# An Analysis of Politicians' Tweets: A multilingual approach

Drena Miftari, Francesco Gandolfi, Yabra Muvdi

Project supervisors: David Rossell, Nandan Rao, Omiros Papaspiliopoulos

June 25th 2020

# Contents

# 1 Introduction

Social media platforms have become a fundamental tool for politicians during their election campaigns. These platforms allow them to reach a broad audience of citizens and mobilize them into voters. Twitter, in particular, has become widely used among politicians. This paper aims to characterize politicians' Twitter communication during elections and study how this communication is shaped by a politician's country and ideological affiliation. To do this, we will analyze a dataset containing tweets authored by politicians or political parties during the month before the 2017 elections in France, the United Kingdom, and Germany. Three main questions will guide our investigation: (i) what are the main themes found in politicians' Twitter communication? (ii) what differences —or commonalities— can be found in how politicians of each country communicate these topics? (iii) how much are politicians using each one of these themes and how is this usage varying along national and ideological lines?

Several authors have conducted in-depth studies on politicians' Twitter usage during electoral periods and its potential impact on voters within a single country (Graham et al. 2013; Kruikemeier 2014; Stier, Bleier, Lietz, et al. 2018). However, these studies have not been extended to study the behavior of politicians in multiple countries jointly. Partly, this could be due to the additional methodological difficulties faced when analyzing texts in various languages. Our paper tries to fill this gap by exploring different tools for multilingual text analysis and formulating a Structural Topic Model (Roberts, Stewart, and Airoldi 2016) that permits the discovery of topics and their relation with selected tweet-level metadata (e.g., country, political party).

The topics estimated from this model provide some clues for answering the questions formulated above. Firstly, we were able to see that most topics were not related to substantial issues (e.g., welfare, health, education) but instead served a basic informational function (e.g., advertising events, inviting people to vote). Secondly, politicians' word usage for substantial topics displayed larger differences across countries than their word usage in informational topics. Thirdly, thanks to the inclusion of tweet-level metadata in the model, we were able to see that topic usage varied significantly both at the country and the political family level. However, topic usage seems to display a higher variation between countries than between political families, suggesting that politicians' Twitter content is more country-specific than ideology-specific.

The remainder of this paper is structured as follows. Section 2 reviews both the literature on political communication and probabilistic topic models. Section 3 describes in detail the data used and our preprocessing strategy. Section 4 develops in full depth the STM model and presents our main findings. Section 5 details two different approaches to verifying some of the results presented in the previous section. Section 6 concludes.

# 2 Literature Review

## 2.1 Analyzing Political Texts

Language has been long recognized to play a central role in the study of politics (Grimmer and Stewart 2013). This interest has only been expanded by the vast amount of digitized political texts that are now available (e.g., speeches, campaign manifestos, social media activity). However, analyzing these large corpora poses new challenges for researchers; their size makes them virtually unreadable and costly to annotate manually. Thus, researchers interested in systematically analyzing large-scale text collections have turned quantitative approaches to text analysis (Grimmer and Stewart 2013; Roberts, Stewart, and Airoldi 2016). The literature provides various examples of the application of these techniques to specific political texts. Grimmer (2010) analyzes the topics of 24,000 press releases from US senators and estimates the attention each senator allocates to these topics using a Bayesian hierarchical topic model. Quinn et al. (2010) analyze US senators' speeches from 1997 to 2004 using a dynamic topic model to understand how the US senate agenda has changed through this period. Focusing on Europe, Proksch and Slapin (2010) use speeches from EU parliamentarians to estimate their relative ideological position. Similarly, Greene and Cross (2017) use a dynamic topic model to analyze the same corpus and detect latent themes and their evolution from 1999 to 2014. These analyses provide a useful starting point for our work. However, they focus on corpora that significantly differ from ours. Thus the need to complement this literature with one that focuses on a similar type of political text.

## 2.2 Political Communication on Twitter

Social media's growth as a tool used by politicians to communicate with a broader audience marked the appearance of a new type of political text. Analyzing this new source of data has generated several relevant research avenues. Ausserhofer and Maireder (2013), for example, analyze Austria's national *Twittersphere* by reconstructing the network formed by users and by investigating the topics of more than 140,000 tweets. The authors manually classified each of the tweets in the sample according to one of 16 predefined political topics and explored the popularity of these topics. Similarly, relying on manual labeling of tweets, Golbeck et al. (2010) studies how members of the US congress communicate through Twitter. All 4,959 tweets collected were labeled into classes showing that *informational* tweets were the most common type. Graham et al. (2013) reach a similar conclusion when analyzing tweets from candidates to the 2010 UK General Election. Analyzing 26,282 manually labeled tweets from 416 candidates, the authors show that candidates mainly use Twitter as a unidirectional form of communication with almost 25% of the collected tweets consisting of just *updates* from the candidate. Finally, and more relevant to our work, Stier, Bleier, Lietz, et al. (2018) study the 2013 German federal election campaign to understand the correspondence between the topics that candidates were discussing in their social media account (Facebook and Twitter) and the most relevant topics to the general public. Although the reviewed literature uses various methods to analyze political communication through Twitter, none of them seem to be entirely adequate for our corpus and questions. We will consequently review the broader literature on methods for statistical text analysis.

## 2.3 Probabilistic Topic Models

Probabilistic topic models are very powerful statistical methods for analyzing and uncovering themes across large sets of documents. These models are part of the larger field of generative probabilistic modeling in which data is treated is if it had been created from a generative process that includes latent – hidden to the observer – variables (Blei 2012). In the case of topic models, these latent variables can be interpret as the *topic structure* of the corpus of interest.

The *latent Dirichlet allocation* (LDA) model proposed by Blei, Ng, et al. (2003) and the *probabilistic latent semantic analysis* (pLSA) proposed by Hofmann (2001) constitute the two foundational probabilistic topic models. Following Boyd-Graber, Y. Hu, et al. (2017) we will focus on the LDA model because of its wide use in the literature and, most fundamentally, because it has served as the basis of several extensions relevant to our analysis. The basic idea of the LDA model is that documents can be represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, Ng, et al. 2003). LDA assumes that documents are generated by a two-stage process where:

1. a random distribution over topics is chosen
2. for each word in the document:
   (a) a random topic from the topics distribution is chosen
   (b) conditioning on the topic, a random word is chosen from the distribution over the vocabulary

In this model, topics are specified before the documents are generated and its defining characteristic is the fact that all documents share the same set of topics but exhibit them in different proportions.

This basic LDA model has been extended in several directions. Wallach (2006) relaxes the bag-of-words assumption, allowing for the order of the words to matter, by having topics that generate words conditioning on the previous word. Blei and Lafferty (2006) extend the model by allowing topics to change over time; relaxing the assumption that the order of the documents does not matter. This model is known as the dynamic topic model. More relevant to our work, several models have been proposed to incorporate a document's metadata into the analysis. Rosen-Zvi et al. (2004) propose the author-topic model where topic proportions are attached to each author and Eisenstein et al. (2010) develop the geographic topic model that allows topical content to vary by regions. In this same direction, Roberts, Stewart, and Airoldi (2016) propose the Structural Topic Model. This model allows a document's metadata to simultaneously affect how words are allocated to topics and the frequency of the terms within each topic. The use of this model will allow us to incorporate the rich metadata contained by a tweet (e.g., country and political party) into the estimation of topics. An in-depth description of the model will be given in Section 4.

None of the extensions of the LDA model covered until now deal with the additional complexity brought in by the analysis of text in multiple languages. Although nothing prevents us from estimating an LDA model for a multilingual corpus, this would inevitably lead to completely language-separated topics since words in different languages do not co-occur. Thus the need to search for a better strategy. Vulić et al. (2015) provide a great overview of the field of probabilistic topic models in multilingual settings. Several researchers have independently formulated models that extend the LDA model to multilingual contexts (De Smet and Moens 2009; Mimno et al. 2009; Ni et al. 2009). Importantly, these models require some type of alignment between the documents in different languages. In the simplest case, the corpus contains parallel documents: each document can be paired with its exact translation in other languages. European Parliamentary acts provide a perfect example of this type of corpus. More interestingly, these models can also accommodate corpora that are only theme-aligned: each document can be paired with documents in the other languages known to contain a similar proportion of themes. Wikipedia articles provide a good example of this type of corpora. Articles for the same subject (e.g., Brexit) in different languages are not exact translations of each other but are known to be covering the same themes. This assumption of thematic-alignment, however, imposes stringent restrictions on the type of corpora that can be analyzed. Boyd-Graber and Blei (2012) address this limitation by proposing a model that only assumes that, within a multilingual corpus, similar themes will be expressed in all languages. The model simultaneously learns the matching between terms across languages and the topics within languages. Again, this assumption does not fit our corpus. We will not expect, for example, the Brexit theme, as discussed in the UK, to have a direct equivalent in France and Germany.

## 2.4 Machine Translation

In the absence of a multilingual probabilistic topic model appropriate for our analysis, we explore the use of machine translation to convert all documents to a common language. Lucas et al. (2015) propose the use of mature translation systems[1] for this purpose and highlight the fact that, given the bag-of-words assumption of the STM, these systems need only to translate correctly the terms in the original document even if the ordering is not adequate. de Vries et al. (2018) evaluate the usefulness of machine translation for automated bag-of-words models, including LDA. Using data from EU parliamentary speeches the authors compare topic models learned on the machine translated text with models learned on the official translation. The authors show that both LDA topic models exhibit very similar topical prevalence and a strong overlap on the topic content. This motivates our use of machine translation for our multilingual corpus. Section 3 will detail our text preprocessing strategy.

## 2.5 Word Embeddings

To finalize this section, it is worth briefly mentioning a completely different modeling approach for working with text data. Instead of treating words as atomic units, word embeddings aim to generate a high-quality continuous vector representation of words. Models such as Word2Vec (Mikolov et al. 2013) or GloVe (Pennington et al. 2014) have become very popular for the quality of the word-vectors they produce. These models can be trained in a corpus in any language. However, each language will have its own continuous word embedding space. Conneau et al. (2018) address this limitation by construction a model that is able to align these spaces. The result is a continuous word embedding space that brings together multiple languages providing a very powerful tool for working with a multilingual corpus. Although much work has been done to produce these high-quality word vectors, the literature on using these continuous word representations as an input to topics models is still very recent (Das et al. 2015; Dieng et al. 2019; W. Hu and Tsujii 2016; Moody 2016; Wang et al. 2015). Exploring the use of these models in the context of cross county social media analysis would be a very relevant future stream of work to complement the results that we present here.

[1] The authors mention both the translation systems produced by Google and Microsoft

# 3 Data

## 3.1 Political Context

2017 was the year of the UK General Elections, German Federal Elections, and the French Presidential Elections. Given that the year was preceded by migration issues, national security threats due to various terrorist attacks in Europe, and skepticism on the sanctity of the European Union, politicians' communication during this time was critical.

The UK General Elections took place on June 8$^{\text{th}}$ 2017. The main parties involved were: The Conservative Party, The Labour Party, The Scottish National Party (Scotland), The Liberal Democrats, the Greens, the UK Independence Party, The Democratic Unionist Party (Northern Ireland), Plaid (Wales), and Sinn Fein (Northern Ireland). The Conservatives won 317 seats with 42 percent of the total vote but, albeit the largest single party with regard to seats and votes, did not manage to reach a parliamentary majority. The Labour Party won 262 seats, or 40 percent of the vote. Due to its small overall majority, The Conservatives formed a minority government with the DUP of Northern Ireland.

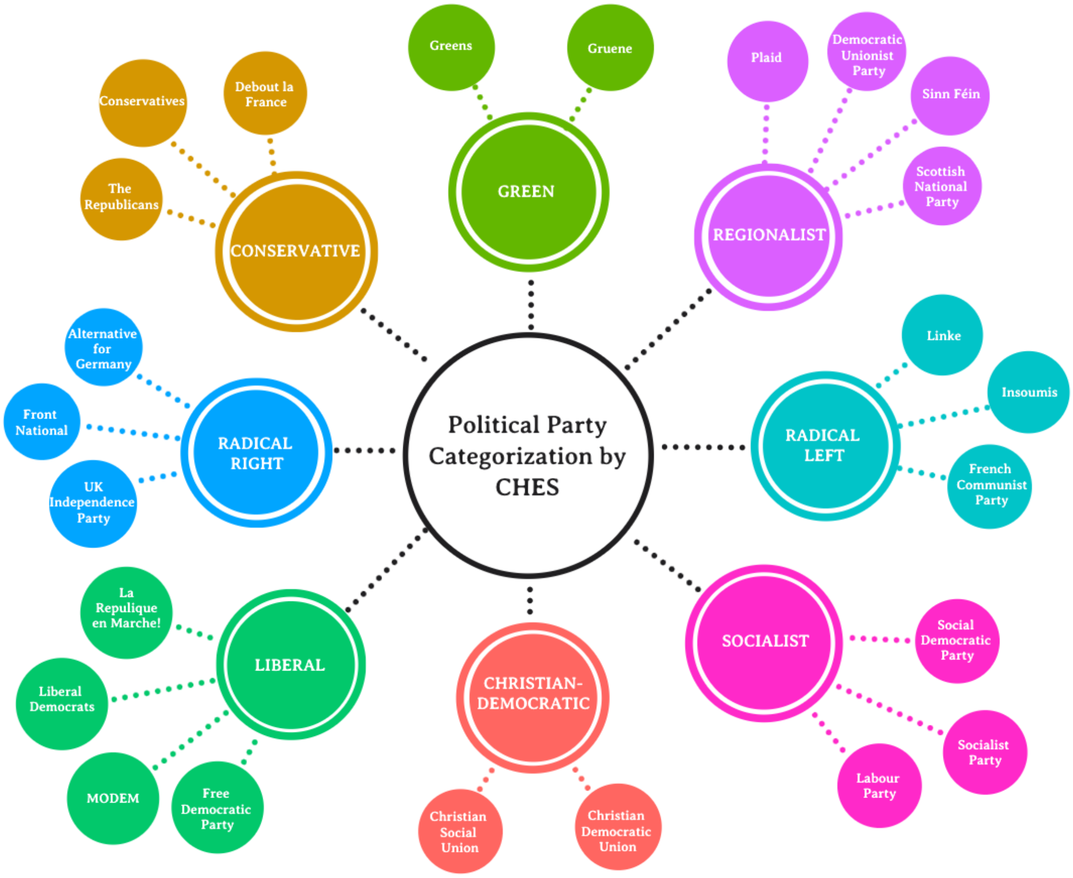
The German Federal Elections occurred on September 24$^{\text{th}}$ 2017, leading to the formation of the country's 19th Bundestag. The main parties involved were The Christian Democratic Union (CDU), Christian Social Union (CSU), The Social Democratic Party (SPD), the Alternative for Germany (AfD), The Free Democrats (FDP), The Left and The Greens. The CSU, led by Angela Merkel, triumphed the elections with a third of the votes, the highest percentage in the country, despite losing 8 percent of swing voters. The SDP's votes fell to 21 percent, and AfD – the country's far-right party – rose to a 12.6 percent vote count, placing themselves as the third political force in Germany.

The French Presidential Elections were held on April 23$^{\text{rd}}$ 2017, and as per the French election system since no party won a full majority, a second round was held on May 7$^{\text{th}}$ 2017. The final round consisted of Emmanuel Macron of the En Marche! party and Marine Le Pen of the Front National (FN) party. Macron won by approximately two-thirds of the votes. In addition to the two finalists, other main parties of this election were Debout la France (DLF), Socialist Party (PS), La France Insoumise (FI), MODEM, the French Communist Party (PCF), and The Republicans (LR).

## 3.2 Data Collection

Twitter enables users to communicate short statements (a tweet) to their audience which has enabled politicians to reach millions of individuals within seconds. In the first quarter of 2017, there were 325 million active users monthly on the platform (Clement 2019). Twitter is accessible to all through its website or mobile application and while it limits tweets to 140 characters in length, it also forces politicians to be concise. Our dataset was composed of 36,059 tweets authored by political parties and political candidates of the 2017 elections in France, Germany and the United Kingdom. These tweets were a subset of a larger dataset composed of over 4 million tweets collected by Majó-Vázquez et al. (2017), who collected all tweets through a Twitter API during the second and third quarter of 2017. The collection technique was structured such that each tweet had to consist of either: a preselected hashtag considered to be important for the elections or posted by accounts that pertained to

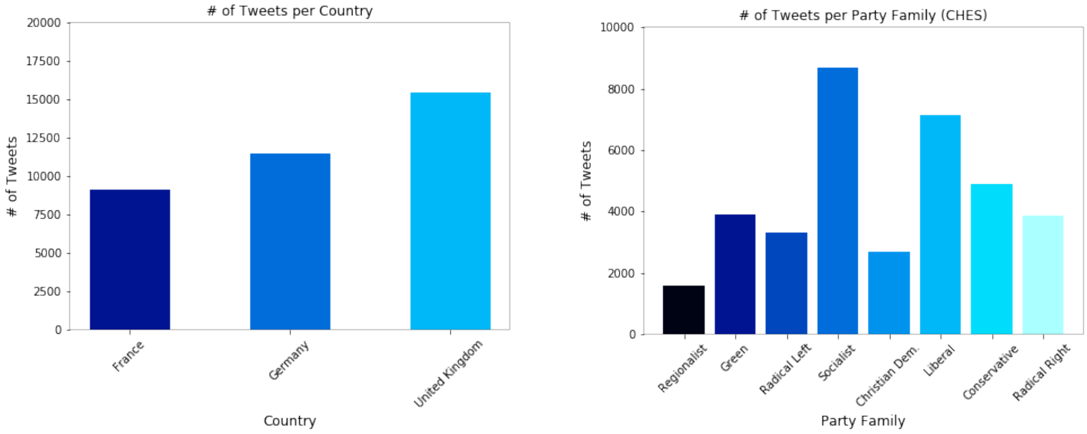
Figure 1: Overview of Political Parties in France, Germany and the UK as categorized by CHES data.



the relevant political parties or candidates. We extended the filtration of the data by selecting tweets that belonged only to a political party or a candidate account. To do this, we obtained information of election candidates (parties and individuals) for all three countries using various sources. For France, we downloaded historical parliamentary data through the NosDéputés (2019) platform, an open data website that collects information on past and current parliamentary members and their respective information. For Germany, we used the existing Bundestag database of Stier, Bleier, Bonart, et al. (2018), whereas for the UK we used an online source (Club 2017) which aggregates political data.

To enrich our understanding of political ideology and rhetoric of the politicians and parties, we merged our data with the Chapel Hill Expert Surveys – CHES – which estimates European party positioning on European integration, ideology, and non-EU policy issues such as immigration, redistribution, decentralization, and environmental policy (Polk et al. 2017). With this data, we were able to assign a predetermined ‘party family’ based on these surveys to our model s to differentiate

Figure 2: Breakdown of Tweet Data by Country and CHES Party Family



the various topics. These categories, which are shown in Figure 1, include a broad overview of 8 main categories: Regionalist, Radical Left, Radical TAN (Traditionalist, Authoritarian, Nationalist) – which can also be referred to as Radical Right –, Socialist, Christian Democratic, Liberal, Green and Conservative. The CHES data provided extra value to the Twitter data by adding already existing cross-country similarities between parties based on political agenda and party history which later allowed us to compare not only parties within countries, but also parties belonging to the same political family.

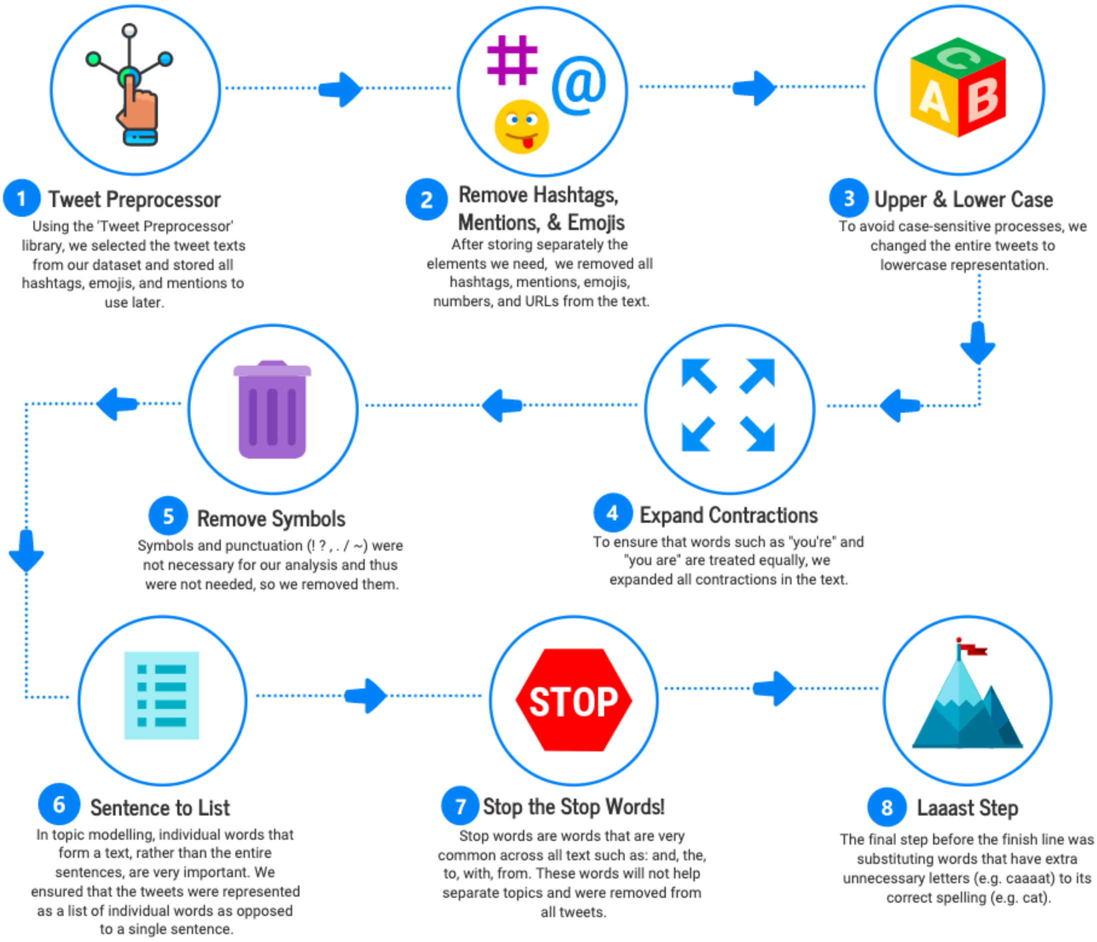
## 3.3 Data Cleaning and Preprocessing

Working with text and topic modeling requires a thorough text cleaning process to be undertaken as the first and most crucial step. Individuals have diverse writing styles and to ensure equality among words and phrases that are expressing similar things, many elements must be changed such as common words, misspelled words or expanding contractions. Figure 3 shows the entire cleaning and preprocessing journey we followed to get the best representation of our tweets. Given its strict character limit per tweet, Twitter incentivizes people to use short-form word representation, acronyms, emojis, or slang to deliver their message while staying within the character limit. Due to this, our data cleaning was essential in having high-quality input data for our model.

## 3.4 Translation vs. Embeddings

Following the cleaning and finalization of the data, there was another challenge to deal with before inputting the data into the model, namely the multilingual nature of the corpus. The goal of this paper was not only to analyze political rhetoric for the 2017 elections but also to analyze the case of Germany, France and the UK in unison. Thus, it required a sense of creativity on feeding

Figure 3: The data cleaning and preprocessing steps undertaken for this paper



1 Tweet Preprocessor

Using the 'Tweet Preprocessor' library, we selected the tweet texts from our dataset and stored all hashtags, emojis, and mentions to use later.

2 Remove Hashtags, Mentions, & Emojis

After storing separately the elements we need, we removed all hashtags, mentions, emojis, numbers, and URLs from the text.

3 Upper & Lower Case

To avoid case-sensitive processes, we changed the entire tweets to lowercase representation.

4 Expand Contractions

To ensure that words such as "you're" and "you are" are treated equally, we expanded all contractions in the text.

5 Remove Symbols

Symbols and punctuation (! ? , . / ~) were not necessary for our analysis and thus were not needed, so we removed them.

6 Sentence to List

In topic modelling, individual words that form a text, rather than the entire sentences, are very important. We ensured that the tweets were represented as a list of individual words as opposed to a single sentence.

7 Stop the Stop Words!

Stop words are words that are very common across all text such as: and, the, to, with, from. These words will not help separate topics and were removed from all tweets.

8 Laaast Step

The final step before the finish line was substituting words that have extra unnecessary letters (e.g. caaaat) to its correct spelling (e.g. cat).

multilingual text to a topic model. The first option considered was using word embeddings to represent the tweets. Most of the pre-trained word embeddings that currently exist are tailored specifically to each language. However, in 2017, the Facebook team created their own pre-trained word embeddings library, MUSE (Conneau et al. 2018), which is a collection of hundreds of languages, 30 of which projected onto the same space. This creates an opportunity to have documents in multiple languages and represent them numerically in an equivalent space. In addition to word embeddings, the other approach explored in this paper is using the Google Translate API, Google's online tool providing instant translations for over 100 languages, to efficiently translate all of the documents into English. Considering that Google Translate, itself, is trained on text found on the web, similar to word embeddings, it does not differ much from the first approach. In fact, Google

Translate is continuously updated, including feedback from the general public and native speakers, which make it a robust method to transform documents into the same language.

The approach chosen for the model in this paper was machine translation via the Google Translate API. This decision aligns with the common practices highlighted in the literature review (de Vries et al. 2018; Lucas et al. 2015). As a last step, we translated the entire multilingual corpus (without emojis, hashtags, and mentions) into English. The word embeddings remained as an exploratory tool for our data.

# 4 Results

## 4.1 Structural Topic Model

In the main analysis, we follow the Structural Topic Model (STM) proposed by Roberts, Stewart, and Airoldi (2016). As highlighted earlier, this model has the strength of allowing to incorporate document-level covariates. In particular, a document's metadata can be included in the model as content covariates or prevalence covariates. *Content covariates* influence the word rates that characterize each topic, while *prevalence covariates* affect the proportion of each document dedicated to each topic. The data generating process of the model is described in equations (1) to (5), and Figure 4 provides a graphical illustration of this process. Equations (1) and (2) characterize topical prevalence, equation (3) topical content, while equations (4) and (5) define the core language model. In all equations $d \in (1 \ldots D)$ identifies the document and $n \in (1 \ldots N_d)$ the position of a word within the document, while $k \in (1 \ldots K)$ indicates the topic, $v \in (1 \ldots V)$ a term in the vocabulary, and where $p$ is the number of covariates.

First, the topic prevalence coefficients $\gamma_k$ are generated as:

$$\gamma_k \sim \text{Normal}_p(0, \sigma_k^2 I_p) \quad \text{for } k = 1...K-1 \tag{1}$$

Given $\Gamma = [\gamma_1|...|\gamma_k]$, and the vector of prevalence covariates $x_d$ for document $d$, the share of each topic in the document is generated as:

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma' x_d', \Sigma) \tag{2}$$

In our model, the vector of prevalence covariates $x_d$ is given by a set of dummies that identify the country and a set of dummies that identify the political family of each document $d$.

Next, the probability of term $v$ of topic $k$ to appear in document $d$ is given by:

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})}{\sum_v \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})} \quad \text{for } v = 1...V \text{ and } k = 1...K \tag{3}$$

where $\kappa^{(t)}_{.,.}, \kappa^{(c)}_{.,.}$, and $\kappa^{(i)}_{.,.,.}$ are topical content model coefficients. In detail, $\kappa^{(t)}_{k,v}$ is the log-transformed rate deviation for term $v$ in topic $k$ over the baseline log-transformed rate for term v ($m_v$), and is shared across all levels of the covariates $Y$. $\kappa^{(c)}_{.y_d,v}$ is the deviation for level $y_d$ of the covariate and each term $v$, over the baseline for term $v$, and is shared across all topics. Similarly, $\kappa^{(i)}_{.,.,.}$ collects the covariate-topic interaction effect.

In our model, the vector of content covariates $y_d$ is given by the set of dummies that identify the country of each document $d$.

Finally, from $\theta_d$ we draw the topic of word $n$ in document $d$ as:

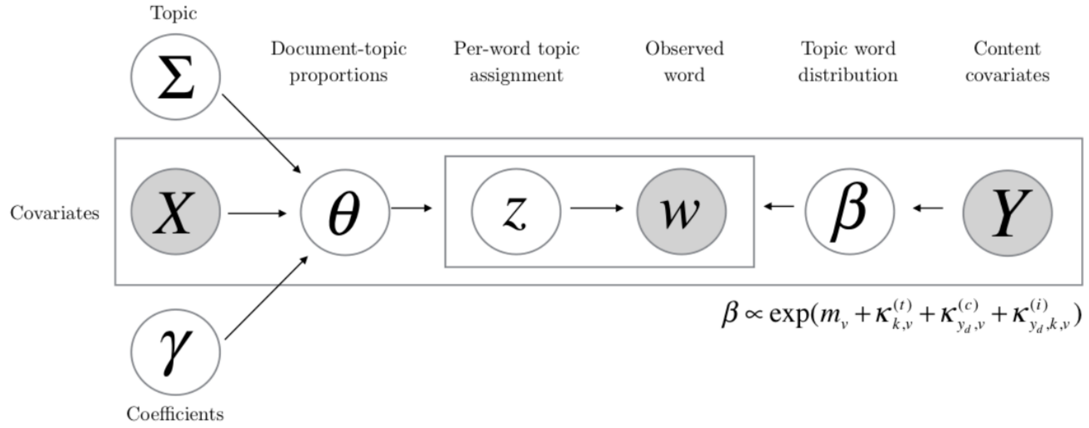
$$z_{d,n} \sim \text{Multinomial}_K(\theta_d) \quad \text{for } n = 1...N_d \tag{4}$$

and the term as:

$$w_{d,n} \sim \text{Multinomial}_K(\mathbf{B}_{d,z_{d,n}}) \quad \text{for } n = 1...N_d \tag{5}$$

As the posterior distribution of this model is intractable, and the logistic Normal prior for $\theta$ is non-conjugate with the multinomial likelihood, Roberts, Stewart, and Airoldi (2016) propose a partially collapsed variational Expectation-maximization algorithm to estimate the model.

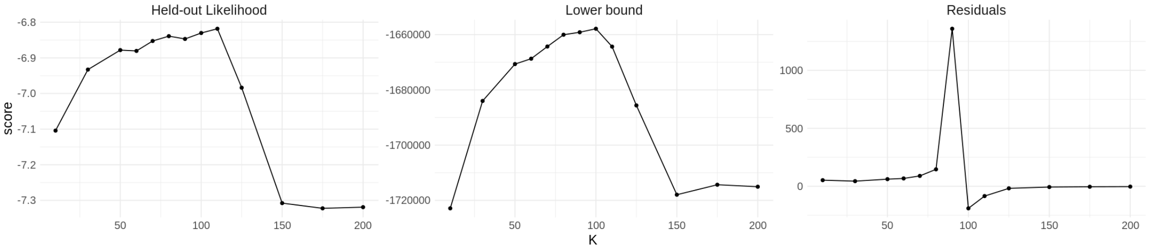
Figure 4: Graphical model of the *STM*


Topic
$\Sigma$
Document-topic proportions
Per-word topic assignment
Observed word
Topic word distribution
Content covariates
Covariates
$X$
$\theta$
$z$
$w$
$\beta$
$Y$
$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$
$\gamma$
Coefficients

## 4.2 Estimating and Interpreting the Model

The number of topics is a major hyperparameter in the model. As for most unsupervised learning methods, there is no unique criterion for selecting the best value of the parameter (Grimmer and Stewart 2013). We start with a data-driven approach to the process by estimating the model for a wide range of topics and comparing their residuals, lower-bound, and held-out likelihood. We use the spectral initialization recommended by Roberts, Stewart, and Tingley (2019) to fit these models. The results can be seen in Figure 5. Using these criteria, we can see that both the held-out likelihood and the lower-bound are maximized near 100 topics. Simultaneously, the residuals have a sharp decline at this point. This behavior of the different scores can be an indication of an adequate number of topics.

Figure 5: Assessing the optimal number of topics



We fitted a model with 100 topics and labeled the topics by exploring its most relevant words according to four different types of word weighting[2]. Additionally, we reviewed the documents with the highest probability for each topic. Despite the large number of topics in the model and the sparsity of our data, we were able to label 58 topics with this procedure. The remaining topics did not display enough cohesion for them to be interpretable and, thus, were excluded from the analysis. A first analysis of these topics, reveals that a majority of them (64%) are not related to substantial issues (e.g., welfare, health, education) but instead serve a merely informational function (e.g., advertising events, inviting people to vote). This result aligns with what has been found elsewhere in the literature (Golbeck et al. 2010; Graham et al. 2013). Furthermore, the model that we estimated allows for the probability of words within a topic to vary across countries. Consequently, for any single topic, we can access the probability distribution over the vocabulary for each country. This helps understand potential differences in the meaning and interpretation of a topic across countries. To investigate this, we calculated the cosine similarity between these probability distributions for each pair of countries (France vs. Germany, UK vs. France and UK vs. Germany). We repeated this process for all our labeled topics. The results are shown in Figure 6. We can observe that most of the topics that exhibit a high degree of similarity — retweeting, take, thanks, support — are also not substantial. On the other hand, topics that can be associated with substantial issues — Brexit, welfare, childcare — appear to have a lower degree of similarity.

We further explored this finding by analyzing two specific topics with low cosine similarity. Figure 7 shows the most characteristic (Roberts, Stewart, and Tingley 2019) words for the *welfare* and *political leaders* topics. While French politicians emphasize words related to “unemployment”, German politicians focus on ”pensions” and their UK counterparts on “taxes” and “care”. This difference in word usage reveals an important heterogeneity on the conception of *welfare* across countries. Similarly, within the topic of *political leaders*, German politicians focus on Angela Merkel and her role as “chancellor” while British politicians highlight Jeremy Corbyn. This is an interesting association given that Angela Merkel was the incumbent chancellor, while Jeremy Corbyn was only one of the candidates.

[2] Namely, highest probability, FREX, Lift, and Score

Figure 6: Cosine distance

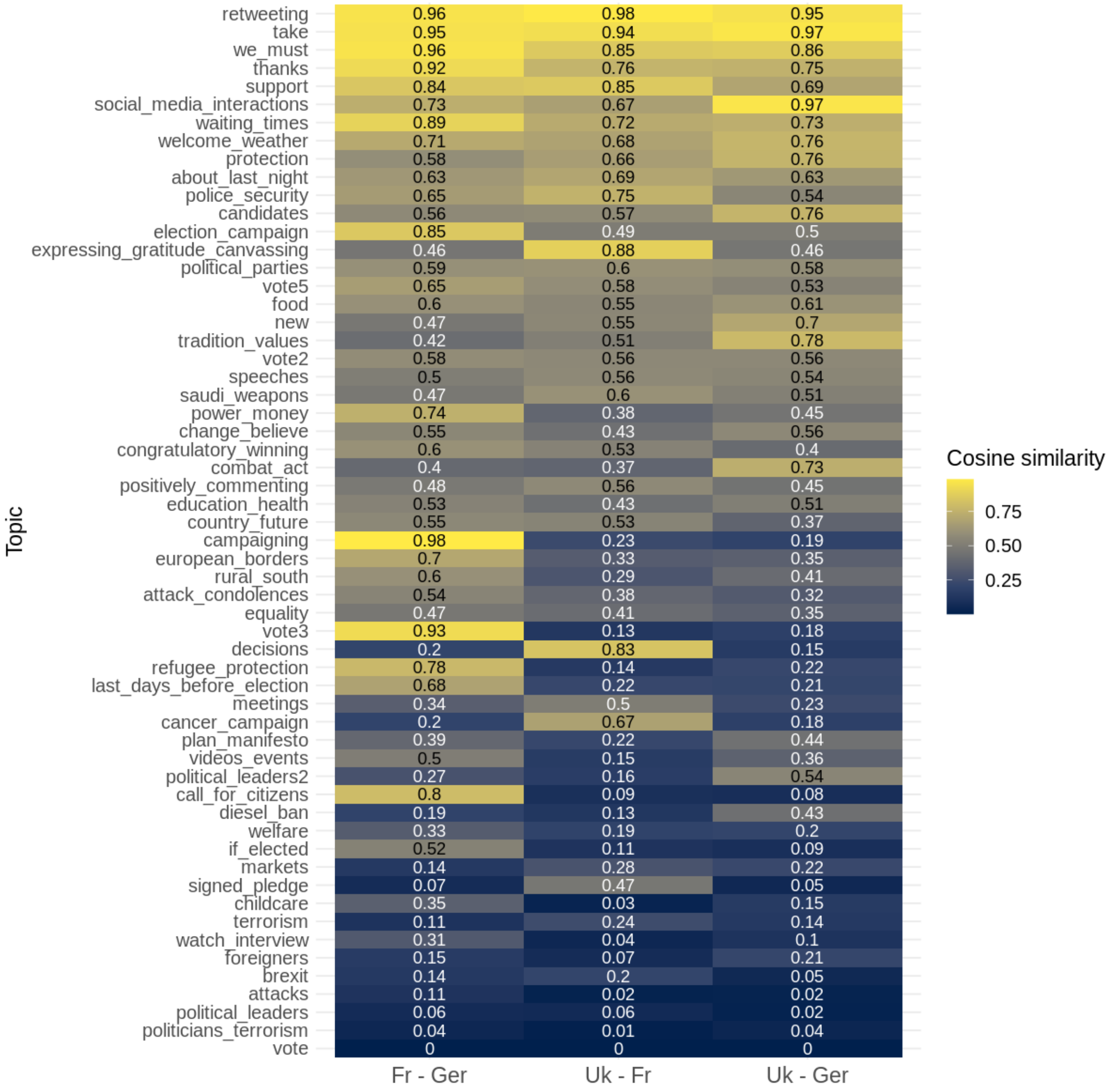
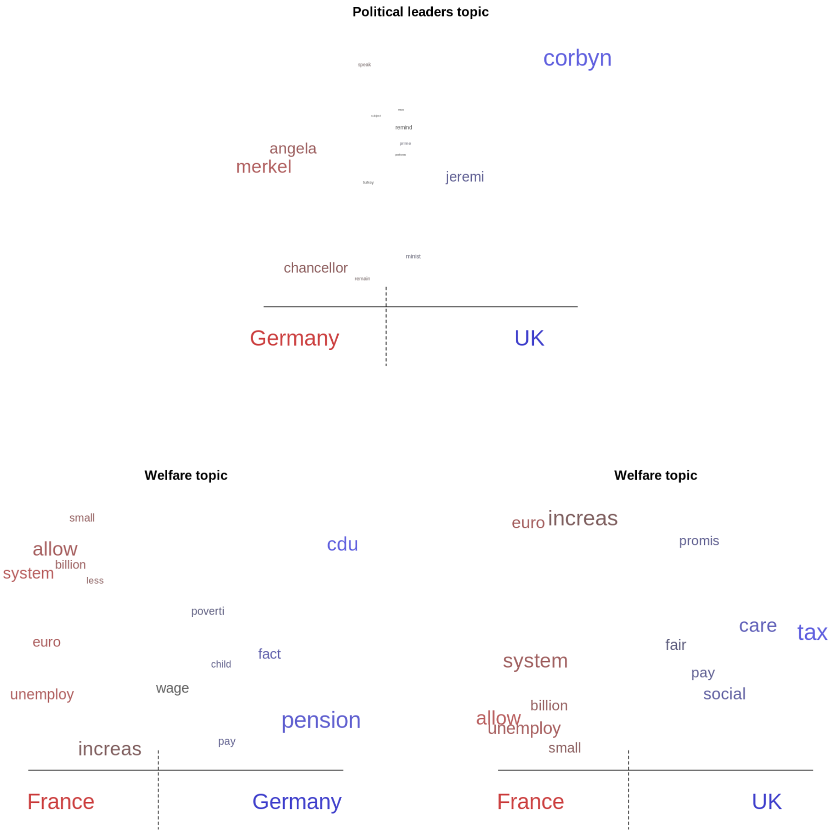


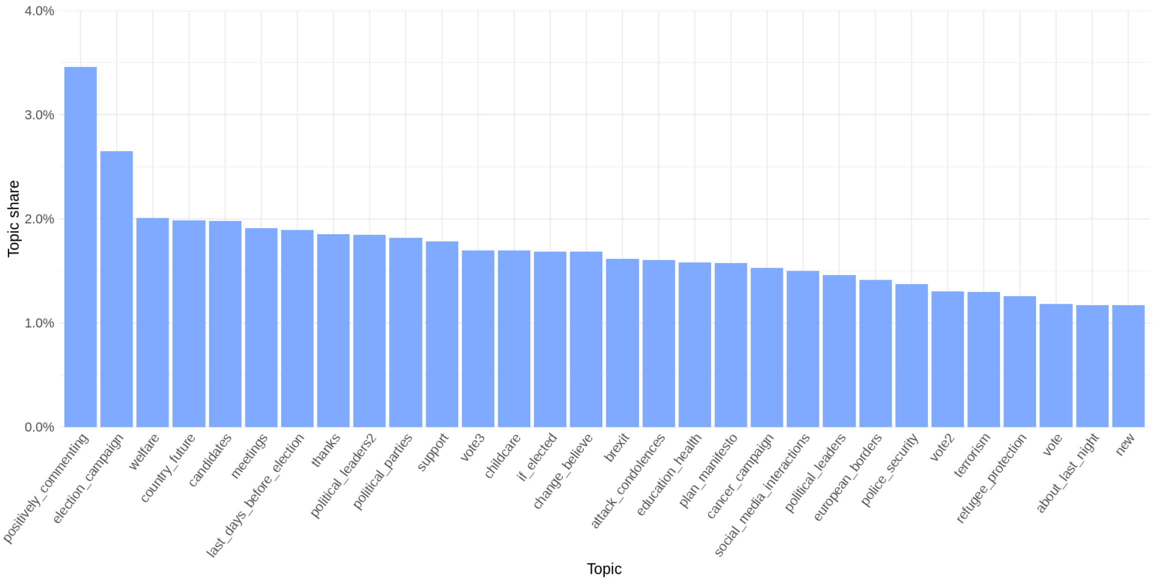
Figure 7: Words with highest probability for topics *welfare* and *political leaders*



After analyzing the content topics used by politicians and their different meanings across countries, we investigated the usage of these topics. Figure 8 shows the topic prevalence of the 30 most used topics in our corpus. In line with our previous findings, the majority of the most used topics do not seem to be referring to substantial issues. Instead, they seem to be capturing different types of campaigning messaging by politicians: expressing gratefulness after rallies, encouraging people to vote, and inviting people to events (both virtual and physical). As mentioned earlier, the STM provides the possibility of understanding how tweet-level covariates affect the prevalence of each topic. Figure 9 shows the estimated topic prevalence by country. Almost all topics exhibit significant differences across countries. In most cases, this difference is mainly driven by the relatively high usage of the topic in one country.

Focusing on the second tweet-level covariate of interest for our analysis, it is possible to understand how the prevalence within each topic changes across different political families. Figure 10 illustrates this. Compared to the variation observed when analyzing the country covariate (figure 9), political families produce less variation within a topic. Nonetheless, most of the topics seem to be mainly used by one specific family, with all the others exhibiting a similar and smaller usage.

Figure 8: Topic prevalence for 30 most used labeled topics



Country differences could partially drive this variation. The topic of *Brexit*, for example, shows a significant difference between the radical right parties and the rest of the parties. However, given the topic's content, it is easy to see that this high usage is mainly due to the British radical right political party (UKIP). These results suggest that although the political family of a candidate influences how he uses topics, the main source of variation seems to be the country. Politician's Twitter content appears to be differentiated more on national than ideological lines.

Figure 9: Topic prevalence for all topics by country

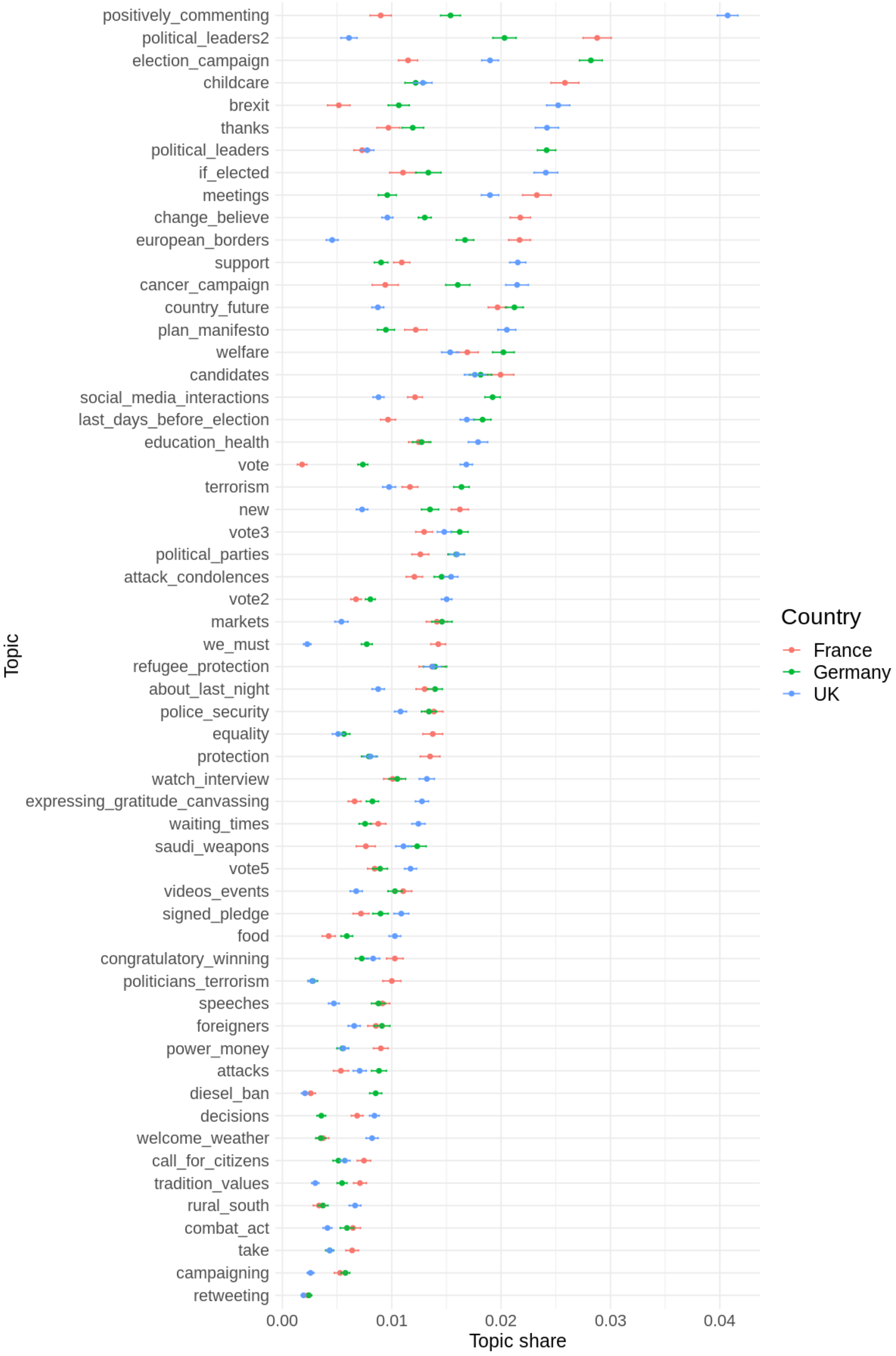
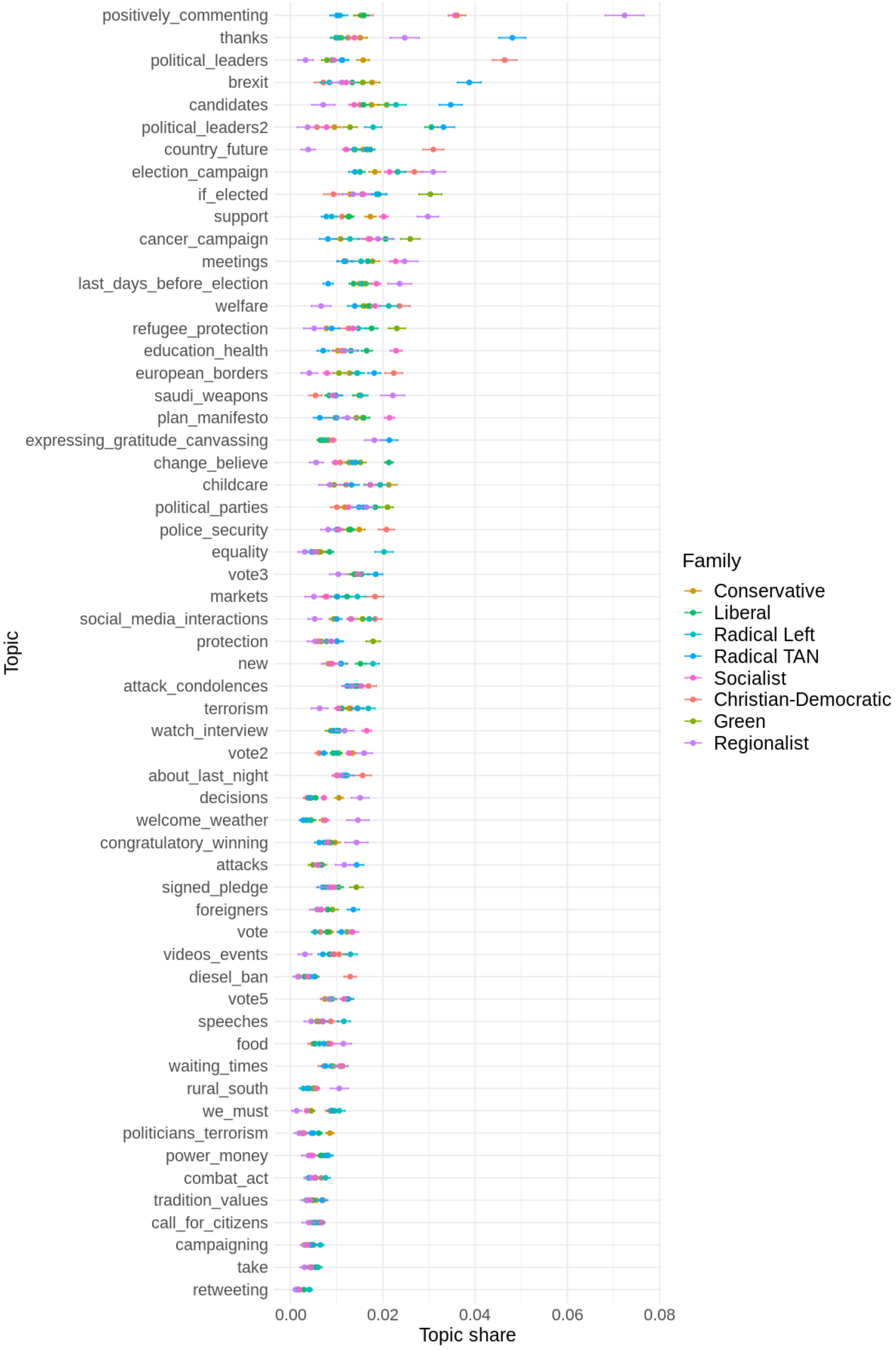


Figure 10: Topic prevalence for all topics by party political family



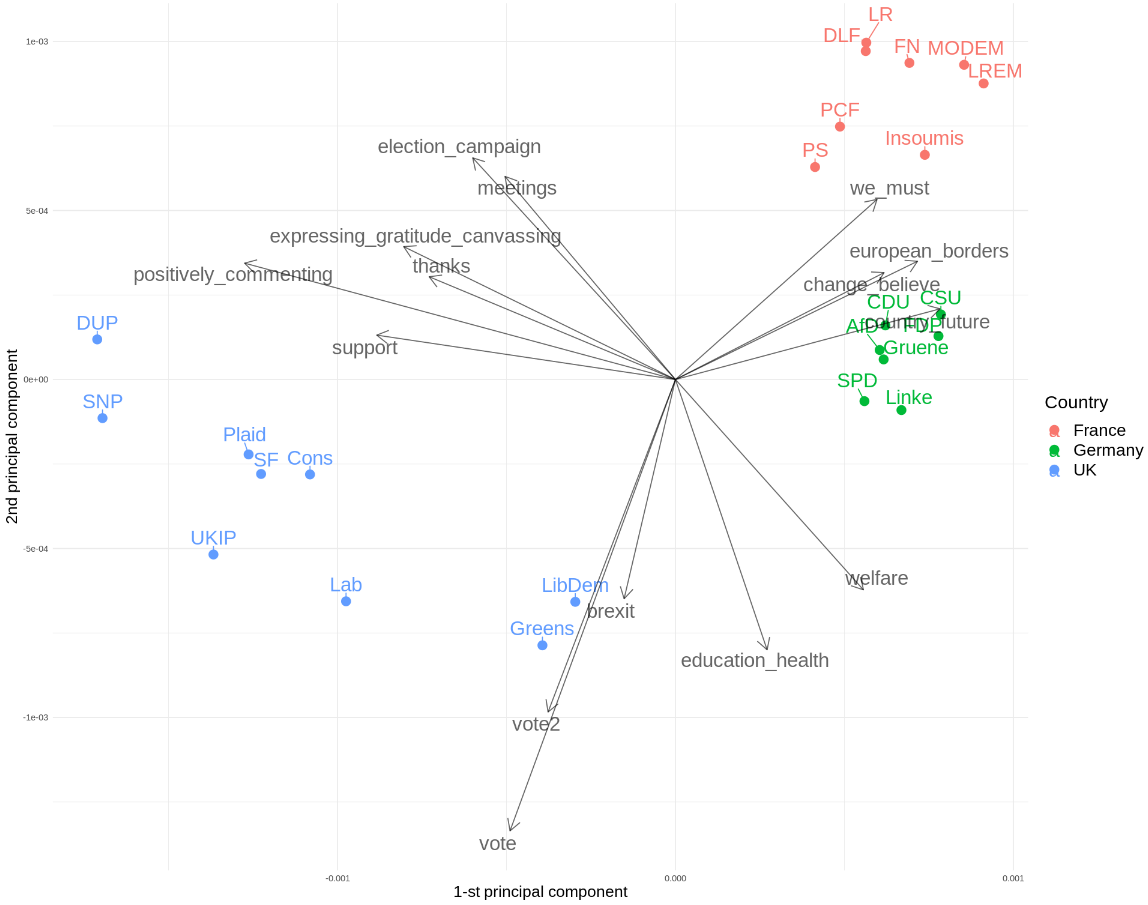
## 4.3 Principal Component Analysis (PCA)

Intending to explore further the main drivers of variation on the usage of topics, we applied the Principal Component Analysis (PCA) on the final results of our model, namely the estimated topic proportions for each tweet. PCA allowed us to reduce the dimensionality of our results and construct two-dimensional plots exploring the principal components with relatively higher variance (PCA biplot).

The first observation from the PCA biplot was the grouping of political parties by their respective country. In Figure 11, we can see a visualized PCA application where we aggregate by political party (by taking the mean) and color each party according to its country of origin, whereas the arrows represent topics and topic usage. Although French and German parties are closer to each other than they are to parties from the UK, it is clear that each political party is more aligned with other parties belonging to the same country. This country alignment strengthens the earlier claim that politicians' Twitter content seems to be much more country-specific than ideology-specific. Besides, we can see that the UK displays the most within-country variability starting from DUP on the top left and continuing to the Liberal Democrats on the bottom right. This suggests that the political agenda of UK parties, which include Regionalist parties situated close together on the left, varies considerably throughout the nation.

The topics displayed in Figure 11, which were selected as the 15 with the highest loadings in the first two dimensions, are another relevant aspect of the PCA biplot. We can see that they are grouped into three main sections: (i) Brexit, education, and welfare (ii) the future of the country and sovereignty, and (iii) positive communication about campaigning. The grouping of topics by direction and the grouping of parties by their country indicate that the political agenda priorities per country are quite disparate. What's more, we can see a dividing line of topics with the UK on one side and France and Germany on the other. Specifically, the topic *Brexit* dominates on the UK side, whereas, on the French and German side, we can see the topics *european borders* and *country future*. This suggests that, instead of tweeting explicitly about Brexit, the countries that would remain in the European Union were tweeting about the future of the union and its borders. On the other hand, the UK is less impacted by topics related to the EU and more by the topic of *brexit*. In addition to Brexit and voting, the other side of UK topics is shaped by campaign messages, events, and meetings. Potentially this could be related to specific political traditions and customs that differentiate how politicians conduct electoral campaigns in the UK compared to France and Germany.

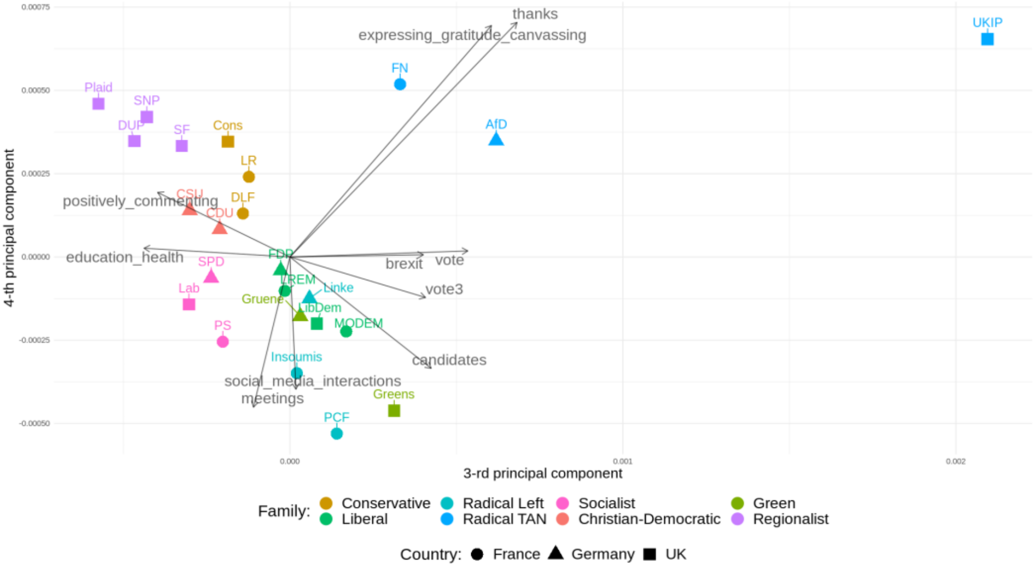
Figure 11: PCA biplot on the 1st and 2nd components



Our interest in variation was not only between countries but also between party families. Particularly, we wanted to capture and observe whether there are cross-country similarities between parties of the same political family. Every country has an average topic distribution representing the Twitter conversation during that time by politicians in that country. However, controlling for that country-level average topic distribution, as shown in Figure 12, the parties start to take a different shape. Figure 12 shows the 3rd and 4th component of the second PCA biplot. At first glance, it is clear that parties of the same political family are closer together than parties of other families. We can see that Plaid, SNP, DUP, and SF, which are all Regionalist parties, are clustered together. In addition, the upper-left quadrant seems to contain more right-wing parties: Conservative, Regionalist, and Christian-Democratic. CDU and CSU, which are colloquially known as the *Union* parties due to their political alliance, are almost perfectly aligned. In fact, center-right parties in the upper-left quadrant are opposite to the center-left parties, which are in the lower-right quadrant.

We can also see a similar pattern with far-right parties in the upper-right quadrant contrasting with the socialist parties on in the lower-left. Lastly, the far-right parties (FN, AfD, and UKIP) are entirely separate from the rest. UKIP stands out by being separated even further within its far-right quadrant, suggesting their topic usage is sui generis.

Figure 12: Demeaned PCA biplot on the 3rd and 4th components



## 4.4 Interactions

From the results above, it is clear that, while we can easily detect common patterns in topic usage in tweets from the same country, commonalities within party families are less clear-cut. To further explore this result, we fit the STM model using interactions between country and party-family as prevalence covariate. By comparing this specification with the baseline one, we can understand whether topic usage is driven by the sum of country and party family effects or whether this is a more nuanced process. It is worth noting that the interpretation of the topics in the two models is the same, with very few exceptions.

As a simple test for differences between the baseline additive model and the interactive specification, for each topic we:

- Regress the share of the topic in each tweet on country and family dummies
- Regress the residuals of this regression on dummies for the interaction between country and family

The presence of statistically significant coefficients associated with the interactions dummies would suggest that the addictive specification does not fully explain topic usage. Results from this test are summarized in Table 1.

Table 1: Significant coefficients in the residual regression

| | sign. 1% | sign 5% |
|---|---|---|
| **Total** | 0.54 | 0.61 |
| **By country** | | |
| France | 0.62 | 0.69 |
| Germany | 0.47 | 0.54 |
| Uk | 0.55 | 0.61 |
| **By family** | | |
| Conservative | 0.64 | 0.70 |
| Green | 0.59 | 0.68 |
| Liberal | 0.63 | 0.70 |
| Radical left | 0.49 | 0.60 |
| Radical tan | 0.68 | 0.74 |
| Socialist | 0.61 | 0.68 |

From this test, we see that more than half of the coefficients are significant at the 1% level (54%), and 61% at the 5% level. Separating the results by country, the interactive model seems to add more to France (69%), followed by the UK (61%), and Germany (54%). Differences are also present across party families (excluding Christian democratic and Regionalist parties because of their presence in only one country), with radical right parties with 74% significant coefficients, radical left with 60% and all other families in the 68%-70% range.

Next, we focused on the change in explained variance between the two specifications. Table 2 shows, for each topic, the explained variance of the additive regression (Column $R^2$ *Add*), the $R^2$ of the regression of the residuals of the first regression on the interaction dummies, and the ratio between the two, which measures the relative improvement in explained variance as a result of including interactive terms in the regression. From the table, we can notice that only a few topics, *thanks* and *brexit* in particular, have a noticeable change in explained variance. In contrast, for the majority of topics, the change is marginal; 72% of the topics that see an increase in explained variance do so by less than 2%. Furthermore, all the topics that benefit from the inclusion of interactions are predominantly used by English politicians, thus reinforcing our findings on the more diverse topic usage in England compared to the other countries considered.

Overall, these results indicate that the interactive model allows for a more nuanced analysis of British politicians' twitter usage because it can capture the specificity of political parties as country-specific representations of a political family. However this is not the case for France and Germany, in light of the negligible increases in explained variance from the interactive model for topics specific to these countries .

Table 2: Explained variance by topic

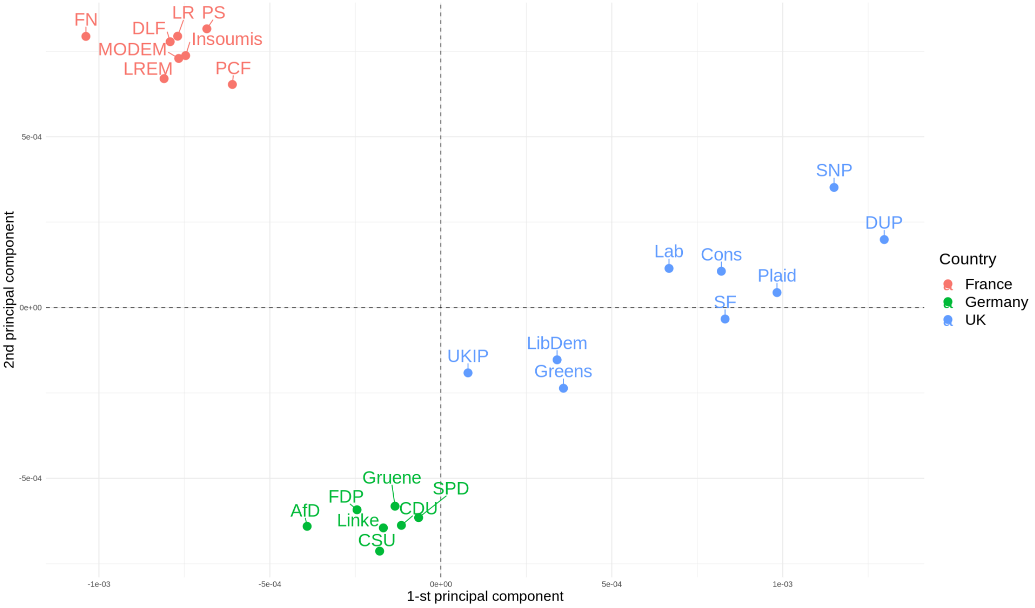
| Topic | $R^2$ Add | $R^2$ Int | Percentage Change |
|---|---|---|---|
| thanks | 0.268 | 0.146 | 0.545 |
| brexit | 0.214 | 0.088 | 0.411 |
| equality | 0.181 | 0.031 | 0.170 |
| expressing_gratitude_canvassing | 0.367 | 0.059 | 0.161 |
| candidates | 0.209 | 0.031 | 0.147 |
| political_leaders2 | 0.336 | 0.049 | 0.145 |
| attacks | 0.112 | 0.010 | 0.092 |
| positively_commenting | 0.502 | 0.037 | 0.074 |
| if_elected | 0.142 | 0.010 | 0.071 |
| refugee_protection | 0.110 | 0.006 | 0.056 |
| cancer_campaign | 0.105 | 0.006 | 0.055 |
| foreigners | 0.122 | 0.004 | 0.033 |
| food | 0.113 | 0.003 | 0.030 |
| signed_pledge | 0.087 | 0.002 | 0.026 |
| speeches | 0.096 | 0.002 | 0.025 |
| welcome_weather | 0.121 | 0.002 | 0.020 |
| vote3 | 0.359 | 0.007 | 0.020 |
| we_must | 0.304 | 0.006 | 0.019 |
| protection | 0.202 | 0.004 | 0.019 |
| political_parties | 0.419 | 0.008 | 0.019 |
| saudi_weapons | 0.147 | 0.003 | 0.017 |
| power_money | 0.107 | 0.002 | 0.016 |
| support | 0.458 | 0.007 | 0.016 |
| politicians_terrorism | 0.101 | 0.002 | 0.016 |
| meetings | 0.263 | 0.004 | 0.016 |
| political_leaders | 0.433 | 0.007 | 0.016 |
| terrorism | 0.327 | 0.005 | 0.015 |
| retweeting | 0.075 | 0.001 | 0.014 |
| waiting_times | 0.187 | 0.003 | 0.013 |
| plan_manifesto | 0.187 | 0.002 | 0.013 |
| european_borders | 0.334 | 0.004 | 0.012 |
| vote2 | 0.585 | 0.007 | 0.012 |
| vote5 | 0.244 | 0.003 | 0.012 |
| watch_interview | 0.159 | 0.002 | 0.012 |
| call_for_citizens | 0.124 | 0.001 | 0.011 |
| welfare | 0.285 | 0.003 | 0.010 |
| change_believe | 0.477 | 0.005 | 0.010 |
| tradition_values | 0.078 | 0.001 | 0.010 |
| attack_condolences | 0.325 | 0.003 | 0.010 |
| last_days_before_election | 0.456 | 0.004 | 0.010 |
| decisions | 0.360 | 0.003 | 0.010 |
| election_campaign | 0.475 | 0.005 | 0.009 |
| childcare | 0.192 | 0.002 | 0.009 |
| combat_act | 0.076 | 0.001 | 0.008 |
| education_health | 0.260 | 0.002 | 0.008 |
| videos_events | 0.169 | 0.001 | 0.007 |
| rural_south | 0.072 | 0.000 | 0.007 |
| markets | 0.163 | 0.001 | 0.006 |
| vote | 0.569 | 0.003 | 0.006 |
| country_future | 0.512 | 0.003 | 0.006 |
| new | 0.281 | 0.001 | 0.005 |
| congratulatory_winning | 0.165 | 0.001 | 0.005 |
| take | 0.225 | 0.001 | 0.004 |
| social_media_interactions | 0.429 | 0.002 | 0.004 |
| diesel_ban | 0.104 | 0.000 | 0.003 |
| campaigning | 0.195 | 0.001 | 0.003 |
| police_security | 0.340 | 0.001 | 0.002 |
| about_last_night | 0.316 | 0.001 | 0.002 |

# 5 Robustness Checks

## 5.1 German Translation

One major concern regarding the results obtained until the moment is that they could be largely driven by our translation strategy. In fact, the higher variation exhibited in the PCA biplots by UK political parties compared to their German and French counterparts, raises the concern of a potential loss in variation in the translation process; given that the tweets from the UK are the only ones that remain in their original language. In order to test this hypothesis, we translated all tweets from British and French candidates to German. Using these tweets we reproduce the initial PCA biplot. Figure 13 shows the result. The overall dynamics seem to remain consistent; political parties seem to be mainly clustered by country and UK parties still exhibit a larger variation. These results give us some confidence that, indeed, we are capturing some signal in the data that is not driven by a preprocessing decision.

Figure 13: PCA biplot on the 1st and 2nd components



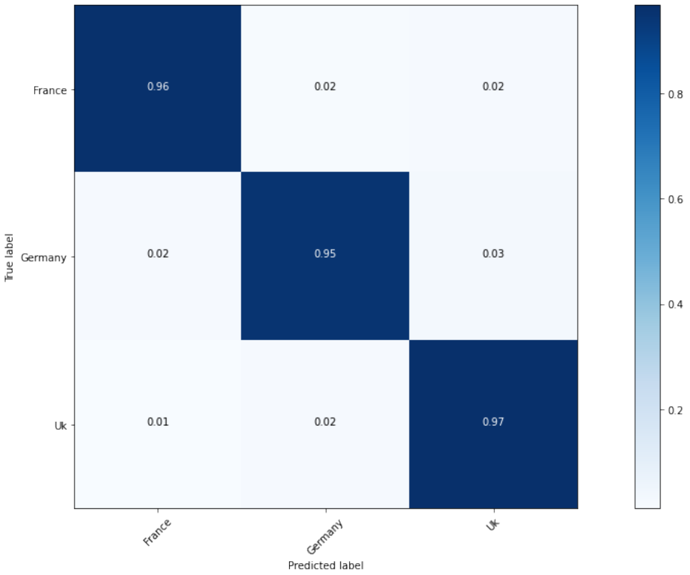
## 5.2 Word Embeddings

As mentioned in our initial literature review, recent developments in the field of natural language processing have led to the appearance of continuous vector representation of words in multiple languages that are aligned in the same space, namely multilingual word embeddings. Using these embeddings allowed us to bypass the translation process and, instead, work directly with the continuous vector representation of each word. Concretely, we used Facebook's pre-trained multilingual word embeddings library MUSE [3] Conneau et al. (2018), which contains hundreds of languages, 30 of which projected on the same space. English, German and French are included among these.

After obtaining all word embeddings for our tweets, we had to decide how to use this information to build a continuous vector representation of each tweet. Adi et al. (2017) show that averaging across the word embeddings that constitute a tweet, is a very effective method for a range of prediction tasks. We will use this approach.

Once we had a continuous vector representation of each tweet, we were able to use these vectors to predict the covariates of interest. If we only had the tweets, could we know the country or political family of the politician who tweeted? We believe this procedure can shed some additional light on the sources of variation in politicians tweets. In order to do this we estimated a multinomial logistic regression and investigated the accuracy achieved when predicting the country of a tweet and its political family. Both a feed-forward neural network architecture and a recurrent neural network architecture were implemented but their results did not differ qualitatively from the ones obtained with the multinomial logistic regression. We chose the multinomial logistic regression for its simplicity.

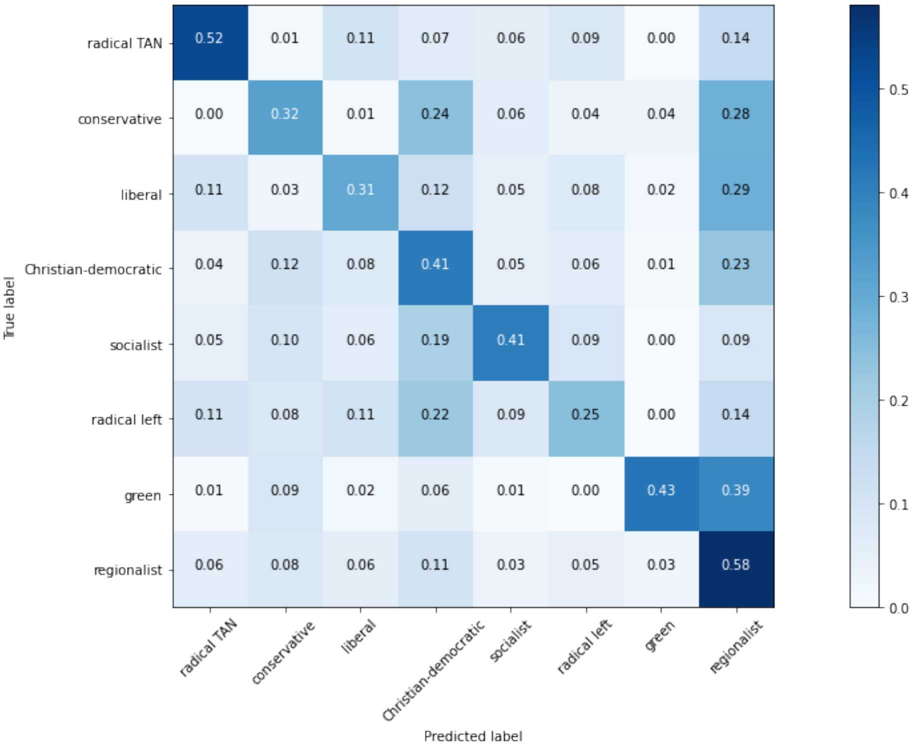
Figure 14: Confusion matrix for country prediction



[3] Available online: https://github.com/facebookresearch/MUSE

Figures 14 and 15 depict the confusion matrix from these two prediction tasks. The classifier can almost perfectly predict (96% accuracy) the country of a tweet in a test set and it performs similarly across the three countries. This finding aligns with the results from the previous section: politicians' Twitter content contains a strong national signal. This allowed us to almost perfectly predict their country with a simple classifier. The panorama, however, is different when predicting the political family of each tweet. Using the same multinomial logistic regression model, we were able to achieve a 42% accuracy on a test set of tweets. Although considerably better than a random guess, this result does not match with the almost perfect prediction achieved for the country of a tweet; emphasizing the idea that what a politician talks about in Twitter is more driven by their country than by their political affiliation.

Figure 15: Confusion matrix for political family prediction



| | radical TAN | conservative | liberal | Christian-democratic | socialist | radical left | green | regionalist |
|---|---|---|---|---|---|---|---|---|
| radical TAN | 0.52 | 0.01 | 0.11 | 0.07 | 0.06 | 0.09 | 0.00 | 0.14 |
| conservative | 0.00 | 0.32 | 0.01 | 0.24 | 0.06 | 0.04 | 0.04 | 0.28 |
| liberal | 0.11 | 0.03 | 0.31 | 0.12 | 0.05 | 0.08 | 0.02 | 0.29 |
| Christian-democratic | 0.04 | 0.12 | 0.08 | 0.41 | 0.05 | 0.06 | 0.01 | 0.23 |
| socialist | 0.05 | 0.10 | 0.06 | 0.19 | 0.41 | 0.09 | 0.00 | 0.09 |
| radical left | 0.11 | 0.08 | 0.11 | 0.22 | 0.09 | 0.25 | 0.00 | 0.14 |
| green | 0.01 | 0.09 | 0.02 | 0.06 | 0.01 | 0.00 | 0.43 | 0.39 |
| regionalist | 0.06 | 0.08 | 0.06 | 0.11 | 0.03 | 0.05 | 0.03 | 0.58 |

Interestingly, the confusion matrix for the political family prediction (Figure 15) reveals heterogeneity in the quality of these predictions. Radical right and regionalist parties show the highest accuracies, 52% and 58% respectively, while radical left parties show the lowest 25%. Potentially, this observed difference can be due to the varying degree of country homogeneity of the parties that belong to these political families. If politicians from certain political families use the identified topics similarly across countries, then predicting the political family of these politicians would be easier than doing so for politicians that belong to families that exhibit large differences across countries. This seems to be the case for radical right parties; they exhibit a higher homogeneity across countries than radical left parties and, thus, knowing only the political family we can achieve better results.

# 6 Conclusion

Exploring politicians' Twitter usage during election periods can help identify main themes in their communication, how their speech differs from other countries or parties, and what contributes to this variance. The 2017 national elections of France, Germany, and the UK provided a conducive context to analyze political speech on Twitter across three European countries at a time when political polarization was increasing across the continent. Although there are several approaches to analyzing text, the Structural Topic Model (STM) chosen for this paper provided an advantageous angle to the analysis, as it allows the inclusion of supplemental political information (i.e., country and party family) as covariates to the model.

Overall, the majority of the 100 topics estimated by the STM displayed enough cohesion to be appropriately labeled. Focusing on this set of labeled topics, we were able to see that most topics were not related to substantial issues (e.g., welfare, health, education) but instead served a basic informational function (e.g., advertising events, inviting people to vote). Interestingly, the word usage for substantial topics displayed larger differences across countries than for informational topics. The specific examples of the *welfare* and *political leaders* topics were shown. Furthermore, thanks to the inclusion of tweet-level metadata in the model, we were able to see that topic usage varied significantly both at the country and the political family level. However, topic usage seemed to display a higher variation between countries than between political families, suggesting that politicians' Twitter content is more country-specific than ideology-specific. This result was reinforced by an analysis of the principal components of the estimated topic proportions. Additionally, two different robustness checks were conducted to assess the potential impact of the translation strategy on the results. The first translated all tweets to German and replicated the principal component analysis; the results remained unchanged. The second deviated from the bag-of-words representation of text in order to make use of continuous vector representation of words (multilingual word embeddings).

A further exploration of how this continuous vector representation of words can be effectively used to undercover the thematic structure in a corpus will positively complement our work. A recent literature on this topic is currently under construction (Das et al. 2015; Dieng et al. 2019; W. Hu and Tsujii 2016; Moody 2016; Wang et al. 2015). Additionally, exploring how the uncovered differences in politician's Twitter usage impact the reaction of Twitter users (i.e., retweets, likes, and responses) and, potentially, how this translates into popularity and electoral results could be a particularly exciting direction for future research.

## References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-Grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *International Conference on Learning Representations (ICLR).*

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society*, *16*(3), 291–314. https://doi.org/10.1080/1369118X.2012.756050

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models, In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania, ACM Press. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Boyd-Graber, J., & Blei, D. M. (2012). Multilingual Topic Models for Unaligned Text. *arXiv preprint arXiv:1205.2657.*

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, *11*(2-3), 143–296. https://doi.org/http://dx.doi.org/10.1561/1500000030

Clement, J. (2019). Twitter: Number of active users 2010-2019. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Club, D. (2017). Candidates and results data. Democracy Club. https://candidates.democracyclub.org.uk/api/docs/

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word Translation Without Parallel Data [arXiv: 1710.04087]. *arXiv:1710.04087 [cs]*. Retrieved April 27, 2020, from http://arxiv.org/abs/1710.04087

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for Topic Models with Word Embeddings, In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1077

De Smet, W., & Moens, M.-F. (2009). Cross-language linking of news stories on the web using interlingual topic modelling, In *Proceeding of the 2nd ACM workshop on Social web search and mining - SWSM '09*, Hong Kong, China, ACM Press. https://doi.org/10.1145/1651437.1651447

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces [arXiv: 1907.04907]. *arXiv:1907.04907 [cs, stat]*. Retrieved June 17, 2020, from http://arxiv.org/abs/1907.04907

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, n/a–n/a. https://doi.org/10.1002/asi.21344

Graham, T., Broersma, M., Hazelhoff, K., & van 't Haar, G. (2013). Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign. *Information, Communication & Society*, *16*(5), 692–716. https://doi.org/10.1080/1369118X.2013.785581

Greene, D., & Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, *25*(1), 77–94. https://doi.org/10.1017/pan.2016.7

Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, *18*(1), 1–35. https://doi.org/10.1093/pan/mpp034

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine learning*, *42*, 177–196. https://doi.org/https://doi.org/10.1023/A:1007617005950

Hu, W., & Tsujii, J. (2016). A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2062

Kruikemeier, S. (2014). How political candidates use Twitter and the impact on votes. *Computers in Human Behavior*, *34*, 131–139. https://doi.org/10.1016/j.chb.2014.01.025

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Majó-Vázquez, S., Nurse, J. R. C., Simon, F. M., & Nielsen, R. K. (2017). Digital-Born and Legacy News Media on Twitter during the German Federal Election. *Reuters Institute for the Study of Journalism.* https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/20171030_RISJ_German_Factsheet_.pdf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space [arXiv: 1301.3781]. *arXiv:1301.3781 [cs]*. Retrieved June 4, 2020, from http://arxiv.org/abs/1301.3781

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, Singapore, Association for Computational Linguistics. https://doi.org/10.3115/1699571.1699627

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec [arXiv: 1605.02019]. *arXiv:1605.02019 [cs]*. Retrieved June 17, 2020, from http://arxiv.org/abs/1605.02019

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2009). Mining Multilingual Topics from Wikipedia.

NosDéputés. (2019). Parliamentary data in opendata. NosDéputés. https://github.com/regardscitoyens/nosdeputes.fr/blob/master/doc/opendata.md

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Polk, J., Rovny, J., Bakker, R., Edwards, E., Hooghe, L., Jolly, S., Koedam, J., Kostelka, F., Marks, G., Schumacher, G. Et al. (2017). Explaining the salience of anti-elitism and reducing political corruption for political parties in europe with the 2014 chapel hill expert survey data. *Research & Politics*, *4*(1), 2053168016686915.

Proksch, S.-O., & Slapin, J. B. (2010). Position Taking in European Parliament Speeches. *British Journal of Political Science*, *40*(3), 587–611. https://doi.org/10.1017/S0007123409990299

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 306–315.

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., & Staab, S. (2018). *Systematically Monitoring Social Media: The Case of the German Federal Election 2017* (preprint). SocArXiv. https://doi.org/10.31235/osf.io/5zpm9

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, *35*(1), 50–74. https://doi.org/10.1080/10584609.2017.1334728

Vulić, I., De Smet, W., Tang, J., & Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, *51*(1), 111–147. https://doi.org/10.1016/j.ipm.2014.08.003

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words, In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania, ACM Press. https://doi.org/10.1145/1143844.1143967

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., & Hao, H. (2015). Semantic Clustering and Convolutional Neural Network for Short Text Categorization, In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2058