Findings from the Hackathon on Understanding Euroscepticism Through the Lens of Textual Data

Federico Nanni[a,*], Goran Glavaš[a], Simone Paolo Ponzetto[a], Sara Tonelli[b], Nicolò Conti[c], Ahmet Aker[d,e], Alessio Palmero Aprosio[b], Arnim Bleier[f], Benedetta Carlotti[g], Theresa Gessler[h], Tim Henrichsen[i], Dirk Hovy[j], Christian Kahmann[k], Mladen Karan[l], Akitaka Matsuo[m], Stefano Menini[b], Dong Nguyen[n,o], Andreas Niekler[k], Lisa Posch[f], Federico Vegetti[p], Zeerak Waseem[d], Tanya Whyte[q], Nikoleta Yordanova[a]

[a]University of Mannheim, [b]Fondazione Bruno Kessler, [c]Unitelma Sapienza, [d]University of Sheffield,
[e]University Duisburg-Essen, [f]GESIS Leibniz Institute for the Social Sciences, [g]Scuola Normale Superiore,
[h]European University Institute, [i]Scuola Superiore Sant'Anna, [j]Bocconi University, [k]Liepzig University,
[l]University of Zagreb, [m]London School of Economics, [n]Alan Turing Institute,
[o]University of Edinburgh, [p]Central European University, [q]University of Toronto

Abstract

We present an overview and the results of a shared-task hackathon that took place as part of a research seminar bringing together a variety of experts and young researchers from the fields of political science, natural language processing and computational social science. The task looked at ways to develop novel methods for political text scaling to better quantify political party positions on European integration and Euroscepticism from the transcript of speeches of three legislations of the European Parliament.

Keywords: Political Text Scaling, Euroscepticism, Text as Data, Hackathon, Computational Social Science

1. Introduction

The unprecedented availability of large amounts of records of digital materials presents tremendous opportunities for political scientists, sociologists, historians, as well as any researchers focused on studying the present times (Grimmer and Stewart, 2013; Graham et al., 2016). However, traditionally trained social scientists and humanities scholars often lack the methodological expertise to examine, manage, and extract information from these large and noisy datasets of primary sources. On the other side of the spectrum, data scientists and natural language processing (NLP) researchers, who work with these resources on a daily basis, usually do not have the background knowledge to identify and address relevant research questions adopting these materials, particularly when it comes to extremely complex phenomena like the impact of the financial crisis in different socio-economical strata, the perception of the migrant crisis across countries and, especially for this paper, the growth of skepticism towards the European Union (EU).

The Goal. For these reasons, we decided to organize a three-day interactive seminar (a 'hackathon') that took place during the first half of December 2017 at Villa Vigoni,[1] with the aim of bringing together PhD and Postdoc researchers from different disciplines and guide them to work together on a series of new datasets. In this context, researchers had the possibility of sharing methodologies, discussing research questions, and cooperating in small interdisciplinary groups.

The focus of the hackathon was to develop new text-scaling algorithms for better understanding how Eurosceptic opinions emerge in institutional debates.

Outline. In the remainder of the paper, we first offer an overview of quantitative approaches for measuring Euroscepticism. Next, we present related work on the adoption of NLP approaches in political science research (in particular for text scaling) and on the benefits of hackathons for enhancing interdisciplinary research. We then describe the datasets adopted, the gold standard, and the addressed task. We share all resources used during the hackathon with the research community to support further work on the topic. The different approaches developed during the hackathons are briefly described before presenting the quantitative evaluation. Finally, the paper is wrapped up with a conclusion.

2. Background: Measuring Euroscepticism

During the last decade, a widespread opposition toward the EU strongly consolidated in several European countries. This phenomenon brought to the rise in consensus of parties critical to the EU from both the left and the right side of the political spectrum (Halikiopoulou et al., 2012). An example of such progressions is given by the results of the last European Parliament (EP) elections, where so-called Eurosceptic parties won 74 seats at the expenses of their mainstream counterparts when compared to previous EP elections in 2009.

Euroscepticism. This complex socio-political phenomenon has generally been labeled using the media-driven concept of Euroscepticism.[2] This term and its academic

*Regarding the order of the authors: first the five organizers, then the participants (alphabetical order).

[1] `http://villavigoni.it`

[2] Euroscepticism was first used by *The Times* in 1985 (hyphenated version of the term). It then spread to the political and academic environment becoming a real subfield of European studies 'a cottage industry of "Euroscepticism studies"' (Mudde, 2012).

study evolved hand in hand with the development of the EU itself. Initially, the study of Euroscepticism was fragmented and limited to countries where the phenomenon was present (Usherwood and Startin, 2013), while, since the 90s, alongside some events crucial to the evolution of the EU (e.g., the signing of the Single European act, the Maastricht Treaty), Euroscepticism has been extensively studied from two main perspectives: mass Euroscepticism and party-based Euroscepticism. The first one deals with voters' attitudes towards the EU, while the second one focuses on political parties' stances on the EU and European integration. In spite of the growing importance of Euroscepticism, scholars still struggle to provide a uniquely valid definition of the concept (Usherwood, 2016).

Issues with the Definition. Euroscepticism was firstly defined by Taggart as a "contingent and conditional opposition to the EU integration as well as a total and unconditional opposition to it" (Taggart, 1998). Since then, and after the dichotomy distinction between 'hard' and 'soft' Euroscepticism was coined (Taggart and Szczerbiak, 2002), a vivid dialogue between scholars in the field emerged to find the best one-size-fits-all definition of the concept (Flood, 2002; Kopecký and Mudde, 2002; Conti, 2003; Rovny, 2004). In parallel with the absence of a precise definition of Euroscepticism, five main problems connected to this concept need to be further stressed. Firstly, the term itself may lead to conceptual confusion since it is composed by a prefix 'Euro' used as a proxy for the EU, a central component 'sceptic' which refers to the contraposition to the pro-EU "religious orthodoxy" (Cotta, 2016), and the suffix '-ism', which is generally used to identify ideologies, even if Euroscepticism cannot be considered as an ideology perse but as a component of other ideologies (Flood, 2002; Vasilopoulou, 2009). Secondly, the term Euroscepticism is clearly negatively constructed (Crespy and Verschueren, 2009) and can thus be misused in political competitions to disparage political challengers both in an inter-party and in an intra-party perspective (Cotta, 2016). Thirdly, Euroscepticism's negative construction implies the recognition of a positive pro-European side that is in turn not well specified, i.e., it is sometimes difficult to draw clear boundaries between which party is or is not Eurosceptic. For example, is a party asking to reform the EU to be considered as Eurosceptic? If this is the case, how can we classify parties rejecting the EU? Fourthly, Euroscepticism generally identifies the EU and the European integration as a monolithic unit without distinguishing between what the EU is (the complex of EU institutions ruling member states, united under a single European community) and what the EU does (the outputs of the EU decision-making process in various policy fields). Lastly, as mentioned above, criticism towards the EU evolved hand in hand with the EU itself, therefore Euroscepticism has changed diachronically and cross-nationally. All the problems connected to the concept of Euroscepticism have led to more recent studies, arguing that it would be better to talk about Euroscepticisms using the plural form (Usherwood, 2016) or to reconceptualise it using the more neutral concept of 'political opposition' (Carlotti, 2017). Besides the above-mentioned problems, Euroscepticism is still widely used to understand both voters' and parties' positioning to the EU.

Traditional Approaches. Various sources of data have been used to estimate the position of political parties on Euroscepticism. Firstly, public opinion surveys, such as the European Election Study, allow measuring voters' perceptions of party positions via issue scales (Adams et al., 2014). Usually, such surveys are conducted periodically with each survey wave sampling new respondents, thus prohibiting a longitudinal analysis of changes in individual perceptions about party positioning. Second, party manifestos for national (Conti, 2003) and European elections (Schmitt et al., 2007) have served to estimate parties' stated preferences. However, as party manifestos are drafted for the purpose of elections, naturally they only offer a snapshot of parties' preferences every four to five years. Third, voting advice applications, such as EU Profiler for the European Parliament elections, offer data on political parties' self-positioning on various issue scales at election time. These data is even scarcer and do not offer more than a glimpse into parties' EU stances either. Fourth, to capture parties' revealed positions on European integration, scholars have relied on expert surveys, such as Chapel Hill Expert Survey (Polk et al., 2017), and surveys of members of parliament (Whitaker et al., 2017), which are conducted once every couple of years. A fifth common measure of parties' EU positions is based on parliamentary roll call votes (Hix et al., 2007). While roll call votes offer fine variation over time, they have been criticized for suffering from the selection bias (Carrubba et al., 2006; Yordanova and Mühlböck, 2015). Vote choice may also not reveal true preferences because it is constrained by party disciplining and the institutional rules (Hug, 2016), as well as strategic behavior on the party of legislators (Mühlböck and Yordanova, 2017).

While all these different approaches have already offered solid insights into the phenomenon, each of these solutions runs the risk of capturing only a few aspects of the overall perception of Continental society towards the European Union, and especially the reasons behind the growth of a widespread opposition to its politics and role. For this reason, the social science practice of survey research has always tried to move beyond these limitations (De Vreese, 2007). However, every social scientist knows the difficulties that survey research brings, both in terms of the time needed to conduct an extensive study and the accuracy of the final results, especially when it comes to analyzing extremely complex topics.

NLP-based Approaches. This brings us to the newest trend in estimating party stances with the use of textual data, i.e., *political-text scaling* (Grimmer and Stewart, 2013), such as from party speeches, press releases, parliamentary questions, etc (Wilde et al., 2014). The major advantages of this methodology are the abundance of such data and the recent developments of NLP approaches precisely tailored for supporting such applications (Glavaš et al., 2016; Glavaš et al., 2017a; Menini and Tonelli, 2016; Menini et al., 2017; Nanni et al., 2016; Zirn et al., 2016, *inter alia*). More substantively, it allows generating time-varying estimates of parties' EU positions. As with any other data, though, researchers have to carefully consider the generation process behind textual data and its implications for the study at hand. For instance, when it comes to parliamentary speeches, parties may strategically decide whom to allow to speak so as to appear unified to the public (Proksch and Slapin, 2015). Also, different ideological dimensions seem to underlie voting behavior and speeches (Proksch and Slapin, 2010). Understanding party and institutional constraints of giving speeches as well as legislators' motivations to speak, is thus essential in judging what speech can tell us about political preferences.

3. Related Work

In this section we briefly present an overview of previous studies on political text scaling and the advantages of organizing shared tasks and hackathons in order to build new bridges between interdisciplinary communities.

Political Text Scaling. The goal of political scaling is to order political entities, such as political parties and politicians, according to the position they expressed in textual content. The type of orientation could be ideological (i.e., left vs. right) as well as policy-specific (regarding economics or welfare). Documents such as parties' election manifestos or transcripts of speeches are commonly used as the data underpinning this type of analysis (Grimmer and Stewart, 2013). Although the idea of estimating ideological beliefs is not new (Abelson and Carroll, 1965), nevertheless the first models able to estimate these beliefs from texts have only appeared in the last fifteen years (Laver and Garry, 2000; Laver et al., 2003; Slapin and Proksch, 2008; Proksch and Slapin, 2010). The seminal works by Laver and Garry (2000) and Laver et al. (2003) are widely considered the starting points of this field of research. These supervised approaches rely on predefined dictionaries of words or reference documents for establishing the position of unlabeled texts. In order to avoid the manual annotation effort (and the biases that this could add to the model), Slapin and Proksch (2008) proposed Wordfish, an unsupervised scaling model which has become the *de facto* standard method for unsupervised political text scaling. This approach models document positions and contributions of individual words to those positions as latent variables of the Poisson naïve Bayes generative model, i.e., they assume that words are drawn independently from a Poisson distribution. They estimate the positions by maximizing the log-likelihood objective in which word variables interact with document variables.

While this previous methodological research has been conducted by the political science community, in recent years works on political text scaling have also been presented by NLP groups (e.g., Nanni et al. (2016)). Among them, in particular Glavaš et al. (2017b), has proposed a new text scaling approach that leverages semantic representations of text, making it suitable both for mono- and cross-lingual political text scaling. The authors of this paper have shown that the semantically-informed scaling models better predict the party positions than Wordfish in two different political dimensions and that the models exhibit no drop in performance in the cross-lingual setting compared to monolingual one.

Gold Standard for Scaling. Generally, expert surveys are regarded as one of the most popular survey-based approaches for the estimation of parties positioning on several issues and as gold-standard for measuring the quality of text scaling algorithms. The rationale behind them is that experts in the field (e.g., political scientists) evaluate parties positioning on several issues on the basis of their domain knowledge. The resulting parties positioning is given by the aggregation of experts' judgments using measures of central tendency (e.g., the mean). Nonetheless, as various experts in the field suggest, the use of the Chapel Hill expert survey, as every expert survey, shows both advantages and drawbacks; this section briefly overviews them. The first problem connected to expert surveys is that it is not clear 'what' experts actually evaluate (Budge, 2000) since they are generally asked questions with 'minimal instructions' (Gemenis, 2015). In other words, experts are asked to provide judgments without having 'reference points', consequently making such judgments interpersonally and cross-nationally incomparable. Steenbergen and Marks (2007) demonstrate that such inter-expert disagreement correlates with certain parties' characteristics like their size and ideological background. However, according to the proponents of expert surveys, such an inter-experts disagreement may be solved through statistical aggregation, since the errors 'will cancel out' (Steenbergen and Marks, 2007). However, such an error component is not only a function of parties' characteristics, but also of experts' personal characteristics such as their knowledge or ideological background. This last consideration is connected to the second main problem of expert surveys: there is a great variance in the criteria used by experts to make their judgments. According to Curini (2010), "the estimation of parties positioning on the basis of survey data (broadly speaking) is not always consistent since respondents tend to place parties they like closer to where they perceive themselves to be, and to place those parties they dislike farther away then the actual position would warrant, thus producing a bias known as rationalization or projection" (see also Granberg and Brown (1992).[3] Since expert surveys aim is to estimate parties' positioning and not to infer the attributes of the experts' population on the basis of a set of actual respondents, relying on them may affect the validity of the obtained results (Curini, 2010).

Besides these problematic aspects, expert surveys are able to provide information in a common, standardized format across a wide range of countries. They are generally regarded as having weight and legitimacy, since they reflect the judgments of experts who are presumably well informed about the topic. Lastly, expert surveys are easily compared to other forms of analyses like the content analysis of parties manifestos or the observation of legislative behavior (through the use of roll call votes), which are in turn not free from biases either. Despite the potential drawbacks of the Chapel Hill expert survey, we relied on it to have a quick and easy way to position parties along the pro-against

[3] More specifically assimilation effects realize themselves when respondents shorten the distance between themselves and the party they favor while widening the distance between themselves and the parties they do not support.

European integration dimension.

Hackathons in DH and CSS. In the last decade, the NLP community has been involved in the organization of several activities aiming to bridge the gap between the field, the digital humanities and the computational social sciences. From workshops[4] and shared-tasks,[5] all the way through seminars,[6] summer schools,[7] and tutorials,[8] large efforts have been made to present and working together on new datasets, tools, and platforms in order to address relevant research questions, following a "more hack, less yack" attitude (Nowviskie, 2014). Among these efforts, some interdisciplinary hackathons similar to ours have been organized in the recent years: the Archives Unleashed[9] series organized five times since 2016, brought together digital archivists, humanities scholars, and computer scientists interested in the use of web archives (such as the Internet Archive) for studying the recent past. Other similar projects have been focused on biodiversity,[10] the 2016 US Elections,[11] and Tibetan studies (Almogi et al., 2016).

Inspired by these previous projects, in the hackathon we organized at Villa Vigoni, we decided to combine this highly interdisciplinary setting with a shared-task focused on developing new algorithms for text scaling.

4. The Hackathon

At the beginning of December 2017, 18 researchers (mainly PhD students and postdocs) with a background in political science, computational social science, or natural language processing took part in the hackathon. Upon arrival, the participants have been divided by the organizers in five interdisciplinary teams, named after national European football teams that did not manage to qualify to the final stage of 2018 World Cup. Then, the participants received an overview of the hackathon's shared-task, which they had 48 hours to address. The task was to develop new text-scaling algorithms tailored for identifying Eurosceptic opinions in institutional debates. Following, the organizers introduced the datasets and evaluation framework, as presented next.[12]

Parliamentary Text Collection. Given the focus on institutional debates, the organizers first crawled and provided to the participants all individual speeches of all European Parliament representatives, in all languages available (i.e., in the original language of the speech and all manual translations to other languages, if existing) from the official website of the European Parliament.[13] The collected materials cover 4 legislations (5th to 8th) and almost 20 years of European politics (1999-2017), and include a large variety of topics, ranging from the advent of Euro, the enlargement of the Union to the economic and refugee crises, and the growth of Euroscepticism. The raw corpus consists of four subparts (one for each legislation period), with one XML document per representative aggregating all speeches that each one delivered over the course of the legislation period (see Fig. 1). Besides the speeches (content and date for each one), for each representative we also obtained the information on the national party and European party group.

Leg. period	# parties	Min. len	Avg. len
5th (1999–2003)	25	14.5K	127.7K
6th (2004–2008)	30	13.9K	96.4K
7th (2009–2013)	24	54.9K	467.0K

Table 1: Per-legislation term statistics of the European Parliament dataset used in the hackaton.

New Benchmark Dataset. For the purpose of the hackathon, we considered only the speeches made or manually translated into English. We concatenated all speeches of all representatives of the same party into a single party-level document. Following previous works (Proksch and Slapin, 2010; Glavaš et al., 2017b), we selected the parties from the five largest European countries: Germany, France, United Kingdom, Italy, and Spain. Finally, we discarded the parties for which the aggregate texts over the whole legislation period ended up being shorter than 10.000 tokens. We decided to use only the data from completed legislation periods, which is why we discarded the ongoing eighth period. In Table 1 we provide some details on the final datasets produced for each legislation period – the number of parties along with the smallest and average party-text length (in number of tokens).

Gold Standard. As gold standard party positions we consider the European integration dimension from the Chapel Hill expert survey (years 2002, 2006, 2010). The Chapel Hill expert survey estimates national parties positioning on a variety of policy issues, European integration included. It is conducted every four years (in the occasion of EP elections) since 1999. The number of included countries increased through time, moving from 14 Western European Countries in 1999 to 31 countries in 2014, thus including those EU member states entering the EU during the various EU enlargement steps. The last wave of the Chapel Hill expert survey includes 268 parties from 31 countries.

Task Formulation. Given a series of documents, each one representing the concatenation of all speeches of the candidates of a European party, develop an algorithm able to place them into a single-dimensional space between 0 and 1, where 0 represents a strongly Eurosceptic position and 1 strongly in favour of European integration. To do so, any external resource could be used (i.e., information from a knowledge base such as DBpedia (Auer et al., 2007)), excluding information regarding the political position of the party to be scaled (e.g., the Italian Movimento 5 Stelle as being an Eurosceptic party). This output had to be derived solely from the textual content of the document.

[4] https://sites.google.com/site/nlpandcss/nlpcss-at-emnlp-2016

[5] https://sharedtasksinthedh.github.io/

[6] https://cds.nyu.edu/text-data-speaker-series/

[7] http://essexsummerschool.com/

[8] http://topicmodels.west.uni-koblenz.de/

[9] http://archivesunleashed.org/

[10] https://www.idigbio.org/content/citscribe-hackathon

[11] https://brown.columbia.edu/election-hackathon

[12] We make all the collections used during the hackathon available at: https://federiconanni.com/hack-vigoni/

[13] http://www.europarl.europa.eu

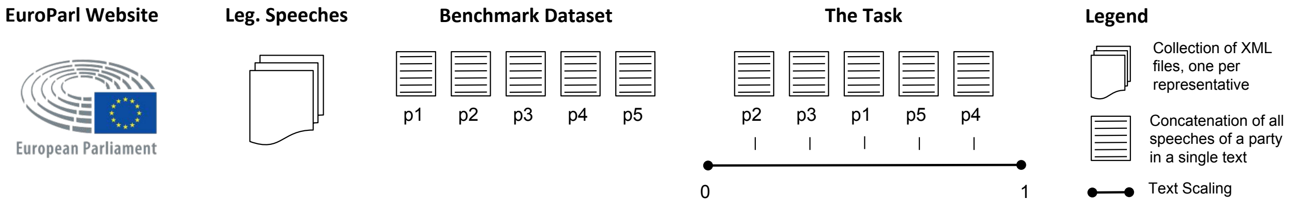

Figure 1: Graphical representation of the creation of the shared task.

5. Overview of the Proposed Approaches

All teams started tackling the scaling problem by manually inspecting the data to observe their structure. Thanks to this, they realized that the majority of the available data were not dealing with issues related to the EU and the European integration project (and therefore to determine if a party is pro or against the EU) but rather with technical aspects belonging to supranational decision-making process (e.g., discussion related to a specific policy issue). To keep only the relevant information, they adopted different filtering strategies. Next, the scaling step has been also approached in different ways. We report below a short overview of the different systems proposed and developed during the hackathon; we also encourage all the participants to continue collaborating on these initial ideas and to present the final results as independent research papers.

Team Greece (Aker, Carlotti, Matsuo, Niekler): To keep only the relevant information, this team used a dictionary for filtering out document irrelevant for the European integration. They identified EU resources (list of terms) available online discussing issues related to the EU and the European integration project, and constructed a dictionary containing only uni-grams and bi-grams. The entries of the dictionary were used in the filtering process. Each speech is regarded as one instance, which consists of multiple sentences. This filtering works at the sentence level. They used the dictionary entries to filter out any sentence within each speech that does not share any entry in the dictionary.

Using the trimmed instances (containing sentences related to the problem) they perform standard bag-of-words feature extraction (with uni-grams and bi-grams) along with feature selection. For feature selection they disregarded any word that occurred in more than 75% of the instances as well as in only 1% of the instances. Furthermore, they used chi-square test to remove further insignificant words leading to a feature vector containing 1500 words.

For each instance they extracted a feature vector containing those significant 1500 terms. The feature values are simple word counts. They used a linear SVM regression model, where the outcome is the true score, with hyper-parameter tuning. The model is capable to score each instance between 1 (pro EU) and 0 (non pro EU). As a number of speech instances are coming from the affiliates of one party, there are a number of predictive scores for each party. They use the median as the final predictive score. For comparison purposes, they repeated the experiments without the filtering process, i.e., feature vectors were extracted without removing any sentence. However, they applied the same feature selection as performed with the dictionary filtering case. They refer to this last experiment as “without dictionary” and the former experiment as “with dictionary”. Against their intuition, the obtained results show that the inclusion of all datasets and sentences performs better on the task than filtering the sentences. This needs further investigation in order to improve and adapt the dictionary to the task.[14]

Team Ireland (Bleier, Menini, Waseem, Yordanova): The team used Will Lowe’s package Jfreq[15] to pre-process the documents: they lowercased, removed numbers and currencies and stemmed all remaining words. A tailored list of stopwords, created considering the specificities of the corpus, was also adopted. Then they used the R implementation of Wordfish (Slapin and Proksch, 2008), from the Austin package, to scale the documents and they tested different word-filtering approaches to improve the results.

Team Italy (Gessler, Hovy, Karan): The team filtered first the speeches based on a list of manually selected keywords, then used paragraph2vec (Le and Mikolov, 2014) on all the speeches (from both scored and unscored parties) to learn distributed party (and word) representations. Then, the resulting matrix of the representation for known parties, together with their respective scores, was used as input to a canonical correlation analysis (CCA) (Hardoon et al., 2004). This step tries to find the first component that explains the variation in the party representations with respect to the observed scores. The fitted CCA model was then applied to the matrix of representations for new parties and the resulting one-dimensional vector was used as score prediction. These scores reach correlation values of up to 0.73 (Spearman) with the gold standard scores.

Team Netherlands (Aprosio, Henrichsen, Nguyen): The team explored a supervised learning approach. The data was segmented into individual speeches. A Linear Regression model was trained based on the labeled data to estimate a score for the individual speeches, and the final score was computed by taking the mean of these scores. The data included speeches covering a wide range of topics. However, speeches about the enlargement were considered the most relevant to a party’s position regarding European integration. Therefore, for the final predictions, only those speeches were included that were about the enlargement based on one of the following words: ‘enlargement’, ‘integration’, ‘accession’, ‘extension’. To overcome the small amount of labelled data, Ridge Regularization was used

[14] Code is available here: `https://github.com/eisioriginal/eu_scepticism_regression`

[15] `http://conjugateprior.org/software/jfreq/`

to prevent overfitting (alpha=1.5) and a small amount of noise was added to the labels. Scikit-learn (Pedregosa et al., 2011) was used to train the models, and the hyperparameters were set using cross-validation on the training set. Only words were kept that appeared in at least 10 speeches and words appearing in more than 10% of the speeches were discarded. Both unigrams and bigrams were used. The results on the validation data were 0.573 (Spearman) and 0.733 (Pearson). The submitted runs included a model trained on both the training and validation data, and a model trained on only the training data.

Team Wales (Kahmann, Posch, Vegetti, Whyte): The approach of the team was based on Party Manifestos data, containing sentences classified (by experts) into different policy categories. The manifestos of UK parties were used because they were the only ones written in English. Some standard pre-processing was applied (lowercasing, removing numbers and stopwords, stemming). The policy categories of interest are European Community/Union (+/−) and National Way of Life (+/−). Based on these sentences a Naive Bayes classifier with three classes was used: (1) not related, (2) pro EU and (3) contra EU. After training this classifier on the manifesto data, it was applied to the test data. Before that, the speeches were split into single sentences and pre-processed. The classifier yielded three values for every sentence of every party. The three values indicate the posterior probability of a sentence belonging to one of the three categories. In order to get a single value for every party, they first excluded all sentences under a certain probability threshold (0.25, 0.5, 0.75) in the first category (not related). Having done this, they calculated a ratio score for every party computed as follows: $\log(\frac{\sum EU_{pro}}{\sum EU_{contra}})$. In the last step they normalized these party scores to the range [0,1].

6. Evaluation

We provided the datasets comprising the 5th and 7th legislation periods as development datasets to the participants. We kept the 6th legislation period (aggregate party texts and gold party positions from Chapel Hill Expert Survey) for final evaluation.

Evaluation Metrics. We use three evaluation metrics for comparing model-produced positions with the gold-standard positions:

- *Pairwise accuracy (PA)* is the percentage of pairs of parties for which the gold scores for the two parties on European integration are in the same relative order as the predicted scores for these two parties. In other words, prediction for a pair of parties A and B is considered correct if party A is more eurosceptic (pro-European) than party B both according to the gold standard and predicted position;

- *Spearman correlation (r_S)* between the set of gold party positions on European integration and those predicted by participants' systems;

- *Pearson correlation (r_P)* between the gold and predicted sets of party positions.

Team/Model	PA (%)	r_S (%)	r_P (%)
Random	51.3	2.6	6.2
WordFish (baseline)	61.8	34.5	29.5
Team Ireland	57.6	17.5	29.4
Team Wales	60.4	28.9	28.2
Team Netherlands	66.8	46.6	59.3
Team Greece	68.5	54.2	64.5
Team Italy	**70.3**	**54.3**	**72.8**

Table 2: Official hackathon results – scaling performance achieved by the best submitted run of each team.

Pairwise accuracy and Spearman correlation capture only the correctness of the ranking of the parties. In contrast, Pearson correlation also takes into consideration the extent to which automated scaling reflects the gold distances between party positions. Put differently, a system that produces the position scores that generate the same party ranking (i.e., the same order of parties from most eurosceptic to most pro-European) as the gold scores will have the perfect PA and r_S, but it will only have perfect r_P if it predicts exactly the same position scores as in the gold standard for all parties. Before evaluating the systems, we linearly scaled both the gold standard scores and system-produced scores to the $[0, 1]$ range.

Results. In Table 2 we show the performance achieved by the best submitted run of each team of participants on the dataset compiled from the 6th legislative period of the European Parliament. Along with the performances of the best runs from all teams, we show the performance of the WordFish model (Proksch and Slapin, 2010), the *de facto* standard model for text scaling in political science. As a sanity check, we also evaluated a baseline that randomly generates party positions.

All teams outperformed the random baseline by a wide margin. Three teams also outperformed the standard scaling algorithm WordFish, with the best-performing approach (Team Italy) outperformed WordFish by 10% in terms of pairwise accuracy, and 20 and 40 points in terms of Spearman and Pearson correlation, respectively.

7. Discussion

In addition to the quantitative outcome of the task, the hackathon made possible that scholars from very different backgrounds, research topics, and methodologies spent two days working together sharing ideas and approaches, each of them excited to think out-of-their-own disciplinary box. While it is generally not so easy to establish such communication channels across disciplines, given the different methodological approaches, research focuses, and even vocabulary (e.g., the meaning of the verb "to code" in computer science and political science), the participants have been incredibly willing to establish a common ground, for cooperating and addressing the task presented to them.

There is still much work that, as organizers of such events, we can do to improve this type of collaborative shared-task, from offering easier-to-digest presentations on the theoretical foundations of the political-science topic under study to establishing methodological debates accessible to the entire audience. Nevertheless, we hope that the collaborations that will bloom thanks to this hackathon will facilitate the communication across research fields and support the future of interdisciplinary research between NLP and political science.

8. Conclusions

In this paper we presented an overview and the results of the first shared-task hackathon organized on the topic of scaling transcripts of speeches from the European parliament regarding Euroscepticism. The hackathon brought together 23 researchers (5 organizers and 18 participants) from 15 institutions with a large variety of backgrounds, from political science to computational social science and natural language processing, which worked together in five small teams for 48 hours in order to develop new approaches for the task. The output of the hackathon has been incredible: in two days these teams developed methods capable of outperforming the most established scaling algorithm in the field, WordFish, by a large margin. This highlights the immense potential of interdisciplinary collaborations and the usefulness of shared-task hackathons for bridging different research communities.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG) for generously supporting the research seminar and the hackathon. This work was also supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4). We thank the staff of the "Villa Vigoni Centro Italo-Tedesco per l'Eccelenza Europea" for their immense hospitality.

9. References

Abelson, R. P. and Carroll, J. D. (1965). Computer simulation of individual belief systems. *The American Behavioral Scientist (pre-1986)*, 8(9):1–24.

Adams, J., Ezrow, L., and Somer-Topcu, Z. (2014). Do voters respond to party manifestos or to a wider information environment? an analysis of mass-elite linkages on european integration. *American Journal of Political Science*, 58(4):967–978.

Almogi, O., Dankin, L., Dershowitz, N., and Wolf, L. (2016). A hackathon for classical Tibetan. *arXiv preprint arXiv:1609.08389*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Budge, I. (2000). Expert judgements of party policy positions: Uses and limitations in political research. *European Journal of Political Research*, 37(1):103–113.

Carlotti, B. (2017). The odd couple: analyzing united kingdom independence party (UKIP) and italian five stars movement's (FSM's) european union (EU)-opposition in the european parliament (EP). *Italian Political Science Review/Rivista Italiana di Scienza Politica*, pages 1–24.

Carrubba, C. J., Gabel, M., Murrah, L., Clough, R., Montgomery, E., and Rebecca, S. (2006). Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-Call Vote Analysis. *British Journal of Political Science*, 36(4):691–704.

Conti, N. (2003). *Party attitudes to European integration: a longitudinal analysis of the Italian case*. Sussex European Institute Brighton.

Cotta, M. (2016). Un concetto ancora adeguato? l'euroscetticismo dopo le elezioni europee del 2014. *Contro l'Europa? I diversi Scetticismi verso l'Integrazione Europea*, pages 233–248.

Crespy, A. and Verschueren, N. (2009). From euroscepticism to resistance to european integration: an interdisciplinary perspective. *Perspectives on European politics and society*, 10(3):377–393.

Curini, L. (2010). Experts' political preferences and their impact on ideological bias: An unfolding analysis based on a Benoit-Laver expert survey. *Party Politics*, 16(3):299–321.

De Vreese, C. (2007). A spiral of euroscepticism: The media's fault? *Acta Politica*, 42(2-3):271–286.

Flood, C. (2002). Euroscepticism: A problematic concept (illustrated with particular reference to France). In *32nd annual UACES conference, Belfast*, pages 2–4.

Gemenis, K. (2015). An iterative expert survey approach for estimating parties' policy positions. *Quality & Quantity*, 49(6):2291–2306.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In **SEM*, pages 125–130.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017a). Cross-lingual classification of topics in political texts. In *NLP+CSS Workshop*, pages 42–46.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017b). Unsupervised cross-lingual scaling of political texts. In *EACL*, pages 688–693.

Graham, S., Milligan, I., and Weingart, S. (2016). *Exploring big historical data: The historian's macroscope*. World Scientific.

Granberg, D. and Brown, T. A. (1992). The perception of ideological distance. *Western Political Quarterly*, 45(3):727–750.

Grimmer, J. and 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.

Halikiopoulou, D., Nanou, K., and Vasilopoulou, S. (2012). The paradox of nationalism: The common denominator of radical right and radical left e euroscepticism. *European Journal of Political Research*, 51(4):504–539.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hix, S., Noury, A. G., and Roland, G. (2007). *Democratic Politics in the European Parliament*. Cambridge University Press, New York.

Hug, S. (2016). Party pressure in the european parliament. *European Union Politics*, 17(2):201–218.

Kopecký, P. and Mudde, C. (2002). The two sides of euroscepticism: party positions on European integration in East Central Europe. *European Union Politics*, 3(3):297–326.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 44 (3):619–634.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97 (2):311–331.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Menini, S. and Tonelli, S. (2016). Agreement and disagreement: Comparison of points of view in the political domain. In *COLING*, pages 2461–2470.

Menini, S., Nanni, F., Ponzetto, S. P., and Tonelli, S. (2017). Topic-based agreement and disagreement in us electoral manifestos. In *EMNLP*, pages 2928–2934.

Mudde, C. (2012). The comparative study of party-based euroscepticism: the sussex versus the north carolina school. *East European Politics*, 28(2):193–202.

Mühlböck, M. and Yordanova, N. (2017). When legislators choose not to decide: Abstentions in the european parliament. *European Union Politics*, 18(2):323–336.

Nanni, F., Zirn, C., Glavaš, G., Eichorst, J., and Ponzetto, S. P. (2016). TopFish: topic-based analysis of political position in US electoral campaigns. In *PolText*.

Nowviskie, B. (2014). On the origin of "hack" and "yack". *Journal of Digital Humanities*, 3(2):3–2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polk, J., Rovny, J., Bakker, R., Edwards, E., Hooghe, L., Jolly, S., Koedam, J., Kostelka, F., Marks, G., Schumacher, G., Steenbergen, M., Vachudova, M., and Zilovic, M. (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*, Online first:1–9.

Proksch, S.-O. and Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, 40(3):587–611.

Proksch, S.-O. and Slapin, J. (2015). *The Politics of Parliamentary Debate: Parties, Rebels, and Representation*. Cambridge University Press, Cambridge.

Rovny, J. (2004). Conceptualising party-based euroscepticism: Magnitude and motivations. *Collegium: news from the College of Europe= nouvelles du Collège d'Europe*, (29):31–48.

Schmitt, H., Braun, D., Popa, S. A., Mikhaylov, S., and Dwinger, F. (2007). *European Parliament Election Study 2014, Euromanifesto Study*. GESIS Data Archive.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Steenbergen, M. R. and Marks, G. (2007). Evaluating expert judgments. *European Journal of Political Research*, 46(3):347–366.

Taggart, P. and Szczerbiak, A. (2002). *The party politics of Euroscepticism in EU member and candidate states*. Sussex European Institute Brighton.

Taggart, P. (1998). A touchstone of dissent: Euroscepticism in contemporary western european party systems. *European Journal of Political Research*, 33(3):363–388.

Usherwood, S. and Startin, N. (2013). Euroscepticism as a persistent phenomenon. *JCMS: Journal of Common Market Studies*, 51(1):1–16.

Usherwood, S. (2016). 2 modelling transnational and pan-european euroscepticism. *Euroscepticism as a Transnational and Pan-European Phenomenon: The Emergence of a New Sphere of Opposition*, page 14.

Vasilopoulou, S. (2009). Varieties of euroscepticism: the case of the european extreme right. *Journal of Contemporary European Research*, 5(1):3–23.

Whitaker, R., Hix, S., and Zapryanova, G. (2017). Understanding members of the european parliament: Four waves of the European parliament research group MEP survey. *European Political Science Review*, 18(3):491–506.

Wilde, P., Michailidou, A., and Trenz, H. (2014). Converging on euroscepticism: Online polity contestation during european parliament elections. *European Journal of Political Research*, 53(4):766–783.

Yordanova, N. and Mühlböck, M. (2015). Tracing the bias in roll call votes: Party group cohesion in the european parliament. *European Political Science Review*, 7(3):373–399.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *PolText*.