Classifying multilingual party manifestos: Domain transfer across country, time, and genre

Anonymous ACL submission

Abstract

Annotating costs of large corpora are still one of the main bottlenecks in empirical social science research. On the one hand, making use of the capabilities of domain transfer allows re-using annotated data sets and trained models. On the other hand, it is not clear how well domain transfer works and how reliable the results are for transfer across different dimensions. We explore the potential of domain transfer across geographical locations, languages, time, and genre in a large-scale database of political manifestos. First, we show the strong within-domain classification performance of fine-tuned transformer models. Second, we vary the genre of the test set across the aforementioned dimensions to test for the fine-tuned models' robustness and transferability. For switching genres, we use an external corpus of transcribed speeches from New Zealand politicians while for the other three dimensions, custom splits of the Manifesto database are used. While BERT achieves the best scores in the initial experiments across modalities, DistilBERT proves to be competitive at a lower computational expense and is thus used for further experiments across time and country. The results of the additional analysis show that (Distil)BERT can be applied to future data with similar performance. Moreover, we observe (partly) notable differences between the political manifestos of different countries of origin, even if these countries share a language or a cultural background.

1 Introduction

Publishing party manifestos in the time frame leading up to an election is a common procedure in most parliamentary democracies around the globe. Summarizing the parties' political agendas for the upcoming electoral period, the published manifestos are intended to serve as guides for voters to reach their decision (Suiter and Farrell, 2011). Since the content of these manifestos also constitutes the foundation for the process of building government coalitions, analyzing them can be very insightful. Janda et al. (1995), for instance, investigate the common assumption that political parties often try to change their images following a poor election result. Other researchers examine if parties learn from foreign successful parties (Böhmelt et al., 2016). Tavits and Letki (2009) and Tsebelis (1999) also investigate their research questions based on political manifestos.

The Manifesto Project[1] covers programs of over 1000 political parties from more than 50 countries over a time frame from 1945 until today (Lehmann, 2022). The database provides access to the raw content of all documents as well as additional annotation for further analysis. Human annotators from over 50 different countries contributed by splitting the documents into quasi-sentences and subsequently classifying each of them according to a coding scheme covering 54 thematic categories. On a more course-grained scale, these 54 categories were further summarized into eight topics. Since manual annotation is extremely time and labor-intensive, requiring annotator training reliability, (partial) automation of the process could yield enormous potential for savings.

Our research explores how methods from the field of Natural Language Processing (NLP), which are more and more frequently used in social science research (Wankmüller, 2021), can be used to classify the quasi-sentences of the political manifestos into the eight topics of the Manifesto coding scheme. Therefore, different NLP methods, namely TF-IDF + logistic regression (LR) as a comparative baseline (cf. Osnabrügge et al. (2023)) and different monolingual and multilingual versions of BERT (Devlin et al., 2019) are used to process and subsequently classify the sequences. In the following, first, the related work (cf. Sec. 2.1) and the data extraction process (cf. Sec. 2.2)

[1] https://manifesto-project.wzb.eu/

will be explained in further detail followed by the experimental setup (cf. Sec. 3), where we delve deeper into the concept of cross-domain classification and motivate the different cross-domain scenarios. The predictive performances of each evaluated model for each of the different scenarios are compared and discussed in Section 4. We conclude the experiments by fine-tuning a multilingual model on the whole corpus.

Contribution: Our main contributions can be summarized as follows: We extend the cross-domain setting introduced by Osnabrügge et al. (2023) along multiple axes. We not only measure transfer across genre (manifestos → speeches) but also across time (2018 → 2022) and country (leave-one-country-out, LOCO). Instead of relying on simple machine learning classifiers, we fine-tune pre-trained language models (Devlin et al., 2019; Sanh et al., 2019) achieving superior performance to simple models. We don't only rely on English texts, but leverage the whole Manifesto database by employing multilingual pre-trained models. This enables us to train one single model which can be used for all languages and countries. The code for our experiments and the trained models are publicly available to nurture further research: *Anonymized for review, please see supplementary material.*

2 Materials and Methods

2.1 Related work

We draw inspiration for our work from the research article "Cross-Domain Topic Classification for Political Texts" (Osnabrügge et al., 2023). The authors employ supervised machine learning (logistic regression, LR) alongside feature engineering techniques for text (TF-IDF w/ n-grams) for the classification of political manifestos and speeches. The analysis was performed on two (labeled) data sets, where each utterance was assigned one of the eight possible categories "freedom and democracy", "fabric of society", "economy", "political system", "welfare and quality of life", "social groups", "external relations" and "no topic". The source corpus consists of manifestos, collected between 1984 and 2018, which were extracted from the Manifesto Project (Krause et al., 2018) for the following seven English-speaking countries: Australia, Canada, Ireland, New Zealand, South Africa, the UK, and the USA. Each document was split into quasi-sentences ($n_{source} = 115,410$) and then labeled by a trained human annotator from the Manifesto Project. In most cases, one quasi-sentence roughly equals one sentence, however, some long sentences containing several statements were split into multiple quasi-sentences. Osnabrügge et al. (2023) use this source corpus for training and for measuring the within-domain performance. The target corpus (n_{target} = 4,165), consists of English speeches held by members of the New Zealand Parliament in the time period from 1987 to 2002. The speeches were extracted from the official record of the New Zealand Parliament (Hansard), and manually annotated according to the same schema by Osnabrügge et al. (2023), who then use it for measuring the cross-domain classification performance.

After the hyperparameter tuning using grid search, they achieve an accuracy of 0.641 on the held-out set of the source corpus and an accuracy of 0.507 on the speeches, showing that cross-domain classification is a reasonable approach. Additionally, the authors create their own, more fine-grained, coding scheme with 44 topic categories for which they report lower performance values for both the within- (0.538) and the cross-domain (0.410) setting. It is important to note, that our performance scores are not perfectly comparable to Osnabrügge et al. (2023), since we download the data ourselves (with slight differences, cf. Sec. 2.2) and thus have a different train/validation/test split.

2.2 Data extraction from Manifesto Project

For conducting the experiments described in Sec. 3, we extract the manifestos ourselves from the Manifesto Project database using its dedicated R-package *manifestoR* (Lewandowski et al., 2020). Thus, as opposed to Osnabrügge et al. (2023), our corpus also includes additional information on the year and country of origin for each utterance. Our data sets include the 2018-2 version of the corpus (Krause et al., 2018), similar to Osnabrügge et al. (2023), as well as the most recent version (2022-1, Lehmann et al., 2022), resulting in $n_{2018,en} = 114,523$ for the seven English-speaking countries mentioned in Sec. 2.1 and $n_{2018,all} = 996,008$ in total. For the 2022 corpus, there are in total $158,601$ English observations and $1,504,721$ for all languages, respectively. Among those, $n_{2022,en} = 27,764$ observations from the period between 2019 and 2022 constitute our test set for the experiments across time for the English language. We observe a difference of 887 samples between the data from Osnabrügge et al. (2023) ($n_{source} = 115,410$) and our data set ($n_{2018,en} = 114,523$), which is probably due to potential changes in the 2018 version the database.

Figure 2 (Appendix A) visualizes the different label distributions for (a) the source corpus of Osnabrügge et al. (2023), (b) our extraction of the 2018-2 corpus, (c) our extraction of the 2022-1 corpus, and (d) the target corpus of the New Zealand speeches (Osnabrügge et al., 2023). While the former three roughly follow the same distribution, with about 57% of the observations assigned to either "*welfare and quality of life*" or "*economy*", the most common class of the latter is "*political system*" (∼26%) followed by "*welfare and quality of life*" (∼19%). Thus, the two main challenges aside from the domain transfer are the overall class imbalance as well as the differences between the source and target domain with respect to the label distribution. Further Figure 3 (Appendix A) shows the distribution of the target classes separated by the language the manifestos are written in. We display the three most frequent languages, which we use for conducting experiments across country (cf. Sec. 3.1), against the distribution in the entire 2018-2 corpus of all manifestos. Here we observe some minor differences, as "*welfare and quality of life*" and "*political system*" are more frequently addressed in German-speaking countries (compared to the overall corpus), "*welfare and quality of life*" and "*economy*" in French-speaking ones, and "*political system*" and "*economy*" in English-speaking ones. Notably, for all three languages, the topics "*freedom and democracy*" and "*external relations*" are addressed less often than in the whole 2018-2 corpus.

3 Experimental Setup

In this section, we introduce the concept of domain transfer in general and in particular the cross-domain classification settings for our application. Further, the methodological background for the employed model architectures will be laid out as follows: First, we briefly review common feature engineering techniques for text data and elaborate on the advantages and disadvantages. These techniques include term-frequency inverse-document-frequency (TF-IDF) weighting, as well as dense word or document embeddings. Second, we introduce two state-of-the-art NLP architectures that we employ in our analysis, namely BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019), both of which do not require prior feature engineering steps but accommodate the whole pipeline in one single model. Finally, we briefly sketch the individual experiments which were carried out over the course of this study.

3.1 Cross-Domain Classification

When talking about *classification* in the context of machine learning, researchers commonly implicitly refer to within-domain/within-distribution classification, implying that the trained model is tested on data from the same origin/distribution as the training data (i.e. the *source domain*). Cross-domain classification, on the other hand, explicitly considers a shift in the domain/distribution/source of the data, i.e. the data-generating process is assumed to be different. Frequently examined cases of domain shift in NLP include a change in language (i.e. training the model on text from one language and evaluating it in another, cf. Conneau et al. (2018, 2019)), topic (e.g. training the model on reviews on restaurants and evaluation it on reviews on laptops, cf. Pontiki et al. (2014)) or genre (e.g. training on texts and evaluation on transcribed audio data, cf. Osnabrügge et al. (2023)). In our experiments, we contribute to this body of research by considering the following different cross-domain settings:

Transfer across genre: We consider party manifestos from all seven (English-speaking) countries as our source corpus $C_{source} = C_{2018,en}$ and evaluate the trained model on a target corpus C_{target} of transcribed parliamentary speeches from New Zealand. This setting is equivalent to the work of Osnabrügge et al. (2023), yet we rely on more elaborated model architectures.

Transfer across time: We use the party manifestos from all countries for all years up until 2018 as source corpus C_{source}[2], while the target corpus C_{target} consists of party manifestos from the year 2019 – 2022. This setting is intended to test the temporal robustness of the fine-tuned models.

Transfer across country: This setup comprises three distinct experiments for different languages (English, German, French), for each of which we include data from all[3] countries, where manifestos in the given language exist in the 2018-2 corpus. The setting for each language consists again of seven

[2] C_{source} is either $C_{2018,en}$, $C_{2018,de}$ or $C_{2018,fr}$

[3] For English we excluded countries with a low n, to stay consistent with Osnabrügge et al. (2023).

	Data set characteristica		Data set splitting		Data set sizes	
Scenario	Corpus	Language(s)	Training set	Test set	Training set	Test set
within-domain	2018-2	En, De, Fr	random split[a]	random split[b]	91,618 / 104,710 / 17,885	11,452 / 13,089 / 2,236
manifestos $\rightarrow$ speeches	2018-2	En	random split[a]	speeches	91,618	4,165
2018 $\rightarrow$ 2022	2018-2	En, De, Fr	random split[a]	future[c]	91,618 / 104,710 / 17,885	27,764 / 30,542 / 343
across country	2018-2	En, De, Fr	$n-1$ countries	held-out country	–[d]	–[d]
Multilingual	2018-2	38 languages	random split[a]	random split[b]	796,806	99,601

[a] Here: .8/.1/.1, i.e. 80% of the 2018-2 data.
[b] Here: .8/.1/.1, i.e. 10% of the 2018-2 data.
[c] *"future"*: data from the 2022-1 corpus recorded after the 2018-2 cut-off.
[d] Different scenarios, test set contains one single country in each experiment.

Table 1: Overview of the investigated cross-domain scenarios, alongside the used corpora, test sets, and languages.

(five and four, respectively) different individual experiments, since for each language we include all but one country as source corpus C_{source} and evaluate the model on a target corpus C_{target} including only the manifestos from the single held-out country. Further, we also inspect a true multimodel model trained on data from all available countries.

Metrics and Training We compare our results, which we measure in terms of Accuracy and Macro-F1 Score, from the cross-domain experiments to the performance we obtain for the within-domain setting. We opt for reporting the macro-averaged version of the F1 Score in order to take into account the class imbalance (cf. Fig. 2). For model training, we conduct a train/validation/test split with proportions .8/.1/.1; all reported performance values are measured on the test set. Note that, depending on the cross-domain setting, also different test sets than the random split are used. Table 1 summarizes the different investigated scenarios in a comprehensive manner, provides an overview of the respectively used corpora for training and evaluation, and specifies with which procedure the respective test sets were created or selected.

3.2 Model architectures

Early feature engineering techniques relying on the bag-of-words (BoW) assumption have in recent years been replaced by more elaborated representation learning algorithms. BoW refers to counting the occurrences of words (or n-grams) in a document and representing it as V-dimensional vector, where V is the vocabulary size. This representation can be enhanced via TF-IDF, as done by Osnabrügge et al. (2023), via a re-weighting using corpus-level occurrence statistics.

With the advent of representation learning, it became possible to represent words (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) and documents (Le and Mikolov, 2014) by dense vectors of a comparably low, fixed dimensionality. These representations were used in a similar fashion in conjunction with a classifier as BoW-based representations. BERT (Devlin et al., 2019) enabled the coupling of these two steps, i.e. it provided one single end-to-end trainable model for learning (contextual) representations and training the classifier. The commonality of BERT and all subsequent architectures is that they all are relying on the Transformer architecture (Vaswani et al., 2017). Based on BERT, DistilBERT models can be trained using model distillation (Buciluǎ et al., 2006; Hinton et al., 2015), a training process during which the smaller student model (DistilBERT) is trained to mimic the larger teacher model's (BERT) behavior. In the case of DistilBERT, the student model, while having half the size of its teacher model, is able to retain approximately 95% of the teacher model's performance on the GLUE benchmark (Sanh et al., 2019).

We use `bert-base-cased` as well as `distilbert-base-cased` for English. For further experiments, we employ `distilbert-base-german-cased`, `flaubert_small_cased` (as no French DistilBERT is available) and `distilbert-base-multilingual-cased`.

3.3 Experiments

In the first step, we stick to the setup from Osnabrügge et al. (2023), extracting similar data, re-running their experiments, and comparing against their LR+TF-IDF baseline. We further compare the performance of BERT against the cheaper DistilBERT for the English within-domain setting and the English cross-domain settings (manifestos $\rightarrow$ speeches, 2018 $\rightarrow$ 2022, and across country) to assess the competitiveness of the latter one. For the cross-domain scenarios in the other languages (German, French) we thereafter conduct all experiments with DistilBERT, since it is the cheaper model. The

	within-domain		manifestos → speeches		2018 → 2022	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
TF-IDF + LR	0.6413	0.5195	0.5059 (↓ 0.1354)	0.4474 (↓ 0.0586)	–	–
English BERT	0.6977	0.5841	0.5613 (↓ 0.1364)	0.5046 (↓ 0.0795)	0.6841 (↓ 0.0136)	0.5707 (↓ 0.0134)
English DistilBERT	0.6866	0.5694	0.5669 (↓ 0.1197)	0.5026 (↓ 0.0568)	0.6784 (↓ 0.0082)	0.5620 (↓ 0.0074)
German DistilBERT	0.6583	0.5628	–	–	0.6559 (↓ 0.0024)	0.5485 (↓ 0.0143)
FlauBERT	0.6087	0.5159	–	–	0.6093 (↑ 0.0006)	0.4783 (↓ 0.0376)
Multilingual DistilBERT	0.6748	0.5941	–	–	0.6311 (↓ 0.0437)	0.5278 (↓ 0.0663)

Table 2: Performance values of TF-IDF + LR (Osnabrügge et al., 2023) versus English BERT and DistilBERT models (upper part) as well as for German DistilBERT and French FlauBERT models (middle part) and the multilingual DistilBERT model (lover part). Absolute change vs. within-domain performance in parentheses.

concluding multilingual experiments on the complete corpus are also conducted using a DistilBERT model, fine-tuning the model on the train set of a random split of *the whole* 2018-2 data set.

4 Results

This section will be structured as follows: First, we will show the superior within-domain performance of pre-trained BERT-based models over the simple baseline from Osnabrügge et al. (2023) and will closely inspect the per-class within-domain performances of the different models. In conjunction with this, we also compare our models to Osnabrügge et al. (2023) on the manifestos → speeches scenario, since we adopt it from their work. This scenario we can, however, only inspect for the English language as the corpus of speeches is from New Zealand. Second, we will verify if and how well experiments across genre and time work for the different monolingual models and the multilingual one. Third, we inspect closely how well performance can be transferred across different countries speaking the same language. Subsequently, we delve deeper into a truly multilingual by fine-tuning a pre-trained multilingual model on the entirety of the corpus and examining its performance for the different countries and languages.

Within-domain performance The results of our experiments comparing different models for within-domain classification, manifestos → speeches, and 2018 → 2022 classification are presented in Table 2. For within-domain classification, the TF-IDF + LR model is clearly outperformed by the deep learning models, where the English models perform better than the German, French, and Multilingual ones. It is notable that in general, the French model exhibits rather low performance values[4] (within-domain as well as across time) compared to all other models, which may for one reason be caused by the relatively small corpus size for this language compared to all other ones (cf. Tab. 1). We also observe the expectedly higher performance of the English BERT model compared to the English DistilBERT, since it generally outperforms DistilBERT in all scenarios except for the accuracy in *manifesto → speeches* transfer. However, the performance gaps between these two models are rather small, which very well justifies the use of DistilBERT for the remainder of the experiments, trading some performance for saving computational expenses.[5]

When further considering the predictive performance separately for each of the eight classes (cf. Tab. 3), we learn that for none of the languages and for none of the investigated scenarios any of the monolingual DistilBERT models was able to predict a single case of the highly underrepresented "*no topic*" class. The obvious reasons for this are the low number of observations as well as the potential ambiguity, heterogeneity, and fuzziness of the manifestos that could not even by the human annotators be classified into one coherent class but were assigned to this collection basin. This peculiarity of the results should always be taken into account when interpreting them since the macro-averaged F1 Score tends to be a rather conservative performance measure as it weighs the performance of this class similarly to all other classes. This also largely explains the quite notable gap between the Accuracies and Macro-F1 Scores (cf. Tab. 2).

[4]Note, that cannot be compared to the English TF-IDF + LR baseline due to different training and test sets.

[5]While training BERT for one epoch took roughly 1h 11min, DistilBERT nearly halved this training time per epoch to about 38min. Adding this up over three epochs amounts to time savings of nearly 100min.

	English			German			French			Multilingual		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
No Topic	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4142	0.1394	0.2086
Freedom / Democracy	0.6258	0.5318	0.5750	0.6631	0.6133	0.6372	0.6533	0.5868	0.6183	0.6165	0.5787	0.5970
External Relations	**0.7395**	0.7517	0.7456	**0.7429**	0.7067	**0.7243**	**0.6688**	0.6913	**0.6799**	**0.7357**	0.7068	0.7209
Social Groups	0.5794	0.5488	0.5637	0.6040	0.5370	0.5685	0.6034	0.4506	0.5160	0.6242	0.5372	0.5774
Political System	0.5629	0.4773	0.5166	0.6088	0.5145	0.5577	0.4407	0.5372	0.4842	0.6012	0.5646	0.5823
Fabric of Society	0.6463	0.6727	0.6592	0.5909	0.6496	0.6189	0.5485	0.4837	0.5140	0.6212	0.6092	0.6151
Economy	0.7269	0.7570	0.7416	0.6882	0.7009	0.6945	0.6270	0.6449	0.6358	0.6934	0.7449	0.7182
Welfare / Quality of Life	0.7293	**0.7793**	**0.7534**	0.6686	**0.7379**	0.7015	0.6604	**0.6990**	0.6791	0.7151	**0.7517**	**0.7330**

Table 3: A detailed performance report for per-class within-domain performance, measured in terms of Precision (P), Recall (R), and F1 Score, for the DistilBERT models in English and German, the French FlauBERT as well as for the multilingual DistilBERT. Best scores (per language) in **bold**, runner-up underlined.

The largest class (in terms of the number of observations) was easiest to classify for the DistilBERT models across all languages, i.e. for "*welfare and quality of life*" overall the highest values in P, R, and $F1$ are observed. Interestingly it is not the second largest class ("*economy*") where the models perform next best, but rather one of the smallest classes ("*external relations*"), which is nicely visualized by the highlighting in Table 3. Nevertheless, the models are capable of predicting also the "*economy*" class quite well. Further, it is interesting to observe that for the classes exhibiting high F1 Scores, the gap between recall and precision is (a) rather small and (b) sometimes even in favor of the recall, while for the low-performance classes, the recall often appears to be notably worse than the precision. This is especially consistently observable for the class "*social groups*".

When compared to the monolingual models, the multilingual one stands out due to two distinct reasons (cf. Tab. 3): First, it is the only one of the four models to detect at least *any* true "*no topic*" observations in its test set. Although the performance for this particular class still is not great, it still seems as if learning from more (and more diverse) data seems to help in this respect. Second, and probably also related to the first advantage, the performance seems to be more stable when comparing the scores across the different classes. While for the other English and French, the ranges (excluding "*no topic*") of the F1 Score were 0.2290, and 0.1957 respectively, this metric is with a value of only 0.1556 comparably small, similar to 0.1666 for the German language.

Transfer across genre and time Inspecting the two cross-domain settings in Table 2 more closely, we see that transfer across the temporal axis works better than across the genre axis. While for the English DistilBERT model the performance on the New Zealand speeches drops by quite a margin (↓ 0.1197 / ↓ 0.0568), it merely changes when evaluated on the data from a different time period (↓ 0.0082 / ↓ 0.0074). Again, comparing BERT to DistilBERT, the latter even seems to be more stable over time since the performance decrease is slightly less pronounced. For the cross-modal transfer scenario, we provide the confusion matrix (cf. Fig. 4 in Appendix B) to enable further error analysis. While the two most frequent classes are still very accurately predicted, the model severely struggles when it comes to distinguishing many of the other classes from the "*political system*" category. Even for the two largest classes, a notable amount of the instances were misclassified into this category. Further, the model's error of confusing a certain category with "*political system*" is even worse for the smaller classes, e.g. "*freedom and democracy*", with fewer samples.

While this comparison of the scenarios across genre and across time can not be made for the other languages and the multilingual scenario, we also observe only very minor drops in performance for the latter scenario there. For the two monolingual models, we record decreases for accuracy of 0.24 percentage points for the German model and even no decrease at all for the accuracy of the French DistilBERT model, as well as decreases of 1.43 (German) and 3.76 (French) percentage points for Macro-F1. The multilingual model, however, exhibits somewhat larger drops in performance of 4.37 percentage points for accuracy and 6.63 percentage points for Macro-F1, respectively.

Transfer across countries The results of our LOCO experiments using the monolingual DistilBERT models for English and German, and a FlauBERT model for French, are presented in Ta-

	n_{random}	$n_{country}$	English-LOCO (DistilBERT) Accuracy	English-LOCO (DistilBERT) Macro-F1	German-LOCO (DistilBERT) Accuracy	German-LOCO (DistilBERT) Macro-F1	French-LOCO(FlauBERT) Accuracy	French-LOCO(FlauBERT) Macro-F1
Australia	1,861	18,480	**0.6304**	0.4877	–	–	–	–
Canada	322	3,047	0.5829	**0.5441**	–	–	–	–
Ireland	2,548	25,357	0.5962	0.4895	–	–	–	–
New Zealand	2,840	28,561	<u>0.6268</u>	0.4761	–	–	–	–
South Africa	628	6,423	0.5997	<u>0.4954</u>	–	–	–	–
United Kingdom	2,182	21,836	0.6080	0.4924	–	–	–	–
United States	1,071	10,819	0.5744	0.4755	–	–	–	–
Austria	3,361	33,818	–	–	<u>0.6071</u>	<u>0.5077</u>	–	–
Germany	6,452	63,413	–	–	0.6039	0.5060	–	–
Italy	63	651	–	–	0.5699	0.4733	–	–
Luxembourg	1,850	19,291	–	–	**0.6114**	**0.5134**	–	–
Switzerland	1,390	13,715	–	–	0.5754	0.4878	–	–
Canada	517	5,386	–	–	–	–	0.4629	0.3822
France	850	8,290	–	–	–	–	<u>0.5624</u>	<u>0.4511</u>
Luxembourg	868	8,662	–	–	–	–	0.5179	0.3993
Switzerland	1	19	–	–	–	–	**0.7368**	**0.7288**
Average			**0.6026**	**0.4944**	**0.5935**	**0.4976**	**0.5700**	**0.4904**

Table 4: LOCO performance for English (7 countries), German (5 countries), and French (4 countries). Best scores per language in **bold**, runner-up <u>underlined</u>. We report both n_{random} for the number of observations in the random test split and $n_{country}$ for the number of observations when the respective country is used as held-out set.

ble 4. We support the results by visualizations (cf. Fig. 1) of how the performance on manifestos from a certain country changes depending on whether we (a) evaluate on its portion of the random test split or (b) on all manifestos of this country as a hold-out set. The most important takeaway from these illustrations is the fact that completely withholding data from a certain country hurts model performance on data from this specific country, but not in equal parts for the different languages. For German-speaking countries (cf. Fig. 1, middle) the decrease from left to right is less pronounced than for the other two languages (Fig. 1, top/bottom).

The overall takeaway from the previous experiments (better performance for English) is not entirely confirmed by these results, also showing a much more nuanced picture regarding interesting inter-country differences per language. For the LOCO scenario within the English-speaking countries, Australia and New Zealand exhibit the highest values for accuracy, while South Africa and Canada outperform the other with respect to Macro-F1[6]. The two European countries and the United States overall show the worst performance with respect to both metrics. Further, it is worth noting that there is a rather high variation among these performance values compared to German and French. Excluding the "*no topic*" class, the values for accuracy exhibit a range of 0.0560, while the Macro-F1 Score has a range of 0.0686. On a final note, it is interesting to see that the performance on New Zealand *manifestos* is among the top-ranking countries in accuracy, while the domain transfer across modalities (to New Zealand *parliamentary speeches*) shows a little bit of a performance decrease.

The German LOCO classification experiments using DistilBERT exhibit somewhat different results compared to the English experiments. While the overall averages are comparable, the ranges (0.0415 for accuracy and 0.0344 for Macro-F1) indicate that the values for all countries are relatively similar, with Luxembourg having the highest accuracy of 0.6114 as well as the highest Macro-F1 Score of 0.5134. We speculate that the reason for this observation might lie (a) in the similarity of the political systems[7] of all these countries and (b) in their geographical and cultural closeness. However, being no experts in political science, we would leave the definite interpretation of such matters to those. Regarding the overall performance, the German model performs no worse than the English model(s) which was not necessarily to be expected due to our conclusions drawn from Tables 2 and 3.

A rather distinct picture emerges when inspect-

[6]Canada has better Macro-F1 Scores than most other countries (except for the top two), but comparably low accuracy.

[7]Despite Luxembourg being a parliamentary monarchy, the country still has a similar landscape of political parties compared to its neighbors, including i.a. social and Christian democrats, liberals, a Green party, as well as different smaller left- and right-wing parties.

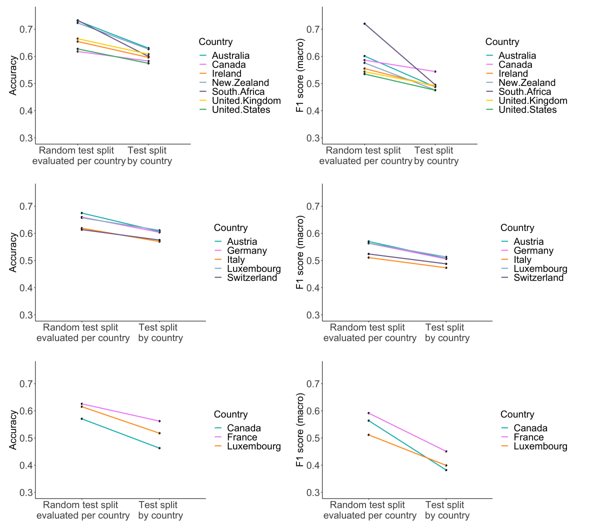


Figure 1: Comparison of the performance on data from specific English- (top), German- (middle), and French-speaking (bottom) countries via the Accuracy (left) and Macro-F1. On the left-hand side of each subfigure, performance is measured on the portion of each country in the random test set, while on the right side, the country-specific LOCO performance is displayed. Lines are drawn between the respective points to visualize the connection within one country. Switzerland is excluded, since there is only one sample in the random test split.

ing the results for the French LOCO classification (still bearing in mind that the performance estimates for Switzerland, with only 19 observations, might make the interpretations rather unreliable). The range for accuracy is 0.2739 and 0.3466 for Macro-F1, which is notably larger than the ranges for both the English-speaking countries and the German-speaking countries. Switzerland exhibits by far the highest values, but it should again be noted that they are based on only 19 observations. The average values are comparable, although a bit lower, to the other two languages, but again strongly influenced by the seemingly strong performance on Swiss manifestos. Regarding the other three countries, France itself stands out from the other two, exhibiting both the highest accuracy as well as the highest Macro-F1 Score among them.

5 Conclusion and Future Work

We showed in a series of extensive experiments that domain transfer along three different axes (genre, time, country) in principal works for this sort of political text. We observed the largest performance drops when attempting to generalize across modalities, however, the models tend to generalize very well across time. While the first finding might be foreseeable, the latter result is insofar kind of interesting since after the time point we chose for splitting the data (2018) quite some new topics, e.g. the global covid-19 pandemic or the Ukrainian war, emerged. Regarding the generalization across country, even within languages (and hence to some extent also cultural backgrounds), there seem to be notable differences between the political communication in the different countries as observed by the large performance differences. To conclude, we can state that a true multilingual approach towards classifying political text looks promising, yielding good and stable performance across numerous countries with different languages.

Interesting starting points for future work are obviously to examine the capacities of the emerging ever more powerful LLMs to tackle challenging tasks like this and to make use of the continuously extending data pool from the Manifesto project. Since new countries and time points are added constantly, there is definitely the potential to extend our work in future research.

Limitations

The advent of large language models (LLMs), in particular ChatGPT (OpenAI, 2022; Bubeck et al., 2023), resulted in a paradigm change in NLP research. Since then, we can loosely categorize existing and newly introduced classification models into several bins: "pre-train/fine-tune", "prompting", and "chatting"While "pre-train/fine-tune" has been (and still widely is) the pre-dominant research paradigm in applied NLP research since $\sim$ 2018, "prompting" has upon the introduction of GPT-3 (Brown et al., 2020) become an exciting approach for tackling (a) multi-task learning and (b) low-resource scenarios via few-/zero-shot learning. Further, accessing a model via prompting might be considered more "human-like" / "natural" than training a model on class labels via gradient descent.

On the other hand, there are still also numerous reasons not to abandon architectures relying on the "pre-train/fine-tune" paradigm (Yang et al., 2023), several of which we consider fulfilled as far as our research question is concerned. First, given the large, annotated training corpus there is no need to rely on few-shot learning but rather to use all of the available data points to achieve maximum model performance. Prompting models would struggle with this amount of data due to context length constraints. Second, given the very custom-defined label set of political topics for this political corpus, for general-purpose prompting models, this label set would always have to be in some way appended to the prompt for the model to be informed about the granularity in the first place. On the one hand, this would probably lead to the model struggling with learning the underlying concepts, on the other hand, it would lead to better adaptive capabilities in case the granularity changes. Third, for domain-specific research questions like this, it might not always be feasible for researchers to access the computational resources for running or prompting such large models, and hence a task-specific, parameter-efficient model that does the trick equally well might be preferable.

We further acknowledge that the performance could potentially still be increased using more elaborate models following the "pre-train/fine-tune" paradigm, e.g. variants of the T5 model family (Raffel et al., 2020; Xue et al., 2020). Using these models, however, comes at the cost of a higher computational expense potentially requiring much more VRAM than the average practitioner has access to. The models we employ can, on the other hand, be fine-tuned comfortably using smaller GPUs with around 16GB of VRAM in an acceptable amount of time. Given the ever-increasing model sizes and thus also the computational requirements, this is an important issue to keep an eye on.

Ethical considerations

To the best of our knowledge, no ethical considerations are implied by our work. The only aspect that is affected in a broader sense is the environmental impact of the computationally expensive experiments. This issue naturally comes with pre-training large language models and is obviously a concern that has to be expressed in every work dealing with this sort of model. But on the other hand, our work rather works against increasing the environmental impact, since we "only" focus on reusing existing pre-trained models and performing the cheap(er) fine-tuning step. Further, we also provide access to our fine-tuned models which can be used by other researchers.

Acknowledgements

Excluded for anonymization reasons.

References

Tobias Böhmelt, Lawrence Ezrow, Roni Lehrer, and Hugh Ward. 2016. Party policy diffusion. *American Political Science Review*, 110(2):397–410.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. arxiv 2016. *arXiv preprint arXiv:1607.04606*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Kenneth Janda, Robert Harmel, Christine Edens, and Patricia Goff. 1995. Changes in party identity: Evidence from party manifestos. *Party Politics*, 1(2):171–196.

Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, and Sven Regel. 2018. Manifesto Corpus, Version: 2018-2. *Berlin: WZB Berlin Social Science Center*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Pola Lehmann. 2022. Manifesto project. Accessed: 2022-10-01.

Pola Lehmann, Tobias Burst, Jirka Lewandowski, Theres Matthieß, Sven Regel, and Lisa Zehnter. 2022. Manifesto Corpus. Version: 2022-1. *Berlin: WZB Berlin Social Science Center*.

Jirka Lewandowski, Nicolas Merz, and Sven Regel. 2020. *manifestoR: Access and Process Data and Documents of the Manifesto Project*. R package version 1.5.0.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Accessed: 2023-01-10.

Moritz Osnabrügge, Elliott Ash, and Massimo Morelli. 2023. Cross-domain topic classification for political texts. *Political Analysis*, 31(1):59–80.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jane Suiter and David M. Farrell. 2011. *The Parties' Manifestos*, pages 29–46. Palgrave Macmillan UK, London.

Margit Tavits and Natalia Letki. 2009. When left is right: Party ideology and policy in post-communist europe. *American Political Science Review*, 103(4):555–569.

George Tsebelis. 1999. Veto players and law production in parliamentary democracies: An empirical analysis. *American political science review*, 93(3):591–608.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

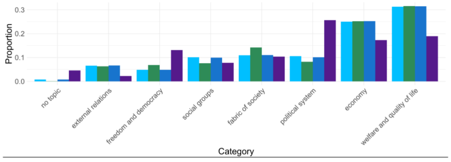
Sandra Wankmüller. 2021. Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research*, page 00491241221134527.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

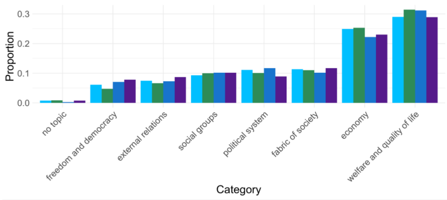
Appendix

A Label distributions



	Data	Number of observations	Description
(light blue)	Source data (ours)	$n_{2018,en}$ = 114,523	Extracted manifestos from the Manifesto Project (version 2018-2)
(blue)	Source data (Osnabrügge et al.)	n_{source} = 115,410	Data from Osnabrügge et al. [18] comprising manifestos (version 2018-2)
(green)	Target data (2022)	$n_{2022,en}$ = 27,764	Extracted manifestos from the Manifesto Project (version 2022-1) after 2018-2
(purple)	Target data (speeches)	n_{target} = 4,165	Data from Osnabrügge et al. [18] comprising speeches from New Zealand

Figure 2: Label distributions for the four different corpora alongside sample sizes and short descriptions.



	Data	Number of observations	Description
(light blue)	English	$n_{2018,en}$ = 114,523	Manifestos from English-speaking countries
(blue)	German	$n_{2018,de}$ = 130,888	Manifestos from German-speaking countries
(green)	French	$n_{2018,fr}$ = 22,357	Manifestos from French-speaking countries
(purple)	All	$n_{2018,all}$ = 996,008	Manifestos from all countries

Figure 3: Label distributions for the three most frequent languages and overall in the 2018-2 corpus alongside sample sizes and short descriptions.

B Confusion matrix

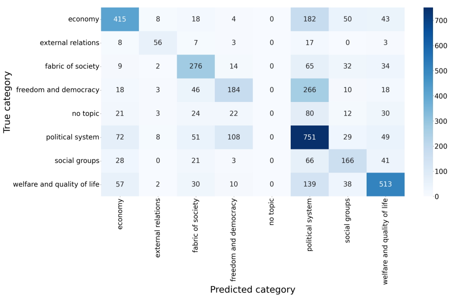


Figure 4: Confusion matrix for the performance of the English DistilBERT model on the test set of the New Zealand parliamentary speeches.