW	0.21	Fine-Tuning	0.48	MSE	**0.49**
Spanish	Parl Questions	Media	BoW	0.25	Fine-Tuning	**0.69**	Fine-Tuning	0.63
Spanish	Parl Questions	Party Manifestos	BoW	0.26	Fine-Tuning	**0.48**	MSE	0.46
Spanish	Party Manifestos	Media	BoW	0.15	Fine-Tuning	**0.69**	Fine-Tuning	0.63
Spanish	Party Manifestos	Parl Questions	BoW	0.24	Fine-Tuning	**0.67**	Fine-Tuning	0.61

Performance Outside the Social Domain: Baseline, Monolingual, and Multilingual Models

Next, we inspect the performance of the different models beyond the context in which they are trained on. RQ3 asked if context-aware approaches, specifically *Fine-Tuning (FT)* and *Multilingual Sentence Embedding (MSE)*, mitigate performance degradation when transitioning from in-domain to out-domain predictions, in comparison to Bag of Words (BoW) models, for both multilingual and monolingual classification tasks.

Table **??** displays the out of domain performance of the monolingual and multilingual approaches, in comparison to the baseline model–representing the out of domain performance of traditional classifiers trained on BoW representations. As can be seen, monolingual models that use the FT framework, are most successful in maintaining performance when transitioning to other domains. In the majority of transitions, these models are the best-performing option, though in several instances scores remain on the low side.

As can be seen in Figure 4a, the performance of classifiers trained on BoW-representations, decreases substantially when transitioning to different topical domains, both for monolingual and multilingual models. In **??**, BoW-based multilingual models are no longer the best choice. Rather, multilingual models using MSE representations seem to experience relatively little performance loss. Consequently, these are the preferred option for a set of models.

With regards to RQ3, we can conclude that indeed, context-aware models are better equipped in handling out of domain shifts.

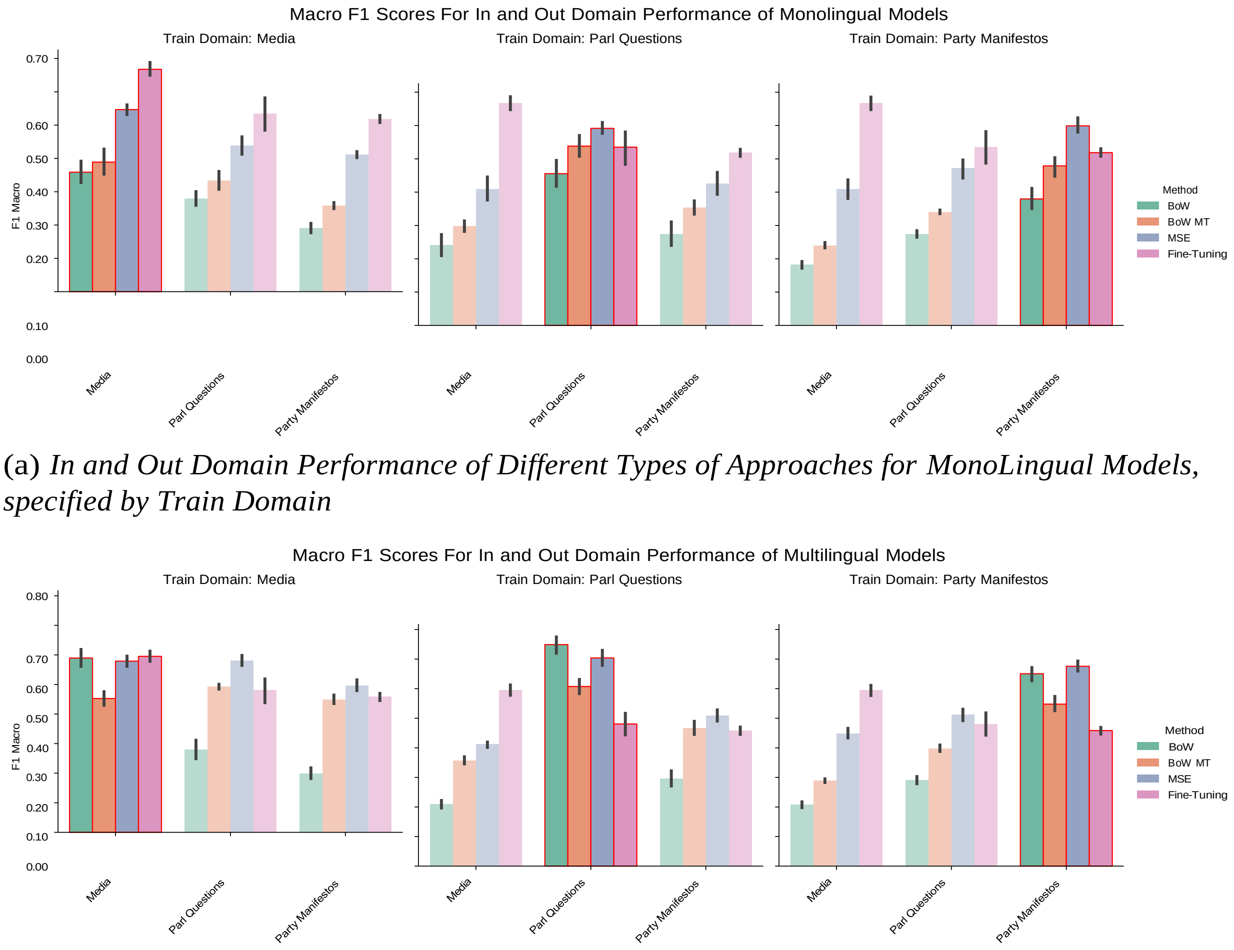
Performance degradation across social domains

We investigate how the performance degradation varies across methods for different domains. An overview is provided in Table 5. Figure 3a summarizes the in and out of domain performance across different techniques for monolingual models, while Figure 3b does the same for multilingual models.

Table 5
Average Performance Decay by Domain Transition

Domain Transition	Average Decay
Media -> Parl Questions	-0.094
Media -> Party Manifestos	-0.149
Parl Questions -> Media	-0.184
Parl Questions -> Party Manifestos	-0.170
Party Manifestos -> Media	-0.159
Party Manifestos -> Parl Questions	-0.126

First, concerning transitions from the e.g., *news media* domain, we observe a moderate average performance degradation (-0.094) when shifting to *parliamentary questions*. The performance degradation is more noticeable when transitioning to *party manifestos* (-0.149). Consequently, moving from the *news media* domain to *party*



(a) *In and Out Domain Performance of Different Types of Approaches for MonoLingual Models, specified by Train Domain*

(b) *In and Out Domain Performance of Different Types of Approaches for MultiLingual Models, specified by Train Domain*

Figure 3

Comparison of In and Out Domain Performance for MonoLingual and MultiLingual Models

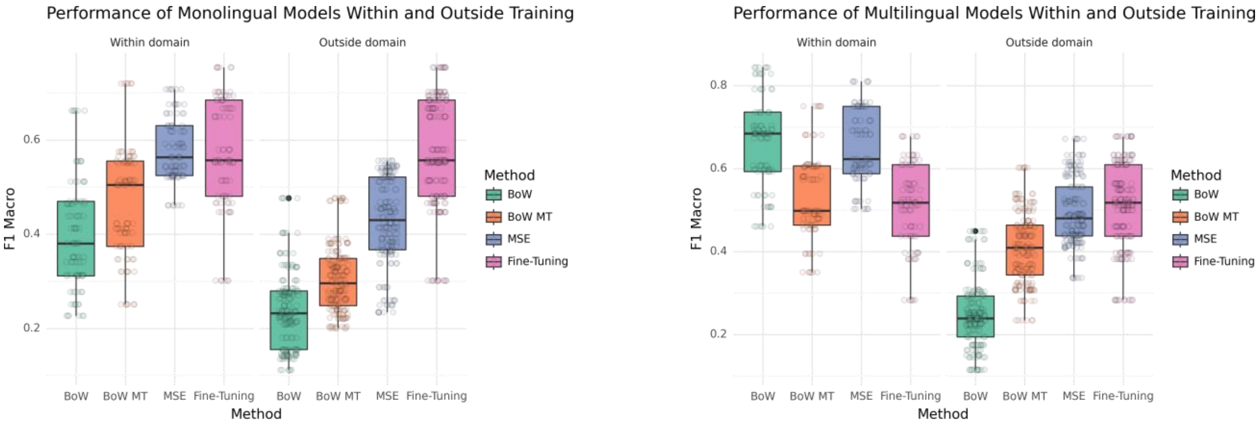
manifestos results in a more significant drop in performance compared to transitioning to *parliamentary questions*.

Secondly, with respect to transitions from the *parliamentary questions* domain, we find a substantial performance degradation (-0.184) when shifting to the *news media* domain. When moving to *party manifestos*, we also observe a significant performance degradation (-0.170). Thus, transitioning from *parliamentary questions* to the *news media* domain leads to a more significant decline in performance than transitioning to *party manifestos*.

Thirdly, we examine transitions from the *party manifestos* domain. Here, we find a notable performance degradation (-0.159) when moving to the *news media* domain and a moderate performance degradation (-0.126) when transitioning to *parliamentary questions*. Consequently, transitioning to the *news media* domain results in a more pronounced drop compared to transitioning to *parliamentary questions*.

Overall, the transition from the *parliamentary questions* domain tends to have the most substantial negative impact on model performance, followed by transitions involving the *party manifestos* domain. Transitioning from the *news media* domain generally leads to less severe performance decay when compared to the other domains.

These variations in performance decay underscore that, even though the data from *party manifestos* originates from a different source (namely CMP) compared to *parliamentary questions* and *news media* (namely CAP), transitions between both domains are feasible. As transitions between the CMP and CAP domains, though with some performance degradation, are possible—especially when using the FT framework with monolingual corpora—it suggests that these sources likely share common constructs or underlying patterns. While differences may emerge when transitioning to different domains, the ability to link CMP and CAP indicates the potential for some level of continuity or overlap in the data they represent. Hence, it is possible to link these two seemingly distinct data sources, which increases their potential for cross-comparative research.



(a) *MonoLingual Models*

(b) *MultiLingual Models*

Figure 4

In and Out Domain Performance of Different Types of Approaches for Monolingual and Multilingual models

Application: The Case of the United Kingdom

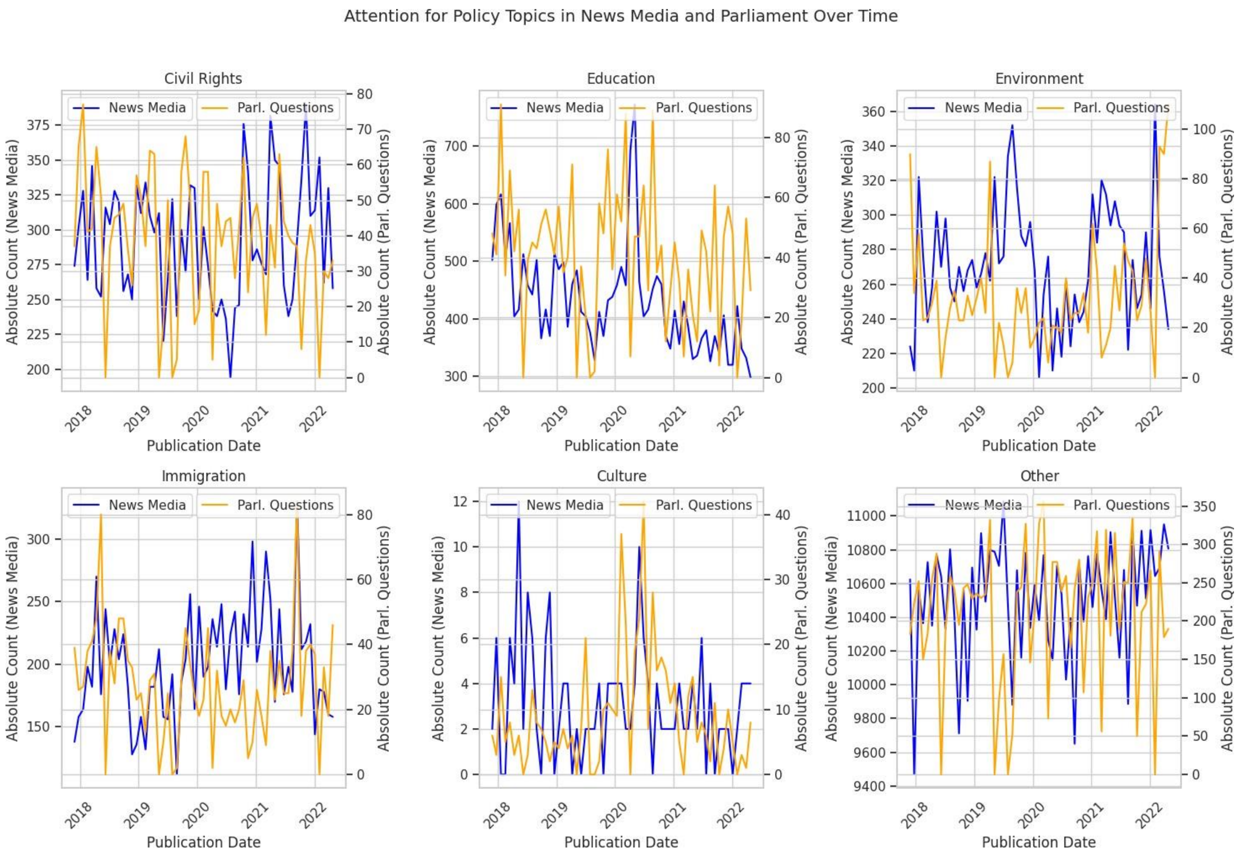


Figure 5

Predicting Topic Attention in News Media and Parliamentary Questions in the UK

In a final step, we explore the effectiveness of the classifiers in the case of studying overtime attention to policy issues on different agendas. More specifically, we collected a comprehensive dataset consisting of all news articles published in The Guardian from 2014 to 2022 (N=1,047,532) and all parliamentary questions raised in the House of Commons (N=35,295).

Drawing from the conclusions presented in this paper, we chose models that demonstrate robust out-of-domain performance to account for potential variations in linguistic patterns over time and across domains. To achieve this, we employ the fine-tuned monolingual BERT model, fine-tuned on annotated media data, for analyzing The Guardian dataset. Additionally, we use the fine-tuned monolingual BERT model, trained on annotated parliamentary data, for the analysis of parliamentary questions.

Descriptive results show that in news media, the large majority of news articles do not pay attention to the substantive topics studied here (other: 89.8%, *n*= 940855). Of the topics, most attention is paid to the topic education (3.7%,38865), followed by civil rights (2.5%, *n*=26367), environment (2.246996, *n*=23538), immigration (1.7%, *n*=17644) and culture (0.02%, *n*=263).

In the political domain, a relatively higher proportion of parliamentary questions could be attributed to the substantive topics, with the "other" category representing 58.6% (n=940,855) of the dataset. Among these topics, education takes the lead with 11.4% (n=20,683), followed by civil rights at 11.2% (n=3,970), environment at 8.3% (n=2,960), immigration at 8.2% (n=2,905), and culture at 2.2% (n=760).

To depict the temporal evolution of attention toward these diverse issues, please refer to Figure 5. The results reveal a strong alignment between both agendas. Notably, it is evident that peaks in one agenda are closely correlated with peaks in the other, and there are instances of overlap. This indicates high levels of face validity and results reveal strong agenda convergence across domains.

Conclusion and Discussion

The current study set out to explore various automated supervised techniques for multilingual and cross-domain classification of policy topics. In particular, the current study investigated two central approaches for automated multilingual content classification: *multilingual* and *monolingual* approaches, incorporating a range of techniques, including bag of words, machine translation, multilingual sentence embeddings (MSE), and the fine-tuning (FT) framework with pre-trained language models. Drawing on annotated data from the Comparative Agendas Project and the Comparative Manifesto Project encompassing three domains (parliamentary questions, media content, and party manifestos) and five languages (English, Dutch, German, Spanish, and Hungarian), our results indicate that multilingual bag of word approaches yielded optimal classification results within social domains. However, when transitioning into out-of-domain predictions, such as predicting topics in *parliamentary questions* using classifiers trained on*news media* data–we found that context-aware techniques, like the FT framework and MSE proved valuable in maintaining performance during out-of-domain transitions. More in particular, here we found that monolingual models employing fine-tuning techniques proved to be the best performing option.

Together, the results suggest that a monolingual approach using the fine-tuning framework is the most suitable option for researchers who aim at classifying content from different social domains. We argue that these models are a relatively safe choice, especially when uncertainty exists regarding the extent of overlap between training and application data. For instance, the application data may differ in terms of the collection timeframe, mentioned sources, and language usage. The results support the idea that monolingual fine-tuned models exhibit stability and maintain performance even in the presence of such variations.

Based on our analyses, we can draw conclusions regarding the potential for linking datasets from the *Comparative Agendas Project* and the *Comparative Manifesto Project*. While these two collections of high-quality annotated datasets cover different social domains, there exists a conceptual overlap that has been previously unexplored empirically. This overlap, however, is far from perfect and matching issues between CAP and CMP is possible for only a minority of cases. The findings suggest the possibility of connecting seemingly distinct data sources and enhancing cross-comparative research capabilities. Notably, transitions from the *news media* domain resulted in moderate performance degradation, especially when shifting to *party manifestos* or *parliamentary questions*, while transitions from the *parliamentary questions* domain had the most substantial negative impact. This indicates that the degradation in performance when transitioning between the domains covered in the *Comparative Agendas Project* and the *Comparative Manifesto Project* datasets is not worse compared to degradation when transitioning between domains within the *Comparative Agendas Project*. This supports our argument that annotation efforts in both areas are aligned and can be combined both conceptually and empirically.

Finally, our study demonstrates that when applied to analyzing policy issue attention in the United Kingdom, the classifiers developed and tested in the current study revealed a strong alignment between *news media* and *parliamentary questions*. This underscores their utility in understanding policy agenda dynamics across different contexts and languages.

Overall, this study sheds light on the potential of the monolingual and multilingual techniques to enhance cross-domain classification performance in multilingual settings. The field of computational social science has primarily focused on the English language, and there is a lack of adequate tools to validly measure concepts in other languages (Baden et al., 2022; Licht, 2023; Lind et al., 2021). The current study helps demonstrate how state-of-the-art models can enable researchers to make the most of available multilingual datasets for classification tasks within and across linguistic and content domains and improve the contextual and semantic understanding of the classifiers. By leveraging available annotated datasets for cross-comparative research, the study’s insight accelerates progress in this field. Finally, one of the prerequisites of successful combining of data from different language and political domains requires systematic linkage and procedures for interoperability as developed in the OPTED infrastructure project.

Acknowledgements

Footnotes

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of bert-based approaches. *Artificial Intelligence Review*, *54* (8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J., & Walgrave, S. (2014). Comparing and combining machine learning and dictionary-based approaches to topic coding. *th Annual Comparative Agendas Project (CAP) Conference, Konstanz, Germany*.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16* (1), 1–18.

Baumgartner, F. R., Green-Pedersen, C., & Jones, B. D. (2006). Comparative studies of policy agendas. *Journal of European public policy*, *13* (7), 959–974.

Baumgartner, F. R., Green-Pedersen, C., & Jones, B. D. (2013). *Comparative studies of policy agendas*. Routledge.

Baumgartner, F. R., & Jones, B. D. (2010). *Agendas and instability in american politics*. University of Chicago Press.

Bestvater, S. E., & Monroe, B. L. (2022). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2022.10

Bevan, S. (2017). Gone fishing: The creation of the comparative agendas project master codebook. version 0.9. 1 beta. *URL: http://sbevan. com/cap-master-codebook. html*.

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14* (2), 83–104.

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The*

ANNALS of the American Academy of Political and Social Science, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR 2020*.

Chun-ting Ho, Justing & Chan, Chung-hong. (2023). Evaluating Transferability in Multilingual Text Analyses. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.2.HO

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based Language Model. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

Delobelle, P., Winters, T., & Berendt, B. (2022). Robbert-2022: Updating a dutch language model to account for evolving language use. https://doi.org/10.48550/ARXIV.2211.08192

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding [arXiv:1810.04805 [cs]]. http://arxiv.org/abs/1810.04805

Eissler, R., Russell, A., & Jones, B. D. (2014). New avenues for the study of agenda setting. *Policy Studies Journal*, *42*, S71–S86.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359.

Karan, M., Šnajder, J., Širinić, D., & Glavaš, G. (2016). Analysis of policy agendas: Lessons learned from automatic topic classification of Croatian political texts. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 12–21. https://doi.org/10.18653/v1/W16-2102

Kroon, A. C., van der Meer, T., & Vliegenthart, R. (2022). Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, *4*(2), 528–570. https://doi.org/10.5117/CCR2022.2.006.KROO

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 1–33. https://doi.org/10.1017/pan.2023.20

Lehmann, P., Franzmann, S., Burst, T., Regel, S., RiethmÃ¼ller, F., Volkens, A., WeÃŸels, B., & Zehnter, L. (2023). The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2023a. https://doi.org/10.25522/manifesto.mpds.2023a

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, 1–14. https://doi.org/10.1017/pan.2022.29

Lin, Z., Welbers, K., Vermeer, S., & Trilling, D. (2023). Beyond discrete genres: Mapping news items onto a multidimensional framework of genre cues [https://arxiv.org/abs/2212.04185]. *International Conference on the Web and Social Media (ICWSM)*.

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. G. (2021). Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research, 3*(3). https://doi.org/10.5117/CCR2021.3.001.LIND

Nemeskey, D. M. (2021). Introducing huBERT. *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, TBA.

Osnabrügge, M., Ash, E., & Morelli, M. (2023). Cross-domain topic classification for political texts. *Political Analysis, 31*(1), 59–80. https://doi.org/10.1017/pan.2021.37

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Ronnqvist, S., Kanerva, J., & Ginter, T. S. F. (n.d.). Is Multilingual BERT Fluent in Language Generation?

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140–157.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models [arXiv:2012.15613 [cs]]. Retrieved September 4, 2023, from http://arxiv.org/abs/2012.15613
Comment: ACL 2021

Sánchez, C., Sarmiento, H., Abeliuk, A., Pérez, J., & Poblete, B. (2022). Cross-lingual and cross-domain crisis classification for low-resource scenarios.

Sebők, M., & Kacsuk, Z. (2021). The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. *Political Analysis*, *29*(2), 236–249. https://doi.org/10.1017/pan.2020.27

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*.

Viehmann, C., Beck, T., Maurer, M., Quiring, O., & Gurevych, I. (2022). Investigating opinions on public policies in digital media: Setting up a supervised machine learning tool for stance classification. *Communication Methods and Measures*, 1–35.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish [arXiv:1912.07076 [cs]]. Retrieved September 4, 2023, from http://arxiv.org/abs/1912.07076

Vliegenthart, R., Walgrave, S., & Zicha, B. (2013). How preferences, information and institutions interactively drive agenda-setting: Questions in the b elgian parliament, 1993–2000. *European Journal of Political Research*, *52*(3), 390–418.

Walgrave, S., & Van Aelst, P. (2006). The contingency of the mass media's political agenda setting power: Toward a preliminary theory. *Journal of communication*, *56*(1), 88–109.

Widmann, T., & Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2022.15