Policy Parsing Panthers at Touché: Ideology and Power Identification in Parliamentary Debates*

Notebook for the Touché Lab at CLEF 2024

Oscar Palmqvist[1], Johan Jiremalm[1] and Pablo Picazo-Sanchez[1,2,*]

[1]Chalmers University of Technology, Gothenburg, Sweden
[2]School of Information Technology, Halmstad University, Sweden

Abstract

Political debates are vital in shaping public opinion and influencing policy decisions. However, understanding the complex linguistic structures used by politicians to ascertain their orientations and power dynamics can be challenging. In this paper we explore Natural Language Processing techniques for identifying political orientation and power structures in parliamentary debates. We introduce a Located Missing Labels-loss in order to train jointly to predict both power and ideology. Furthermore, our proposed method also trains to predict a third synthetically generated polarity label. Finally, we combine this training method with pre-processing steps including back-translation and meta data inclusion. Our results show that our method manages to improve upon conventional methods of fine-tuning.

Keywords

Political Debates, NLP, CLEF, Touché

1. Introduction

Parliamentary debates play a vital role in political communication and society [1]. During these debates, representatives from diverse parties and ideologies share their opinions, arguments, and stances on issues impacting society. Making debates more accessible and easy to follow not only serves to inform people but offers a basis for seeking further information and engaging in the democratic process [1].

The ability to detect and classify political motives from speech may also be utilised when the speaker is not forthcoming with their political agenda. Detecting hidden political motives in media such as news reporting and advertisements may benefit society by providing transparency [2].

The complex nature of politics makes these debates challenging to understand [3]. According to a recent survey, 65% of Americans say they always or often feel exhausted when thinking about politics [4]. In addition, 42% of adults in the U.S. reported to have watched none, or very few, of the presidential debates in 2020. Moreover, analysing political speeches and making classifications can be challenging due to complex rhetorical strategies such as metaphors, parallelism, and suggesting answers [5]. Political context and the speaker's background influence how messages are conveyed and interpreted.

The challenge of analysing and making classification on political speech may be approached from the perspective of Natural Language Processing (NLP). NLP is a field in artificial intelligence that focuses on analysing, understanding, and processing natural language data using computers [6]. Sub-tasks within NLP involve, among others, text summarisation, machine translation, and sentiment analysis. More recently, the field of NLP has surged in popularity with the development of chatbots such as ChatGPT. Besides the massive Large Language Models (LLMs) such as GPT-4, there have also been multiple other different approaches for NLP such as rule-based and probabilistic approaches [7]. The incredible performance of these LLM can be applied to the complex political realm with great success [8].

Conference and Labs of the Evaluation Forum (CLEF) [9] hosts an open competition in 2024 called Touché as part of one of their so called *labs* [10]. The task, *Ideology and Power Identification in Parliamentary Debates*, is one of the four competitions as part of Touché's presence at CLEF 2024 [11]. This

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
0000-0002-0303-3858 (P. Picazo-Sanchez)

document is an entry to that competition. It will, therefore, investigate which NLP tools are best suited for identifying ideology and power in parliamentary debate speeches.

Research goals. The research goals[1] for this project are inspired by and correspond to the two sub-tasks of the Touché competition *Ideology and Power Identification in Parliamentary Debates* [11, 12].

RQ1 Investigate what the best methods and practices are for identifying the political orientation in a parliamentary speech.

RQ2 Investigate what the best methods and practices are for identifying whether a parliamentary speech is made by a speaker in opposition or in power.

Evaluation The results of both research questions are evaluated against a test set provided by Touché using a macro-averaged F1-score [13], as it is the performance metric of the Touché competition [11].

Paper structure The rest of this document is organised as follows. In Section 2 we discuss a selection of related work which is closely related to our problem at hand and which influenced our method. We outline and describe the datasets set in Section 3. We present our full proposed method in Section 4 and share the results from its application in Section 5. Finally, we discuss and explain these findings in Section 6 and conclude the document in Section 7.

2. Related Work

Here, we discuss previously explored techniques used for political classification within NLP and their relevance to our project. We also examine a range of studies showcasing the performance of models such as RoBERTa and Bidirectional Encoder Representations from Transformers (BERT) in tasks such as multi-lingual political orientation classification and stance classification of political tweets. Additionally, we explore domain-specific pre-training, back-translation, multi-label learning with missing labels, and other methodologies relevant to our research objectives.

2.1. Model and training method

Fine-tuned RoBERTa has demonstrated significant superiority over zero- and few-shot GPT-3 for multi-lingual political orientation classification [14]. For the task of implicit ideology prediction, fine-tuned RoBERTa has also been shown to outperform GPT-4, Llama-2-13B and Llama-2-70B using in-context learning, as well as Llama-2-70B utilising Low Rank Adaptation (LoRA)-fine tuning [15]. Furthermore, it is noteworthy that in the same study, the only model which beat fine-tuned RoBERTa for Named-Entity Recognition (NER) was the LoRA fine-tuned Llama-2-70B.

It has been shown that even when only fine-tuning a classification head, BERT has approached the performance of few-shot GPT-3 models for stance classification of political tweets [16].

In a similar manner, a comparison of the performance of "Small Language Models" and modern LLMs for sentiment analysis tasks has been performed [17]. The study compares $T5_{large}$ (770M parameters) to Flan-T5, Flan-UL2, text-davinci-003, GPT-3.5-turbo (11B, 20B, 175B, undisclosed, parameters respectively) across 13 different sentiment analysis tasks and 26 datasets. For context, the base version of BERT has 110M parameters whilst the large version totals 340M parameters. The T5-model was trained on the entire training dataset whilst the other LLMs utilised zero- or few-shot classification. They divide their sentiment analysis tasks into three categories, and find that the smaller fully trained T5-model achieves the best results for all categories, outperforming zero-, one-, five-, and ten-shot classification versions of the larger models. The authors conclude that whilst the larger LLMs perform adequately in simpler tasks, they are outperformed in complex tasks which require structured sentiment information or deeper understanding.

[1] Also referred to as sub-tasks.

RoBERTa has been shown to be state-of-the-art for propaganda classification [18]. The same RoBERTa-model outperforms fine-tuned versions of GPT-3 and multiple in-context learning versions of GPT-4 [19], yielding a micro average F1-score of 63.4% as compared to that of the best GPT-model (base GPT-4) of 58.11%.

New and large decoder-only models have been compared to encoder-only models for the tasks of intent classification and sentiment analysis [20]. The study reveals that, in general, encoder-only models provide superior performance, at a fraction of the computational demand, for natural language understanding tasks.

2.2. Domain-specific pre-training

Domain-specific pre-training refers to the process of training a model on domain-specific texts before fine-tuning for a specific task within that domain [21]. Domain-specific pre-training has shown great utility for domains with abundant unlabeled text, such as the biomedical field [22].

For the task of multi-lingual political orientation classification, however, some authors recently showed that domain-specific pre-training does not greatly impact results [14]. Also, after a threshold of approximately 10,000 sentences, the general-domain pre-training seems to be sufficient as to not benefit from additional domain-specific pre-training.

2.3. Back-translation

Back-translation involves the process of translating one text into another language, and then translating the new text back into the original language [23]. In our case this technique will be used to create artificial data that is similar to the original data, as further explained in Section 4.1.1. Back-translation has shown widespread utility for machine translation tasks [24]. Furthermore, using back-translation to artificially extend datasets has also shown promising results for hate speech detection tasks [25]. Back-translation for classification tasks has also shown itself to be particularly useful when there is less training data [26].

2.4. Multi-label learning with missing labels

Multi-label Learning with Missing Labels (MLML) has shown great utility for image classification tasks [27]. In these tasks, a "missing" label most often refers to a false negative, and the challenge is to differentiate between true negatives and false negatives caused by incomplete or faulty annotation. Many methods have been proposed to handle these missing labels [28, 29, 30]. As will be shown in Section 4, our two sub-tasks can be combined into a single multi-label classification problem with located missing labels. Traditional MLML methods are, however, not suited for our task since the locations of our missing labels are known and not hidden as false negatives. A study on MLML for image and facial-expression classification from 2014 shares our definition of MLML where missing labels are located but the technique is not appropriate for our project since it is tailored for a label-space magnitudes larger than our own [31].

Our two sub-tasks can be combined into a single multi-label classification problem, as will be shown in Section 4. Similar approaches of combining tasks have shown to be beneficial. For the task of peer-assessment evaluation, multi-task learning BERT has been shown to outperform its single-task counterparts [32]. For the multi-task learning BERT, three separate classification heads were added to the same base BERT model and the loss for fine-tuning was the sum of the Cross Entropy (CE)-loss from each classification task.

2.5. Performance of models in similar contemporary competitions

In 2021, amongst other years, CLEF organised a competition called EXIST [33]. The two task for the competition were:

	Task1 Ranking	Task 2 Ranking	Average Ranking	Bert	Beto	mBert	XLM-R	RoBERTa	RF	LR	SVM	fastText
Ai-UPV	1	1	1	x	x	x						
SINAI-TL	2	3	2,5	x	x							
AIT FHSTP	3	5	4				x					
Multiaztertest	4		4	x	x							
LHZ	8	2	5				x					
nlp uned team	5	9	7				x					
QMUL-SDS	11	4	7,5				x					
Alclatos	10	6	8		x			x				
ZK	9	8	8,5			x	x	x				
GuillemGSubies	7	14	10,5	x	x							
IREL hatespeech group	14	7	10,5					x				
Codec		12	12	x	x							
S_exist	12	13	12,5					x			x	
MiniTrue	13		13	x	x	x						
UMUTeam	16	11	13,5	x	x							
Free	6	24	15			x		x				
ZZW	15	18	16,5				x					
Zimtstern	17	16	16,5			x						
LaSTUS	18	15	16,5			x						
Recognai	26	10	18		x							
Andrea Lisa	21	17	19			x						
CIC	19	19	19	x					x		x	
MessGroupELL	20		20			x		x				
MB-Courage	22	22	22	x								
Nerin	24	20	22							x	x	
Soumya	23	23	23						x		x	
UNEDBiasTeam	25	21	23							x		
BilaUnwanPk1	27	25	26									x
Almuoes3		27	27						x	x	x	
ORDS_CLAN	29	26	27,5									x
Uja	28		28	x	x							

Table 1
Results and techniques used in the EXIST 2021 competition [33] sorted in order of average ranking. Random Forest (RF) stands for Random Forest, Logistic Regression (LR) stands for Logistic Regression.

- **Task 1: Identifying Sexist Content** In this task, the system is supposed to perform binary classification. It must determine whether a given text (tweet or gab) exhibits sexism, whether directly, by describing a sexist scenario, or by criticising sexist behaviour.
- **Task 2: Categorising Sexist Content** Following the identification of sexist content, the subsequent task involves categorising the content based on the type of sexism present.

The results, and techniques used in the two sub-tasks in the competition are compiled in Table 1. Some takeaways from the approaches in the competition comes from the datasets including Spanish and English text. When participants used Beto (Spanish version of BERT), it was exclusively used to analyse the Spanish texts which means that it had to be combined with other models for English. The same is true for BERT, almost all participants that used BERT for the English texts also ended up using other models for the Spanish texts. Lastly, the most common and best performing LLMs for handling multiple languages in this competition were mBERT and XLM-R.

	# of speeches	% of task data	% of all data
Left	58,146	39.0	16.2
Right	90,797	61.0	25.3
Power	111,127	53.1	31.0
Opposition	98,114	46.9	27.4

Table 2
Distribution of the opposition and power labels in the power dataset as well as Left and Right labels in the orientation dataset.

3. Dataset

Touché provides a dataset [34] for our task which contains a selection of speeches from the ParlaMint corpora [35]. This dataset contains data from parliamentary speeches in multiple European parliaments—we include the countries covered in the dataset in Appendix A. More precisely, the dataset consists of two separate subsets, one for each sub-task. These subsets are also divided into multiple sub-subsets each containing data from only one country.

The organisers altered the dataset to provide less information than the original one, but also includes an automatic translation to English for most non-English texts. The provided training dataset consists of 6.5GB of text files [34] divided among the orientation and power datasets and contains the following fields:

id is a unique (arbitrary) ID for each text.

speaker is a unique (arbitrary) ID for each speaker. There may be multiple speeches from the same speaker.

sex is the (binary/biological) sex of the speaker. This information is collected from varying sources (typically data published by the respective parliament), and in some cases it may be unspecified or unknown.

text is the transcribed text of the parliamentary speech. Real examples may include line breaks, and other special sequences escaped or quoted.

text_en is an automatic English translation of the corresponding text. This field may be empty for speeches in English. There might be missing translations for a small number of non-English speeches.

label is the binary/numeric label. For political orientation, 0 is left and 1 is right. For power identification 0 indicates coalition (or governing party) and 1 indicates opposition.

Data imbalance There is an uneven distribution of data where some countries have more data than others. In addition, the label distribution among the countries is also skewed. For instance, in Figure 1 we can see that Serbia has more data for speeches connected to right wing parties while for the power dataset Serbia has more speeches from speakers in power. Table 2 represents the overall label distribution of the dataset.

The text lengths of each country also vary and are displayed in Figures 2 and 3 for the orientation dataset. Shorter texts may contain less helpful information for the predictions and thus, decrease the performance. Moreover, the limited context length of the models may not be able to capture all relevant information in the longer texts. For instance, BERT has a context length of 512 tokens which means that it can not process the entirety of most texts at once.

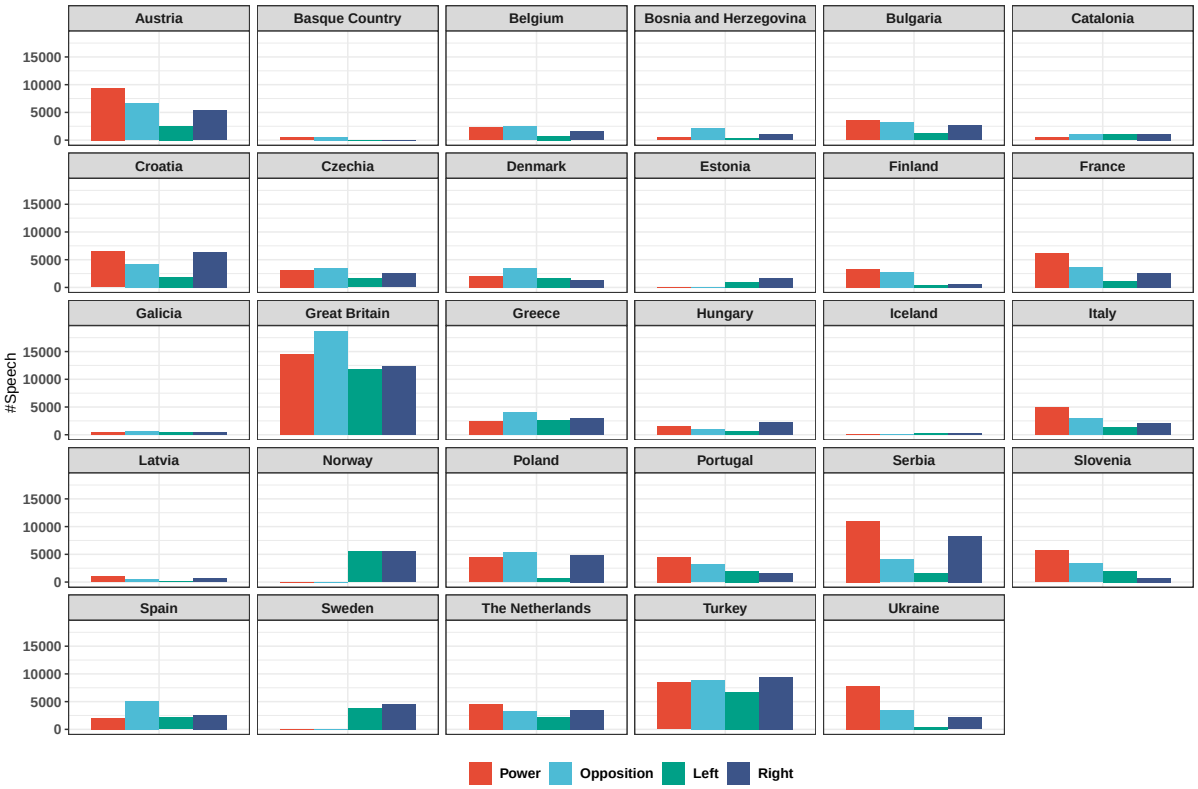


Figure 1: Distribution of the Opposition, Power, Left and Right-wing labels for each country in the orientation dataset.

Shared information Even though the datasets are split, 47.5% of the speakers that appear in one of the datasets appear in both datasets, as shown in Figure 4. Moreover, 51.4% of speeches in the power dataset are made by a speaker who also appears in the orientation dataset. This amount equates to 72.2% of the total number of speeches in the orientation dataset.

Isolated speeches The dataset does not include the date of when the parliamentary speeches occurred. This means that modelling changes in party ideas over time or fully encapsulate how politics shift is infeasible. Moreover, simply connecting a speech to a certain ideology will not be as useful when it comes to predicting whether the speaker is currently in a governing position or opposition. This is because a country with a left leaning government one year could have a right leaning government the next.

Even though speeches are originally part of debates and exchanges in parliament, the dataset contains no data for connecting multiple speeches to a single conversation or exchange. This prevents approaches which would model an entire debate and label the participants in the debate rather than the individual speeches themselves.

Privacy in the dataset The dataset uses arbitrary codes for people's names. This makes it difficult to check if our results match up with real-world politicians. Additionally, the test set for the orientation task does not contain any speakers in the original orientation dataset. As a result, a solution which attempts to connect specific speeches to the correct political parties using the speakers identity is infeasible.

Test set The test set for the orientation sub-task contains randomly sampled speeches by speakers who do not appear in the training set. The test set for the power sub-task does, to a large extent, contain speakers that also appear in the training set. However, speakers recurring in the test set will tend to

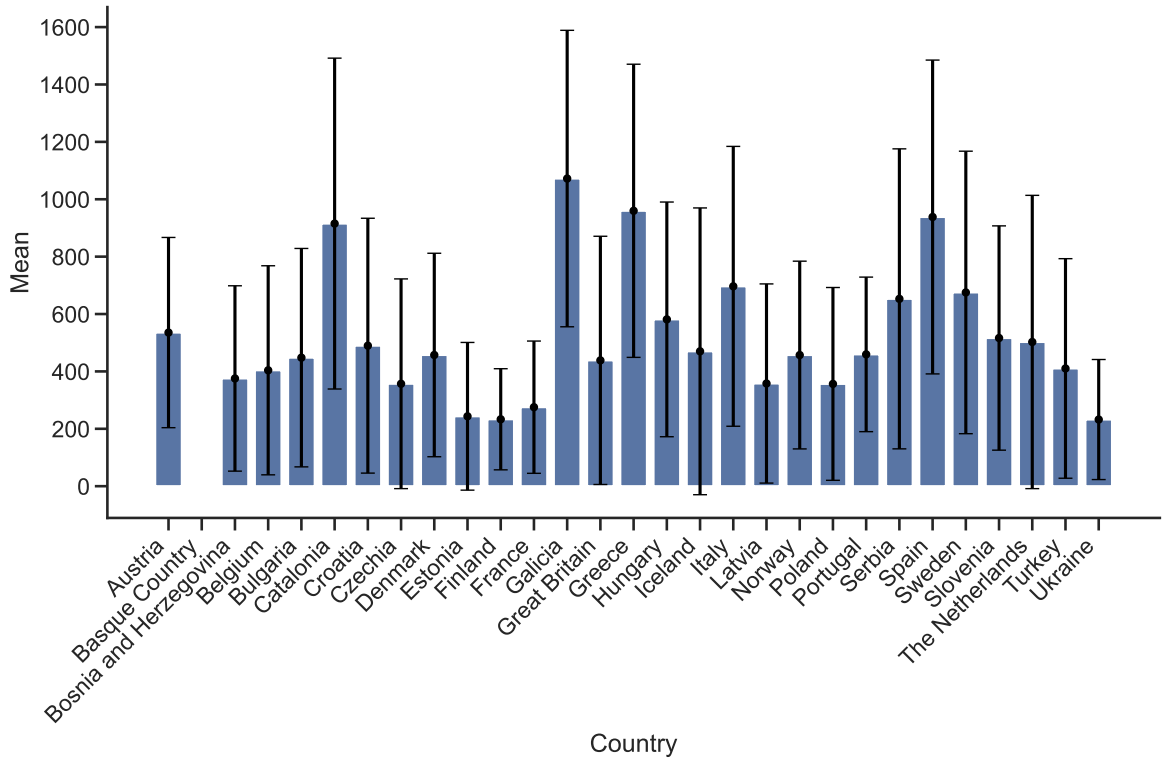


Figure 2: Mean and standard deviation for text lengths per country in the orientation dataset.

have a different label distribution compared to the training set. For both sub-tasks, the test set contains approximately 2,000 speeches for each parliament, whose general label distributions resemble those in the training set. The test data follows a similar structure to the training data apart from the speaker_id and label field being hidden.

4. System Overview

In the following, we describe how we processed and prepared the data as well as how we selected and trained models. An overarching view of our method is illustrated in Figure 5.

4.1. Dataset and Preprocessing

We decided not to extend the dataset using external sources. This was partly due to other parliamentary debates in our selection of countries either being unavailable or already in the original ParlaMint corpora. It would also require a lot of work in order to create properly labelled datasets in the same format as our base dataset. Finally, the amount of available data is of substantial size and a larger amount would put further pressure on the need for computational resources.

We generated the training dataset for each task by sampling 70% of the provided datasets, leaving the remaining 30% of the datasets for the validation set. Note that the 70/30 split is a commonly used rule-of-thumb which has shown some empirical optimality [36, 37].

We used this split for evaluating the individual and combined parts of our model. Once we got the most effective techniques, we switched to a different split for our final ensemble model as explained in Section 4.6.

Missing translations Firstly, some English translations were missing in the provided Finnish dataset. Specifically, there were 271 translations missing in the training dataset and 88 in the test set. These

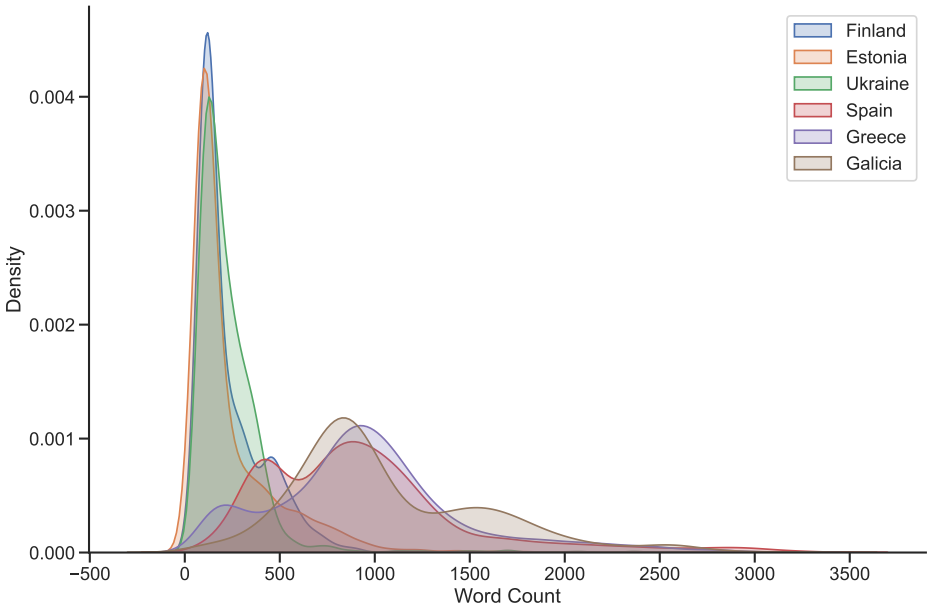


Figure 3: Density distribution of text lengths for the 3 countries with the highest, and lowest average text lengths in the orientation dataset.

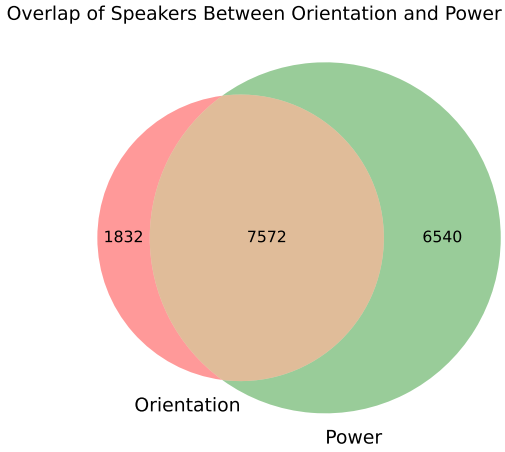


Figure 4: Venn diagram where the red region represents the number of speakers that only appear in the orientation dataset ($Orientation \setminus Power$). The green region represents the number of speakers that only appear in the power dataset ($Power \setminus Orientation$). The centre region represents the number of speakers that appear in both datasets, in other words the intersection ($Orientation \cap Power$).

texts had to be manually machine translated. For the translations we used a Python package called `mrTranslate`.

4.1.1. Back-translation

To address country distribution imbalance, we applied back-translation to the data from countries with fewer than 15,000 entries in the power and orientation datasets combined. We chose this threshold in order to strike a balance between improving the amount of speeches available for parliaments with less representation in the dataset and not increasing the training time excessively. The back-translation process involved translating English text to the original language and back, followed by translating the

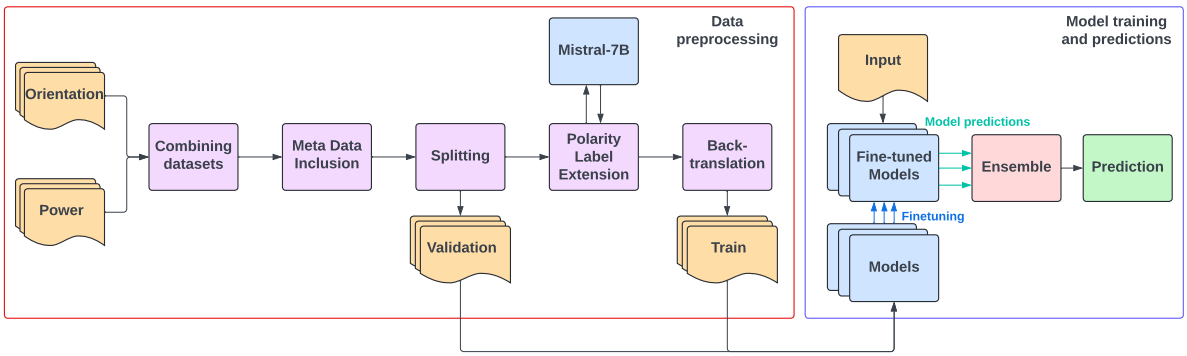


Figure 5: Overarching method illustration. The method is divided into a data preprocessing stage and a model training and prediction stage. The test set is inserted as input, but it could potentially be any speech.

original text to English and back to the original language. We also used *mrTranslate* for the translations and appended the resulting data to the dataset, keeping all other fields unchanged from the original entry. Sometimes, however, the translations would fail. In these cases, we manually translated the texts using Google Translate.

4.1.2. Meta data inclusion

By prepending each text with the corresponding country and gender of the speaker, i.e., "Germany, Female", models got access to the available contextual information not included in the speeches themselves. We hypothesise that since parliaments and contexts of debates vary, so should their analyses. By giving the model access to all available metadata, i.e., all available context, we suspect that models might be able to better adapt their predictions. An example of adapting to such a context might be adapting the prediction of someone advocating for a certain law depending on what the law is currently in a given country. There might also be useful information in the meta data itself, such as gender making a politician more likely of belonging to a certain ideology in some countries. These examples are of course purely speculative and therefore part of the reason why we chose to include all meta data instead of selecting fields based only on our own speculations.

4.2. Combined training

We merged the datasets by combining them where the label column was split into orientation and power labels. Despite being separate, both datasets contain shared elements, allowing the extraction of some data in the power dataset data into the orientation dataset, as explained in Section 3. We first verified that each speaker in the orientation dataset consistently had the same label. Once a speaker was confirmed as label y in the orientation dataset, all texts by that speaker in the power dataset were also classified as label y for orientation. When the datasets shared a text, the text from the orientation dataset was removed (since the text from the power dataset already had received the orientation label). This increased the number of orientation entries by 72.2% which equates to 51.4% of the original power dataset.

4.2.1. Located missing labels loss function

Since the combined dataset has many missing labels, we needed to create a custom loss function. We calculated a filter tensor for each batch and label, marking rows with true labels as one and rows without true labels as zero. Then, we used this filter tensor as weights for the entries in the batch when computing the CE-loss. To account for the number of incomplete labels, we divided each label-specific loss by the sum of its filter tensor. By summing the loss for our two labels, we had created a multi-label loss function which could account for Located Missing Labels (LML).

Let us consider this loss function for a single label, i.e., orientation. Formally, let C be the set of classes (i.e., left and right) and p_i the output predictions for all classes $c \in C$ for entry i in the batch. $p_{c,i}$ is then the predicted probability of class c for entry i in the input batch such that $p_{c,i} \in [0, 1]$ and $\sum_c p_{c,i} = 1, \forall i$. Also let y be the true labels such that $y_i \in \{-1\} \cup C$ (where -1 corresponds to a missing label) represents the true label of entry i. Then our custom LML-Loss can be expressed as Equation (1).

$$\begin{aligned} \textbf{\textit{LMLLoss}}(p, y) &= \frac{1}{\sum_i f_i} \sum_i f_i \textbf{\textit{CELoss}}(p_i, y_i) \\ f_i &= \begin{cases} 1, & y_i \in \{C\} \\ 0, & y_i = -1 \end{cases} \end{aligned} \tag{1}$$

As each label-specific loss is divided by the sum of its filter tensor, the resulting loss maintains a consistent size regardless of the number of samples with a true label in the batch. Consequently, the final summed loss represents the combined loss for each task, regardless of its prevalence in the batch.

4.3. Polarity label extension

If training to predict both the power and orientation labels at the same time yields increased performance, then training to predict a third label might yield further improvements. We chose polarity as the label to add to our dataset, i.e., whether a text carries a positive, negative or neutral sentiment. We chose polarity since it is an effective metric for identifying trends in parliamentary speeches.

To obtain polarity labels for our dataset, we used version 0.2 of the instruction fine-tuned Mistral-7B, as available under *mistralai/Mistral-7B-Instruct-v0.2* on Hugging Face. We chose this model since it outperforms other open source LLMs of similar or larger size, such as 7- and 13-billion parameter versions of Llama-2 [38]. We chose one example text for each polarity label and had GPT-3.5 explain why it assigned that label to that text. With these examples, we constructed our in-context learning prompt using 3-shot classification. The base textual prompt, which we then formatted into the instruction chat format of the model, can be found in Appendix B.

We double-quantised the Mistral model to a 4-bit normal float with a 16-bit float compute type to fit the model in memory and for faster inference. To fit the entire base prompt along with the text to label, we defined a context length of 4,096. To generate from the model, we used sampling beam search with 3 beams along with forcing the output to contain “Positive”, “Negative”, or “Neutral”. Finally, we assigned the polarity label as 0, 1 or 2 depending on whether the first word of the output was “Negative”, “Positive” or “Neutral” respectively, assigning -1 otherwise.

To perform the polarity classification, we added three output nodes to our classification head, corresponding to each polarity class. To account for holes in the polarity data caused by failed generations, we also used our LML-loss to calculate the loss from the polarity predictions. We only added half of the polarity LML-loss to the base LML-loss in order to prioritise our two core tasks.

4.4. Models

We restricted the models to those we could effectively train. Therefore, we excluded more extensive and capable models, such as the 70-billion parameter version of Llama-2 [39]. Furthermore, we were forced to limit hyperparameters, such as batch size and learning rate, to less-than-ideal values to comply with our limited computational resources. Our selection of models was also influenced by the notion that encoder only models outperform modern large decoder models for similar tasks, at a lower computational demand [20].

We compared different modern transformer-based models to find which models performed the best for our task. Different models necessitated different hyperparameter values due to their differing sizes and designs. For all models, any implemented classification heads took the place of the last layer of the model as provided by their Hugging Face sequence classification implementation, leaving the method of pooling as implemented in the base model.

BERT and mBERT We evaluated the uncased version of BERT [40], available under *bert-base-uncased*, using the provided English translations. We chose this model since it is competent while being much smaller than other modern models (see Section 2.1) and since it had previously shown outstanding performance in similar competitions (see Section 2.5).

We also evaluated the uncased version of multilingual BERT, available under *bert-base-multilingual-uncased*, on the speeches in their original languages. We chose this model because it is a multilingual version of BERT and because it demonstrated excellent results for multilingual tasks in similar competitions (see Section 2.5).

RoBERTa and XLM-RoBERTa We evaluated the large version of RoBERTa, available under *FacebookAI/roberta-large*, since it has been shown to be state-of-the-art for similar tasks (see Section 2.1). We also evaluated XLM-RoBERTa, available under *FacebookAI/xlm-roberta-large*, which is RoBERTa pre-trained for multi-lingual tasks [41].

DeBERTa V3 DeBERTa V3 is an improvement upon the original DeBERTa model [42]. The original DeBERTa model outperforms the large version of RoBERTa using less training data on a wide range of NLP tasks [43]. The DeBERTa family of models also utilise *distangled attention* and relative position embeddings which allow it to process longer sequences than BERT and RoBERTa. Due to these factors, we chose to evaluate DeBERTa V3, as available under *microsoft/deberta-v3-large*.

Gemma The 7-billion parameter version of Gemma outperforms similar models of equal size such as Mistral-7B and Llama-2-7B [44]. Limited by our computational resources, we evaluated the smaller 2-billion parameter version of Gemma, as available under *google/gemma-2b*. This smaller version still necessitated using techniques such as LoRA and double quantising the model to 4-bit normal float. We applied LoRA to all matrices in the self attention- and mlp-layers of Gemma.

4.5. Hyperparameters

Due to our limited computational resources, unfortunately, no experiments could be exhaustive. When we discovered that a certain hyperparameter value worked well, for instance using a warm-up period, we could not then afford to repeat experiments for all previous models to include this choice. Efforts where instead directed as to for each model balance the necessity to fit to the data in a manageable time frame with the desire not to cause excessive unlearning in the base model.

The main parameters which had to be adapted depending on the size of the models was the learning rate and the warm-up ratio. For instance, if a large model was trending downwards by the second epoch, then the learning rate might be lowered and/or a warm-up period added, in this way experiments where exploratory.

We set specific hyperparameters to increase training speed and reduce memory consumption to accommodate larger models. For instance, all models except BERT and mBERT used 16-bit floating point mixed-precision training to accelerate the training process. BERT and mBERT used regular full-precision training. Additionally, Gemma required LoRA and 4-bit quantisation to make fine-tuning feasible for us.

All models trained using a maximum sequence length of 512 tokens. Whilst we would have preferred to train with longer sequence length for the models which can handle it (DeBERTa-V3 and Gemma), this was not computationally feasible. However, to still utilise the longer context lengths DeBERTa-V3 can handle, it was re-evaluated using a maximum sequence length of 4096 tokens after training was complete. In this way we could train and evaluate the models more efficiently with 512 tokens and then afterwards leverage longer context lengths for only the final version of the model. We would have preferred to also re-evaluate Gemma using a longer context length, however, due to unforeseen limitations on computational resources, only DeBERTa could be re-evaluated using a longer sequence length.

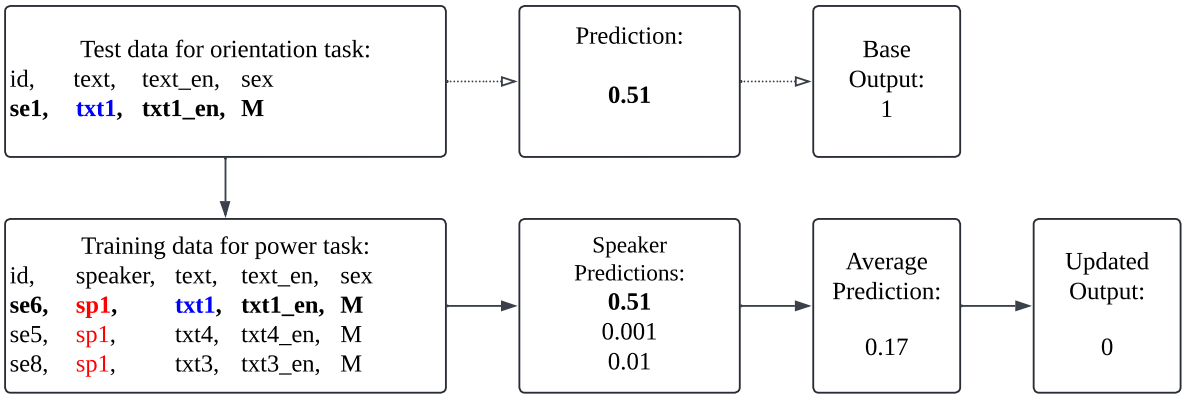


Figure 6: Visualised process for extracting the speaker for texts in the test data for the orientation task. The bottom flow chart then shows how we take the average prediction for all texts by speaker *sp1*. Note that we illustrate the predictions as ranging from 0 to 1 for simplicity whilst in actuality the logits were used.

4.6. Ensemble modelling

After identifying the best performing models and training methods using our validation set, we created new training and validation sets which we used to re-train the selected models for ensemble modelling. These new validation sets contained disjoint selections of 10% of the available data, with a minimum of 5 samples for each country and label.

We chose this approach to increase the amount of data available to our models. By each model in the ensemble having a separate validation set, each model could be monitored for over-fitting whilst the ensemble as a whole had still trained on the entire dataset. We speculate that by using this approach, our ensemble will be able to leverage the entire training set. This is since if a single model has not been able to learn something useful due to the required data being in its validation set, the other models of the ensemble will have had access to that data.

We also chose to decrease the ratio of the validation set as we deemed the necessity of it being representative and reliable to be diminished once we had already determined and validated our method.

We created the ensemble by selecting the best performing multilingual and English-only models. We then trained two instances of the best performing of the two models as well as once instance of the other model using our newly created ensemble training and validation sets. For a given prediction, we ran each of these models and averaged their output logits before applying the sigmoid function and rounding to receive a final prediction.

4.7. Additional data extraction for test set

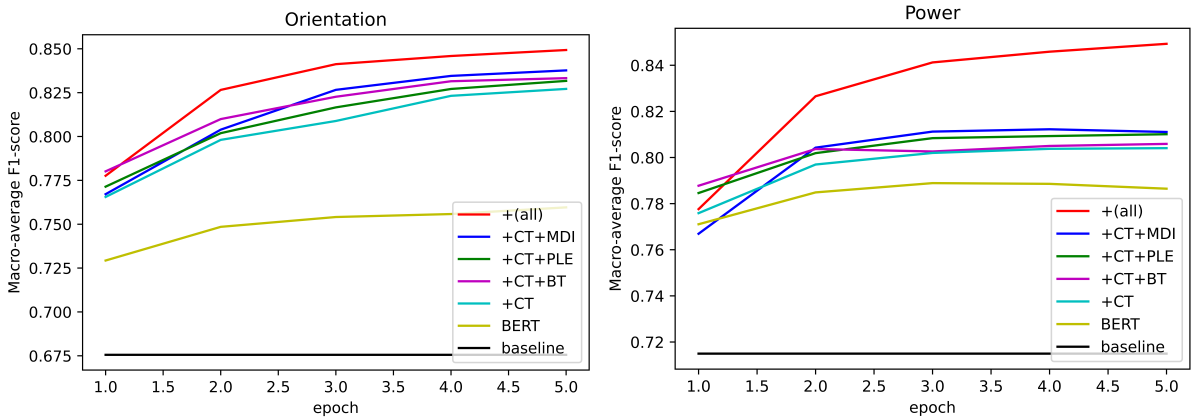
Roughly 23% of the texts appearing in the provided test data for the orientation task also appear in the training data for the power task. Using this information we can extract the speaker id for the overlapping texts. Then, since the orientation label is always the same for each speaker, we can use these additional speeches to influence our prediction on the test set. For a given speech in the orientation test set, we averaged the logits of the examined model on that speech and the logits of our best performing model on all other speeches by the same speaker. In other words, predictions on the test data were averaged with those produced by our best model on speeches by the same speaker. This way if the text in the test data lacks clear ideological signals, we can instead rely on other texts by that speaker to make our prediction. The process is visualised in Figure 6.

5. Results

In the following, we present the results of our method application to the provided dataset and evaluated models.

Sub-task	Macro-average F1-score
Orientation	0.6755
Power	0.7149

Table 3
Baseline macro-average F1-scores on our validation set for Orientation and Power tasks from provided linear logistic regression model.



(a) Orientation (b) Power

Figure 7: Comparison of base BERT and with different combinations of the components of our method over the training epochs. CT stands for combined training, BT for back-translation, PLE for polarity label extension and MDI for meta data inclusion.

We fine-tuned the models using the transformers library for Python as provided by Hugging Face, running on a NVIDIA V100 with 32GB of memory. Also, if not mentioned, we set the hyper-parameter values to their default values as provided by the library.

The competition organisers provided a baseline of a simple linear logistic regression model. When we fitted this model to our training set and then applied this model to our validation set, we achieved the macro-average F1-scores as shown in Table 3.

5.1. Method components

To understand the effects of each component of our method, we fine-tuned BERT multiple times with different combinations of the components in our base method. These components, as presented in Section 4, were combined training, back-translation, polarity label extension and meta data inclusion. All runs used the same hyper-parameters, which can be found in Appendix C.1. Results are illustrated over the training epochs in Figure 7 and summarised in Table 4.

The results as a whole show that the combined training beats both the conventionally trained BERT and the baseline. The other components all individually improve performance of the combined training further. Finally, all components together yielded the best performance.

Combined training (CT) Training to predict both labels at once using our LML-loss showed significant improvements in comparison to when only training for a single label at a time.

Back-translation (BT) Our results indicate that back-translation yielded an improvement in the orientation task over all epochs whilst only yielding an non-marginal improvement in the first epochs of the power task when training for both tasks using CT. To further investigate if back-translation helped improve the performance of countries with less data, we also visualised the results for each

Model	Orientation	Power
baseline	0.6755	0.7149
BERT	0.7596	0.7865
+CT	0.8271	0.8041
+CT+BT	0.8333	0.8059
+CT+PLE	0.8317	0.8101
+CT+MDI	0.8377	0.8111
+(all)	**0.8493**	**0.8152**

Table 4
BERT component study showing how combinations of components in our method impacted the macro-average F1-score for the two tasks compared to conventional fine-tuning. CT stands for combined training; BT for back-translation; PLE for polarity label extension, and; MDI for meta data inclusion.

parliament individually. Results can be found in Appendix D.2 and show that, on average, parliaments with back-translation saw a significant improvement whilst the remaining parliaments did not.

Polarity label extension (PLE) Extending the combined training by adding a third label yielded an increase in performance over all epochs.

Meta data inclusion (MDI) Including the available meta data by prepending it to each text resulted in a improvement by the second epoch and onwards for both tasks. However, for the first epoch it caused a decrease in performance for the power task whilst not impacting the orientation task.

Method components conclusions The examination of the components in our method seems to indicate that all components of our method are beneficial for BERT. This is especially clear due to the combination of two factors. The first factor is that the conventionally trained model had seemingly started to stagnate or over-fit whilst our proposed method was still improving through out all training epochs. The second factor is that our method exceeds the conventional fine-tuning already by the first epoch. In combination then, we may reason that our method provides an intrinsic advantage since it both converges faster, by the second factor, and does seems to cause less over-fitting or unlearning, by the first factor. These factors may also be reasoned as to guaranteeing that we actually make better update steps instead of just smaller (by the second factor) or larger (by the first factor).

In order to to validate that the improvement in performance on the orientation task from our method is not only due to the increased amount of data, we conventionally fine-tuned RoBERTa on the full set of available orientation data. We compare this to RoBERTa fine-tuned using the same hyperparameters, as found in Appendix C.2, but using our full method. Results are illustrated in Figure 8 and show that even when using the same training data, our method exceeds conventional fine-tuning over all epochs for orientation classification. As discussed in Section 2.1, fine-tuned RoBERTa has been shown to outperform very capable models and to be state-of-the-art for similar tasks. It is therefore very encouraging to note that our method managed to significantly improve upon the performance of fine-tuned RoBERTa for political orientation classification.

5.2. Models

The highest attained scores resulting from the application of our method on various models are shown in Table 5. Corresponding hyperparameters can be found in Appendix C.2. Results indicate that DeBERTa-V3 was the best performing model, with XLM-RoBERTa being the best performing multilingual model. Gemma, which trained using LoRA and quantisation, manages to exceed the performance of BERT and mBERT but falls short of the other models.

To investigate the impact of re-evaluating DeBERTa-V3 using a longer sequence length, we also evaluated different sequence lengths. The results are shown in Table 6 and indicate that there was a

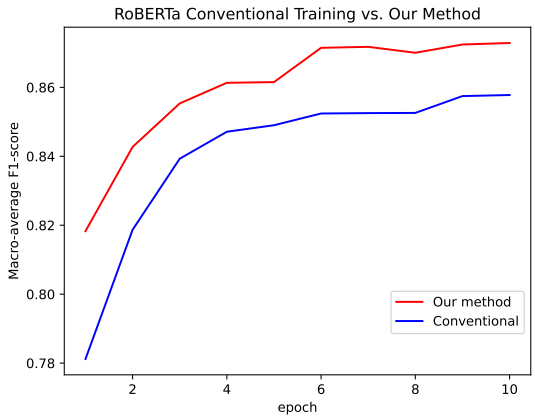


Figure 8: Comparison of RoBERTa trained for political orientation classification using conventional fine-tuning and our method.

Model	Language*	Orientation	Power
BERT	Translation	0.8493	0.8152
mBERT	Original	0.8251	0.7941
RoBERTa	Translation	0.8729	0.8440
XLM-RoBERTa	Original	0.8621	0.8379
DeBERTa-V3	**Translation**	**0.8870**	**0.8630**
Gemma	Translation	0.8541	0.8358

*Translation corresponds to training on automatic translations to English instead of the original language.

Table 5
Highest attained macro-average F1-scores of our examined models.

Sequence length	Orientation	Power
512	0.8788	0.8526
1024	0.8847	0.8610
2048	0.8869	0.8627
4096	**0.8870**	**0.8630**

Table 6
Macro-average F1-scores of DeBERTa when evaluated using different sequence lengths.

significant improvement in performance from increasing the sequence length initially but that these increases diminish. The improvement by going from 512 to 1024 tokens was noticeable (+0.0059 and +0.0084) whilst the improvement by going from 2048 to 4096 tokens was minor (+0.0001 and +0.0003). These findings are not surprising, after all, successive increases in maximum sequence length add fewer and fewer tokens since more speeches become fully covered.

5.3. Translated vs. multilingual

To investigate whether models pre-trained for multilingual tasks outperform their mainly English-comprehending base models, we compared BERT and mBERT as well as RoBERTa and XLM-RoBERTA. Each pair used the same hyperparameters internally (see Appendix C.2). The multilingual models processed the original texts whilst their counterparts processed the automatic translations. Results indicate that the multilingual models lag behind by a consistent amount. The macro-average F1-scores over the training epochs can be found in Appendix D.1.

Model	Orientation	Power
XLM-RoBERTa	0.8621	0.8379
DeBERTa-V3	0.8870	0.8630
Ensemble	**0.8972**	**0.8697**

Table 7
Macro-average F1-scores on our base validation set of base models and the ensemble of them.

Model	Orientation	Power
baseline	0.5603	0.6401
Ensemble	**0.7945**	**0.8271**

Table 8
Macro-average F1-scores on test set of our final ensemble and the provided baseline model

5.4. Ensemble modelling

In order to validate that ensemble modelling was a beneficial approach, we averaged the output logits of DeBERTa-V3 and XLM-RoBERTa, fine-tuned on our base training set, on our base validation set. The macro-average F1-scores of the yielded predictions, as seen in Table 7, show that the predictions of our best performing stand-alone model could be improved by also considering the outputs of our best performing multilingual model.

5.5. Test set results

Baseline The macro-average F1-scores attained by the baseline model on the competition test set are shown in Table 8. This baseline model was fitted to the entirety of the original provided datasets. When comparing this baseline to the baseline on our validation set, we see a decrease in macro-average F1-score of 0.1152 and 0.0748 for the orientation and power task respectively. This indicates that the test set is much more challenging, which is not surprising due to the nature of its construction. For the orientation task the test set contains speakers that do not appear in the training set, and for the power task it contains speakers which appear with a different role than they do in the training data. The test set also does not share the same distributions in parliament representation, for further details see Section 3.

Final ensemble model Our final ensemble consisted of two DeBERTa-V3 models and one XLM-RoBERTa model, fine-tuned using disjoint selected validation sets, as detailed in Section 4.6. We have made these models available on Hugging Face under *oscpalML/DeBERTa-political-classification*, *oscpalML/DeBERTa-political-classification-alternative* and *oscpalML/XLM-RoBERTa-political-classification*. Our final ensemble, averaging the logits of these models, yields macro-average F1-scores as seen in Table 8.

Additional data extraction The additional data extraction improved the performance on the orientation task. Our ensemble, without considering the other available speeches by a speaker, yielded a macro-average F1-score for the orientation task of 0.7854 whilst when utilising the other speeches the score increased to 0.7945.

6. Discussion

In the following chapter, we discuss and reason about the effectiveness of the proposed solution. We further discuss the task itself and the limitations of our project.

Method effectiveness We show that our method improves performance for BERT and RoBERTa. Whether or not these findings translate to other models and tasks is, of course, a prudent question. The limitations of this project, enforced upon us by our limited computational resources, prevent us from examining this quandary fully. However, due to the relatively similar architectures of our examined models and the nature of our method, we hypothesise that the benefits of our method does translate. This is because our method does not closely depend on the internals of a model, instead aiming to provide a more representative loss function and better input data. We leave more extensive empirical confirmation for future research.

Multi-task training The combined training yields an increase in performance for the orientation task, which is not surprising since we extract additional orientation labels and therefore provide the model with more data. On the other hand, the fact that there is also a significant improvement for the power task is very intriguing. We speculate that this improvement is due to our LML-loss providing a more representative loss, which incentivises extracting features which are useful for both tasks and discourages over-fitting. This seems to be supported by the additional increase in performance provided by also adding the prediction of a polarity label. This increase in performance is even more impressive when considering that the polarity label was synthetically generated and likely to add at least some amount of noise. Since polarity likely shares some important similarities with features useful for our tasks, we speculate that our LML-loss was improved as to further incentivise cross-task useful features.

Data preprocessing The observed impact of prepending available meta data to each speech, as shown in Figure 7, is not inexplicable. We suspect that the prepended sentence is very different from the pre-training material of the base model, since it is simply two words and does not follow the form of a regular phrase or sentence. This disruption, we speculate, might essentially confuse the model until it is able to learn it in later epochs. Once the model has understood how to interact with the prepended sentence however, it is able to leverage it into making better predictions.

It would be interesting for future research to compare the difference between adding new tokens representing the meta data and prepending the meta data in English, as we did. It might be the case that base models are able to leverage prior understanding of countries, be it their general political environment or some other aspect. On the other hand, it might also be the case that prior bias hurts the models ability to predict accurately and fairly.

Nature of tasks On the validation set, it is interesting to note that the linear baseline performed better on the power task than the orientation task given that the models utilising our method show the opposite behaviour. We may also note that whilst the difference is small, the power task benefited more from a longer sequence length. It is therefore not entirely unreasonable to suggest that the power task might rely more on specific words and phrases, as the linear baseline does. In other words, it might be the case that specific words are more important for predicting political power whilst how you speak in general is more indicative for predicting political orientation.

Difference between validation and test scores Another notable aspect is the discrepancy between the achieved scores on the validation and test sets. Without using additional data extraction, our top-performing single model achieved a macro-average F1-score on the validation set that was 10.1 percentage points higher than that of our best ensemble model on the test set. In contrast, the gap for the power task was just 3.6 percentage points. We attribute the majority of this gap to two factors: the difference in the distribution of the amount of parliamentary data and the nature of the test sets' construction.

Since all parliaments have the same amount of data in the test sets whilst having greatly differing amounts of data in the training and validation sets, it is not surprising that the achieved scores differ. In the likely case that the models perform better on parliaments with more data, then the test set represents an increased presence of the harder-to-predict parliaments and a decreased presence of the

easier parliaments. Regardless of this factor, just the difference in distribution itself likely also introduces a challenging condition. In other words, that there is a drift in label and parliament distributions is likely detrimental, even if the nature of that drift was not suspected to be particularly damaging. This is because the model likely to some extent relies on the statistical trends, such as favouring to predict the more common label when a speech is ambiguous.

The nature of the test sets' construction may also account for the difference in discrepancy between our two tasks. The power test set largely contains the same speakers as our training data, just with a different power label. Given that we saw a relatively small decrease in performance, it seems that our models have been able to avoid over-fitting to a specific speakers power label. Such a case was likely aided by speakers exhibiting multiple power labels in the training data.

This behaviour is not shared with the orientation task. Since the orientation test set largely consists of speakers who do not appear in the training data, the gap in performance may indicate that our models have overfit to specific speakers. An additional factor is that the new speakers may cover topics our models have previously not encountered. If a speaker is mainly exhibiting a political idea which the model has not previously learnt the ideological connotations of, then the classification likely becomes much more challenging. It would therefore be interesting to investigate whether these previously not encountered speakers are contemporary to and covering the same topics as the speakers in the training set.

Weak ideological signals Some parliamentary speeches might not indicate strong political beliefs. They could solely cover practical proceedings, not expressing any opinions or making any arguments. These types of speeches are likely more challenging to classify, especially if the provided dataset does not exhibit clear rhetorical or linguistic differences for labels within a given class. This likely introduces an upper limit on the performance of any model performing this task with similar data.

Limitations As previously discussed, our limited access to computational resources determined what methods and models we could examine. This prevented us from examining large models such as the 70-billion parameter version Llama-3 or even the the 7-billion parameter version of Gemma. Not only were we limited in the selection of models, but also in the number of experiments which we could perform. More time and computational resources would have allowed us to attempt more techniques and further search for optimal hyperparameters. Techniques which could not be examined include using different learning rates for different layers and balancing the loss-function.

We were also limited by the data which we had access to. In real life, these speeches are not stand-alone but most often parts of exchanges and debates. The problem of weak ideological signals could likely be mitigated by considering all the speeches a speaker makes in an exchange together. By representing speeches as parts of a larger debate, not only could a model base its prediction on all of the speakers speeches, but also on the speeches made by the other participants in the debate.

7. Conclusion

In this study, we proposed a method for improved fine-tuning of LLMs for ideology and power identification. Our research questions where as stated below.

RQ1 Investigate what the best methods and practices are for identifying the political orientation in a parliamentary speech.

RQ2 Investigate what the best methods and practices are for identifying whether a parliamentary speech is made by a speaker in opposition or in power.

In answering our research questions, we found that modern LLMs are an effective approach for identifying both ideology and power in parliamentary debates. We further found that ideology and power likely share useful features and therefore fine-tuning to predict them jointly yields improved

performance for both tasks. This improvement also extends to fine-tuning to predict synthetic labels, in our case polarity.

We also note that performance can be improved by making the context of the speech, as available by meta data, available to the models. Furthermore, back-translation can be utilised to boost performance of countries with a smaller presence in a given dataset. Finally, we found that English models predicting on automatic translations tend to outperform multilingual models predicting on the original languages but that an ensemble of both types of models is the best approach.

Acknowledgments

We would like to acknowledge our supervisor, Pablo Picazo-Sanchez, for his continuous guidance, feedback and assistance in this project. Furthermore, we want to express our gratitude for the computational resources provided by the Data Science and AI division at Chalmers University of Technology and University of Gothenburg. Finally, we would like to thank our examiner Moa Johansson for providing feedback on early and intermediary versions of this paper.

References

[1] S. Coleman, Meaningful political debate in the age of the soundbite, in: Televised election debates: International perspectives, Springer, 2000, pp. 1–24.

[2] H. Wasmuth, E. Nitecki, (Un)intended consequences in current ECEC policies: Revealing and examining hidden agendas, Policy futures in education 18 (2020) 686–699.

[3] M. J. Hinich, M. C. Munger, Analytical politics, Cambridge University Press, 1997.

[4] Pew Research Center, Americans' Dismal Views of the Nation's Politics, https://www.pewresearch.org/politics/2023/09/19/americans-dismal-views-of-the-nations-politics/, 2023. Accessed on November 29, 2023.

[5] M. K. David, Language, power and manipulation: The use of rhetoric in maintaining political influence, Frontiers of Language and Teaching 5 (2014) 164–170.

[6] Ö. Sahin, Ö. Sahin, A gentle introduction to ML and NLP, Develop Intelligent iOS Apps with Swift: Understand Texts, Classify Sentiments, and Autodetect Answers in Text Using NLP (2021) 1–15.

[7] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (2014) 1093–1113.

[8] P. Törnberg, ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning, arXiv preprint arXiv:2304.06588 (2023).

[9] Université Grenoble Alpes, CLEF 2024 - conference and labs of the evaluation forum, https://clef2024.imag.fr/, 2024. Accessed on January 17, 2024.

[10] Webis Group, Touche, https://touche.webis.de/, ????. Accessed on January 17, 2024.

[11] Webis Group, Ideology and Power Identification in Parliamentary Debates 2024, https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html, ????. Accessed on January 17, 2024.

[12] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[13] L. Derczynski, Complementarity, F-score, and NLP Evaluation, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 261–266.

[14] M. Bosley, M. Jacobs-Harukawa, H. Licht, A. Hoyle, Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research (2023).

[15] H. Yu, Z. Yang, K. Pelrine, J. F. Godbout, R. Rabbany, Open, closed, or small language models for text classification?, arXiv preprint arXiv:2308.10092 (2023).

[16] Y. Chae, T. Davidson, Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning, Open Science Foundation (2023).

[17] W. Zhang, Y. Deng, B. Liu, S. J. Pan, L. Bing, Sentiment analysis in the era of large language models: A reality check, arXiv preprint arXiv:2305.15005 (2023).

[18] M. Abdullah, O. Altiti, R. Obiedat, Detecting propaganda techniques in English news articles using pre-trained transformers, in: 2022 13th International Conference on Information and Communication Systems (ICICS), IEEE, 2022, pp. 301–308.

[19] K. Sprenkamp, D. G. Jones, L. Zavolokina, Large language models for propaganda detection, arXiv preprint arXiv:2310.06422 (2023).

[20] A. Benayas, M. A. Sicilia, M. Mora-Cantallops, A comparative analysis of encoder only and decoder

only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance (2024).
[21] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, R. Arora, Pre-training BERT on domain resources for short answer grading, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6071–6075.
[22] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.
[23] Smartling, What is back translation and why is it important?, 2023. URL: https://www.smartling.com/resources/101/what-is-back-translation-and-why-is-it-important/.
[24] S. Edunov, M. Ott, M. Ranzato, M. Auli, On the evaluation of machine translation systems trained with back-translation, arXiv preprint arXiv:1908.05204 (2019).
[25] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.
[26] S. Shleifer, Low resource text classification with ulmfit and backtranslation, arXiv preprint arXiv:1903.09244 (2019).
[27] Y. Yu, Z. Zhou, X. Zheng, J. Gou, W. Ou, F. Yuan, Enhancing label correlations in multi-label classification through global-local label specific feature learning to fill missing labels, Computers and Electrical Engineering 113 (2024) 109037.
[28] Y. Zhang, Y. Cheng, X. Huang, F. Wen, R. Feng, Y. Li, Y. Guo, Simple and robust loss design for multi-label learning with missing labels, arXiv preprint arXiv:2112.07368 (2021).
[29] X. Zhang, R. Abdelfattah, Y. Song, X. Wang, An effective approach for multi-label classification with missing labels, in: 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), IEEE, 2022, pp. 1713–1720.
[30] Z. Ma, S. Chen, Expand globally, shrink locally: Discriminant multi-label learning with missing labels, Pattern Recognition 111 (2021) 107675.
[31] B. Wu, Z. Liu, S. Wang, B.-G. Hu, Q. Ji, Multi-label learning with missing labels, in: 2014 22nd International conference on pattern recognition, IEEE, 2014, pp. 1964–1968.
[32] Q. Jia, J. Cui, Y. Xiao, C. Liu, P. Rashid, E. F. Gehringer, All-in-one: Multi-task learning bert models for evaluating peer assessments, arXiv preprint arXiv:2110.03895 (2021).
[33] U. N. Group, EXIST: sEXism Identification in Social neTworks, 2021. URL: http://nlp.uned.es/exist2021/, accessed on January 28, 2024.
[34] Çöltekin, Ç., M. Kopp, V. Morkevičius, N. Ljubešić, K. Meden, T. Erjavec, Training data for the shared task Ideology and Power Identification in Parliamentary Debates, https://doi.org/10.5281/zenodo.10450641, 2024.
[35] CLARIN ERIC, ParlaMint: Harmonised Parliamentary Corpora, 2021. URL: https://www.clarin.eu/parlamint, accessed on November 24, 2023.
[36] K. K. Dobbin, R. M. Simon, Optimally splitting cases for training and testing high dimensional classifiers, BMC medical genomics 4 (2011) 1–8.
[37] Q. H. Nguyen, H.-B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, I. Prakash, B. T. Pham, Influence of data splitting on performance of machine learning models in prediction of shear strength of soil, Mathematical Problems in Engineering 2021 (2021) 1–15.
[38] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[39] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
[41] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[42] P. He, J. Gao, W. Chen, DeBERTav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[43] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).
[44] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).

A. List of countries in the dataset

- Austria (at)
- Bosnia and Herzegovina (ba)
- Belgium (be)
- Czechia (cz)
- Denmark (dk)
- Estonia (ee) [only political orientation]
- Spain (es)
- Catalonia (es-ct)
- Galicia (es-ga)
- Basque Country (es-pv) [only power]
- Finland (fi)
- France (fr)
- Great Britain (gb)
- Greece (gr)
- Croatia (hr)
- Hungary (hu)
- Iceland (is) [only political orientation]
- Italy (it)
- Latvia (lv)
- The Netherlands (nl)
- Norway (no) [only political orientation]
- Poland (pl)
- Portugal (pt)
- Serbia (rs)
- Sweden (se) [only political orientation]
- Slovenia (si)
- Turkey (tr)
- Ukraine (ua)

B. Polarity base prompt

Label the polarity of the following text, similarly to the provided examples. Your answer needs to start with “positive”, “negative” or “neutral”, followed by a short justification for your answer. It is important that you only assign a positive or negative label if you are sure of your answer. Here is your first example.

Text: The south-west was cut off from the UK last winter and Network Rail performed miracles in getting that line back up and running. I therefore find it extraordinary that reasons such as the weather have been used to excuse the chaos and incompetence of this debacle, particularly out of King's Cross. Why did the Secretary of State feel that it was not necessary for Ministers to ask for a basic reassurance that an overrun on any of the big programmes could be managed? Why were contingency plans not in place, and why was the rail regulator warning not adhered to?

Negative. The text expresses frustration and criticism towards the handling of infrastructure issues, particularly the failure to address problems with the rail system despite previous incidents. It highlights perceived incompetence and lack of planning, suggesting a negative sentiment towards the situation.

Here is your second example.

Text: We are committed to ensuring that claimants receive high-quality, objective, fair and accurate assessments. The Department monitors assessment quality through independent audit. Assessments deemed unacceptable are returned to the provider for reworking. A range of measures, including provider improvement plans, address performance falling below expected standards. <p> I do agree with the hon. Lady, which is why we have been trying to work more strategically with Motability, thrashing through the issues I am very aware of on appeals and on matters such as when an individual leaves the country. We are looking to reduce the amount of time that appeals take and at what we can do with the running of the scheme so that the precise scenario she outlines does not happen.

Neutral. The text describes the commitment to ensuring quality assessments for claimants and outlines measures taken to monitor and address assessment quality. Additionally, it mentions efforts to work with Motability to improve processes and reduce appeal times. The tone is informative and focused on addressing issues, without expressing overt positivity or negativity.

Here is your third example.

Text: I congratulate the hon. Gentleman on bringing this much needed debate to the Floor of the House. Will he join me in paying tribute to local MND associations across the United Kingdom for the invaluable support they provide? I know of the excellent work of my local Leicestershire and Rutland association, having heard at first hand from a constituent and friend of mine, Ruth Morrison, about her tragic personal experience. The support that is available is of immense value and I hope the hon. Gentleman will join me in paying tribute to the work of those associations.

Positive. The text expresses gratitude and admiration for the efforts of local Motor Neurone Disease (MND) associations, highlighting the invaluable support they provide. It also encourages acknowledgment of their work, suggesting a positive sentiment towards their contributions.

Now here is your text to label:

C. Hyperparameters

C.1. BERT method components

The hyperparameters which where not left to their default values as provided by the transformers library are shown in table 9.

Hyperparameter	Value
Learning rate	3e-5
Epochs	5
Batch size	40
Weight decay	0.001
Input length	512

Table 9
Non-default hyperparameter values for the BERT method components experiment.

Hyperparameter	BERT	mBERT	RoBERTa	XLM-R	DeBERTa-V3
Learning rate	3e-5	3e-5	1e-5	1e-5	2e-5
Epochs	5	5	10	10	10
Warm-up ratio	0	0	0	0	0.2
Batch size	40	40	20	20	10*
Weight decay	0.001	0.001	0	0	0
Train input length	512	512	512	512	512
Eval. input length	512	512	512	4096	2048

*Gradients where accumulated for 2 steps to simulate batch size 20 for DeBERTa-V3

Table 10
Best performing examined non-default hyperparameter values for various models

Hyperparameter	Gemma
Learning rate	3e-5
Learning rate schedule	constant
Epochs	3
Warm-up ratio	0.1
Batch size	12
Train input length	512
Eval. input length	512
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Quantisation type	4-bit NormalFloat
Quantisation compute type	16-bit float
Double quantisation	True

Table 11
Best performing examined non-default hyperparameter values for Gemma

C.2. Model hyperparameters

The best performing hyperparameters used for models which did not utilise LoRA can be found in table 10, and the ones for models which did utilise LoRA in table 11. Not mentioned hyperparameter values where left to their defaults as provided by the transformers library.

D. Figures and illustrations

D.1. Multilingual model performance figures

The comparison of models and their multilingual counterparts is illustrated in fig. 9.

D.2. Back-translation impact on specific parliaments

The difference in macro-average F1-score when fine-tuning combined training BERT with and without back-translation is visualised for each parliament in fig. 10.

D.3. Power dataset illustrations

The power dataset is illustrated inf fig. 11 and fig. 12.

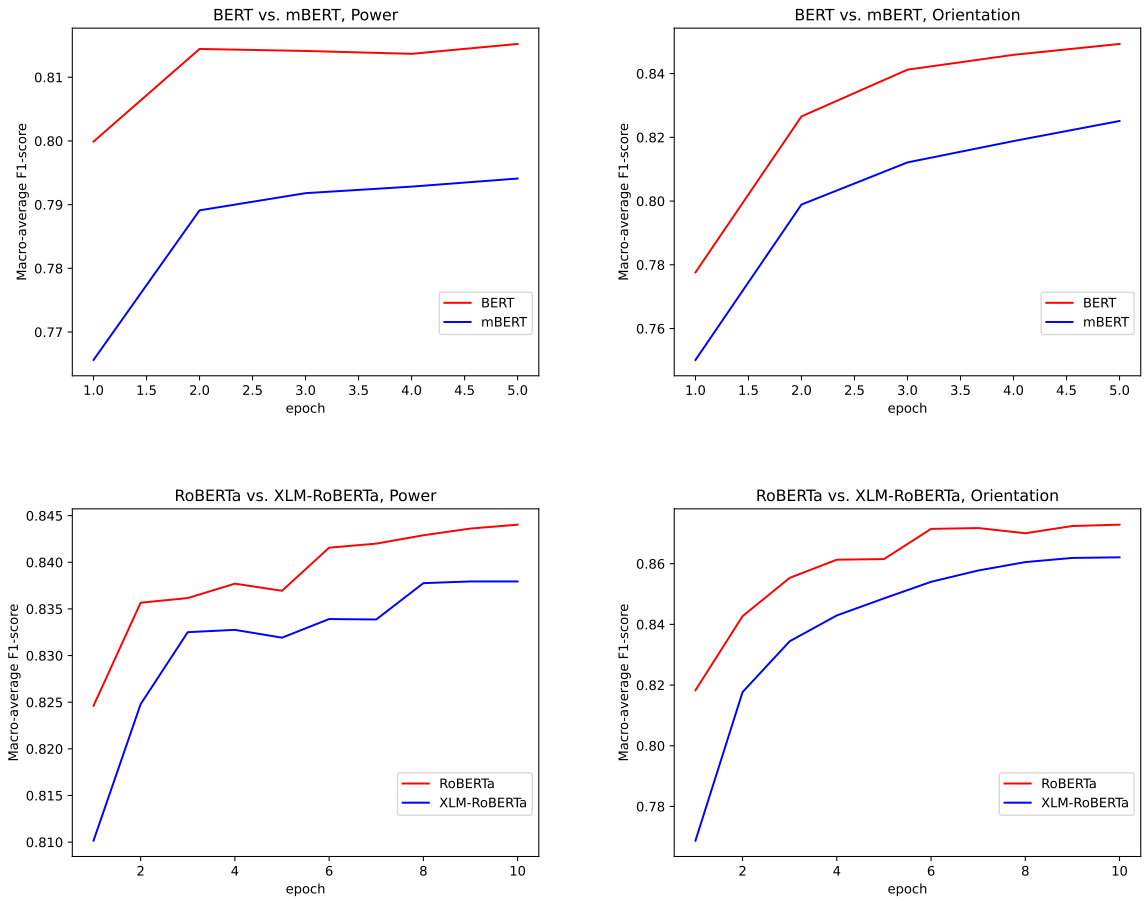
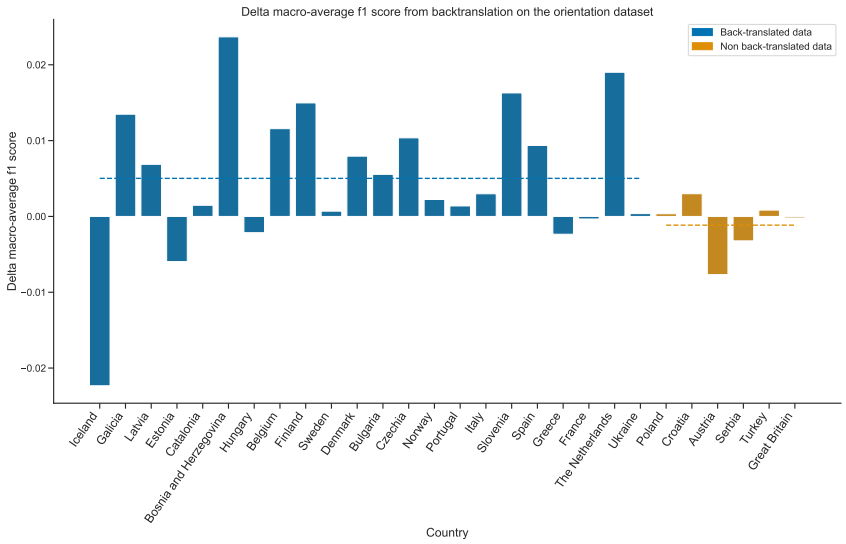
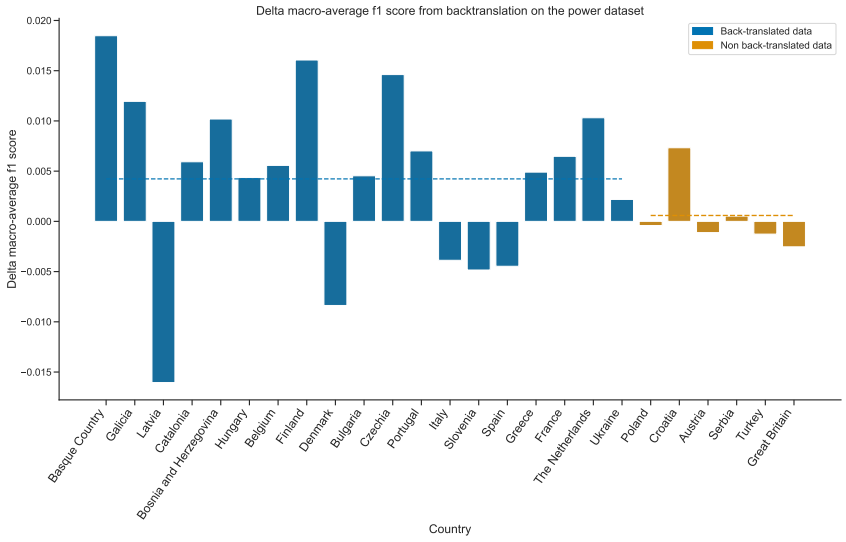


Figure 9: The comparison of BERT and RoBERTa fine-tuned on the English translations vs. mBERT and XLM-RoBERTa fine-tuned on the original texts.



(a) Orientation



(b) Power

Figure 10: Difference in macro-average F1-score comparing combined training BERT with and without back-translation on the orientation task. The countries are ordered by the number of speeches before back-translation, from left to right in increasing order.

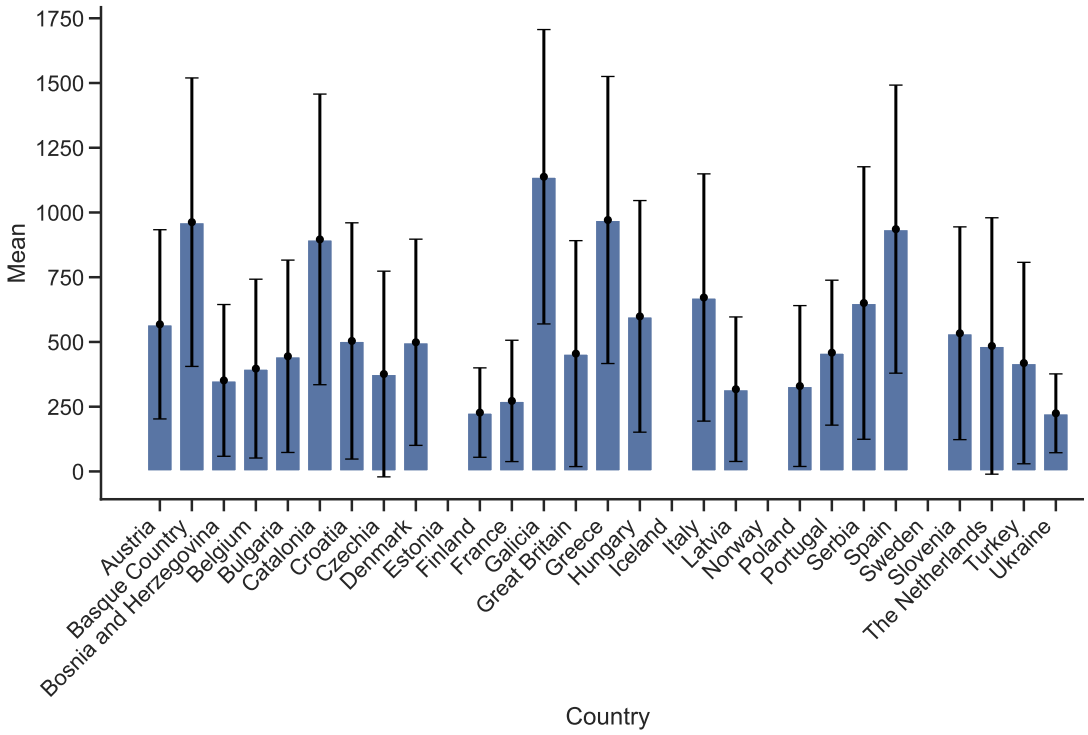


Figure 11: Mean and standard deviation for text lengths per country in the power dataset.

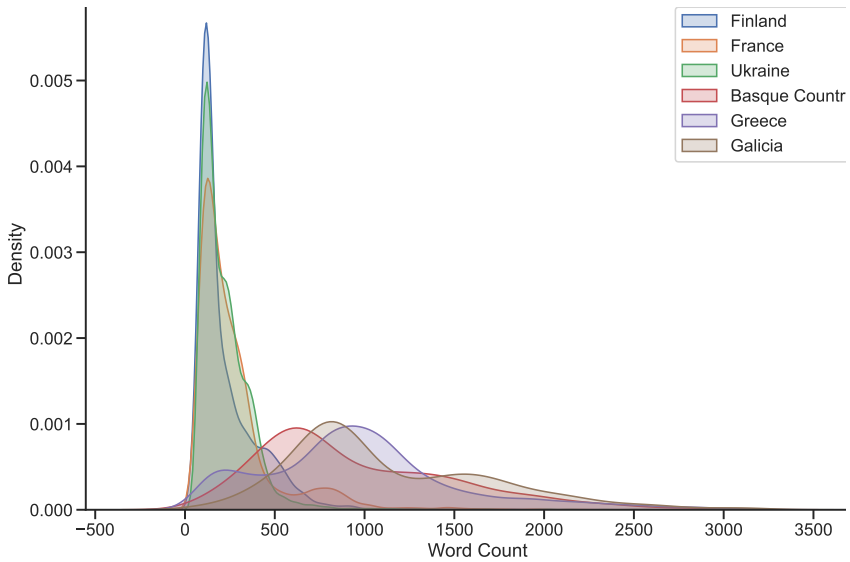


Figure 12: Density distribution of text lengths for the 3 countries with the highest, and lowest average text lengths in the power dataset.