Proceedings of the 1st Workshop on Computational Linguistics for Political Text Analysis (CPSS-2021)
Düsseldorf, Germany

Ines Rehbein ♠, Gabriella Lapesa ♣, Goran Glavaš ♠, Simone Ponzetto ♠
University of Mannheim ♠, University of Stuttgart ♣

Sep 6, 2021

Conference Program

September, 6

9:00 – 10:00 Welcome & Invited Talk

10:00 – 11:00 Oral Session 1

11:00 – 11:15 Coffee Break

11:15 – 12:15 Twin Panel

12:15 – 13:15 Lunch Break

13:15 – 14:15 Oral Session 2

14:15 – 15:00 Invited Talk II

15:00 – 15:15 Coffee Break

15:15 – 16:15 Poster Session

16:15 – 17:00 Invited Talk III

17:00 – 17:45 Open Discussion

INVITED TALK III
Comparability and interoperability of parliamentary corpora: Easier said than done

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

The talk presents the ParlaMint corpora containing transcriptions of the sessions of 17 European national parliaments with half a billion words. The corpora are uniformly encoded, contain rich meta-data about the 11 thousand speakers, and are linguistically annotated following the Universal Dependencies formalism and with named entities. Samples of the corpora and conversion scripts are available from the project's GitHub repository, the complete corpora are deposited on the CLARIN.SI repository under the CC BY license, and available through its NoSketch Engine and KonText concordancers for exploration and analysis. The corpora are the result of the CLARIN ParlaMint project (2019-2021), and the talk presents the project, the corpus compilation workflow, the Parla-CLARIN-based encoding of the corpora and their distribution. We concentrate on the most difficult aspect of the project, which was the goal to make the corpora interoperable while at the same time having a large number of partners each one in charge of producing their own corpus.

Small data problems in political research: a critical replication study

Hugo de Vos
Institute of Public Administration
Leiden University
h.p.de.vos@fgga.leidenuniv.nl

Suzan Verberne
Leiden Institute of Advanced
Computer Science (LIACS)
Leiden University
s.verberne@liacs.leidenuniv.nl

Abstract

In an often-cited 2019 paper on the use of machine learning in political research, Anastasopoulos & Whitford (A&W) propose a text classification method for tweets related to organizational reputation. The aim of their paper was to provide a 'guide to practice' for public administration scholars and practitioners on the use of machine learning. In the current paper we follow up on that work with a replication of A&W's experiments and additional analyses on model stability and the effects of preprocessing, both in relation to the small data size. We show that (1) the small data causes the classification model to be highly sensitive to variations in the random train–test split (2) the applied preprocessing causes the data to be extremely sparse, with the majority of items in the data having at most two non-zero lexical features. With additional experiments in which we vary the steps of the preprocessing pipeline, we show that the small data size keeps causing problems, irrespective of the preprocessing choices. Based on our findings, we argue that A&W's conclusions regarding the automated classification of organizational reputation tweets – either substantive or methodological – can not be maintained and require a larger data set for training and more careful validation.

1 Introduction

In[1] 2019, the Journal of Public Administration Research and Theory (JPART) published a paper on the use of Machine Learning (ML) in political research (Anastasopoulos and Whitford, 2019) (A&W). With this paper, A&W attempt 'to fill this gap in the literature through providing an ML "guide to practice" for public administration scholars and practitioners' (Anastasopoulos and Whitford, 2019, p. 491). A&W present an example study, in which they aim to 'demonstrate how ML techniques can help us learn about organizational reputation in federal agencies through an illustrated example using tweets from 13 executive federal agencies' (Anastasopoulos and Whitford, 2019, p. 491). In the study, a model was trained to automatically classify whether a tweet is about moral reputation or not. According to the definition scheme by A&W, a tweet addresses moral reputation if it expresses whether the agency that is tweeting is compassionate, flexible, and honest, or whether the agency protects the interests of its clients, constituencies, and members (Anastasopoulos and Whitford, 2019, p. 509). The conclusion of the example study was that 'the Department of Veterans Affairs and the Department of Education stand out as containing the highest percentage of tweets expressing moral reputation.' (Anastasopoulos and Whitford, 2019, p. 505).

A&W also provided a concise, but more general, introduction to machine learning for Public Administration scientists, of which the example study was an integral part illustrating how machine learning studies could work. The concise overview on supervised machine learning makes the paper a valuable addition to the expanding literature on machine learning methods in political research. However, the example study contains several shortcomings that are not addressed by A&W. A possible undesired result is that practitioners or researchers unfamiliar with machine learning will follow the wrong example and consequently conduct a flawed study themselves. It is for this reason that we zoom in on the data used in the example study and the validation that is reported by A&W, showing the problems with their study.

A&W train a Gradient Boosted Tree model with bag-of-words features on the binary classification

[1] All data and scripts are published at: https://anonymous.4open.science/r/Critical_Replication_ML_in_PA-3F20/README.md

task to recognize whether a tweet is about moral reputation or not. The model is first trained on a data set of 200 human-labeled tweets and evaluated using a random 70-30 train–test split. The trained model is then used to automatically infer a label for 26,402 tweets. Based on this larger data set, A&W analyze to what extent specific US institutions work on their moral reputation via Twitter.

The core problem with this set-up is that the training data set is too small to train a good model. We show that this results in a model that is of drastically different quality when the random split of the data is varied, an effect that we will call model (in)stability. The consequences of these mistakes are that the model by A&W can not reliably be used for data labeling, because data generated with this model can not be assumed to be correct. Although the mistakes can only be solved with a larger data set, the flaws could have been detected if the model would have been validated more thoroughly by the authors.

The consequences for the conclusions in the A&W paper itself might be relatively small, because it is only one example without overly strong substantive claims. However, more importantly, the weaknesses of the paper might also influence any future research based on the study; the paper was published in a high-impact journal and has been cited 49 times since 2019.[2]

In this paper, we replicate the results by A&W, and analyze their validity. We perform what Belz et al. (2021) call a *reproduction under varied conditions*: a reproduction where we "deliberately vary one or more aspects of system, data or evaluation in order to explore if similar results can be obtained" (p. 4). We show that the A&W results can indeed be reproduced, yet only in very specific circumstances (with specific random seeds). We demonstrate that the methods have flaws related to data size and quality, which lead to model instability and data sparseness. This means that the 'guide to practice' that A&W aim to provide requires careful attention by any follow-up work.

We address the following research questions:

1. What is the effect of small training data on the stability of a model for tweet classification?

2. To what extent do changes in the preprocessing pipeline influence the model quality and stability in combination with the small data size?

We first make a comparison between the data set of A&W and other text classification studies in the political domain (Section 2). We then report on the replication of A&W's results, followed by an analysis of the model stability under the influence of different random data splits (Section 3). In Section 4 we conduct additional experiments varying the preprocessing pipeline to further analyze the implications of the small data size on the usefulness of the data for the classification task. We conclude with our recommendations in Section 5.

2 Related work on political text classification and data size

In the field of political science, text mining methods (or Quantitative Text Analysis (QTA) as it is called in the Political Science community) have been used for about a decade. One of the first major papers on the use of automatic text analysis in the field was Grimmer and Stewart (2013). In this seminal paper the pros and cons of using automatic text analysis are discussed.

Another major contribution to the field is the Quanteda package (Benoit et al., 2018) in R. This R package contains many tools for Quantitative Text Analysis such as tokenization, stemming and stop word removal and works well with other (machine learning) R packages like topicmodels (Grün et al., 2021) and xgboost (Chen and Guestrin, 2016). This package that has been developed by and for Political Scientists and Economists has already been widely used in the community.

A&W used the tm package (Feinerer and Hornik, 2021) for text mining in R. The data set used to train their machine learning model consists of a total of two hundred tweets. Eighty two of those were manually labeled by the authors as being about moral reputation and 118 as not being about moral reputation.[3] The average length of a tweet in the data set is 17.7 words with a standard deviation of 4.4.

In comparison to other studies that used machine learning for tweet classification, 200 tweets is notably small. The issue of the small data size is

[2] According to Google Scholar, June 2021

[3] Originally, they also had the tweets annotated via crowd sourcing, but the resulting annotations had such a low intercoder reliability that they decide not to used them due to the poor quality.

aggravated by the short length of tweets: They contain few words compared to other document types such as party manifestos (Merz et al., 2016; Verberne et al., 2014) or internet articles (Fraussen et al., 2018). Because tweets are so short, the bag-of-words representation will be sparse, and in a small data set many terms will only occur in one or two tweets. This makes it difficult to train a generalizable model, as we will demonstrate in Section 4.

Based on the literature, there is no clear-cut answer to how much training data is needed in a text classification task. This depends on many variables, including the text length, the number of classes and the complexity of the task. Therefore we can not say how many tweets would have sufficed for the goal of A&W. What is clear from related work, is that it should be at least an order of magnitude larger than 200. Elghazaly et al. (2016), for example, used a set of 18,278 hand-labeled tweets to train a model for recognizing political sentiment on Twitter. Burnap and Williams (2015) used a set of 2,000 labeled tweets to train a model that classifies the offensiveness of Twitter messages. Amador Diaz Lopez et al. (2017) used a total of 116,866 labeled tweets to classify a tweet about Brexit as being Remain/Not Remain or Leave/Not Leave.

Most, if not all, of the recent work in the field of computational linguistics uses transfer learning from large pre-trained language models for tweet classification, in particular BERT-based models (Devlin et al., 2018). In these architectures, tweets can be represented as denser vectors, and the linguistic knowledge from the pretrained language model is used for representation learning. The pretrained model is finetuned on a task-specific dataset, which in most studies is still quite large. Nikolov and Radivchev (2019), for example, used a training set of 13,240 tweets (Zampieri et al., 2019) to fine-tune a BERT model to classify the offensiveness of a tweet. This resulted in an accuracy of 0.85.

A more general point of reference about sample sizes for tweet classification is the SemEval shared task, a yearly recurring competition for text classification often containing a Twitter classification task. For example, in 2017 there was a binary sentiment analysis task where participants could use a data set of at least[4] 20,000 tweets to train a model (Rosenthal et al., 2019).

These studies show that even in binary classification tasks using twitter data, a lot of data is often needed to achieve good results, despite that those tasks might look simple at first glance. In the next section, we empirically show that the A&W data is too small for reliable classification.

3 Replication and model stability

A&W report good results for the classifier effectiveness: a precision of 86.7% for the positive class ('about moral reputation'). In this section we present the results of an experiment that we did to validate the reported results. In addition to that we will also assess the stability of the model. By this we mean how much the model and its performance changes when the data is split differently into a train and test set. We argue that if an arbitrary change (like train test split) leads to big changes in the model, the generalizability of the model is poor, because it shows that changes in data sampling results in changes in model quality, and hence in different classification output.

3.1 Exact replication

We first completed an exact replication of the experiment of A&W to make sure we started from the same point. We followed the data analysis steps described in A&W exactly. Thanks to the availability of the data and code, the study could be replicated with ease. The exact replication yielded the same results as reported in A&W.

3.2 Varying the random seed

In their experiments A&W make a random 70-30 train–test split of the 200 labelled tweets: 140 tweets are randomly sampled to be the train set and the remaining 60 tweets form the test set. In their paper, they present the result of only a single random split. For reproducibility reasons A&W use a single random seed for the train–test split.[5]

In order to assess the generalizability of the model, we generated a series of one thousand random seeds (the numbers 1 to 1000). This resulted in a thousand different train-/test splits of the tweets. We reran the experiment by A&W with all the random train–test splits, keeping all other settings unchanged. In all cases, the train set contained 70% (140) of the labeled tweets and the test set 30% (60) of the labeled tweets. For each of the thousand

[4] There were other tasks where more training data was available.

[5] In their case this seed is 41616

runs we calculated the precision, in the same way that A&W did.

If a model is robust, most of the different configurations should yield approximately the same precision. Inevitably, there will be some spread in the performance of the models but they should group closely around the mean precision which indicates the expected precision on unseen data.

3.3 Results of varying the random seed

Our experiment resulted in precision scores that ranged from 0.3 to 1.0. The mean precision was 0.67 with a standard deviation of 0.14. The median was 0.69. The mean and standard deviations of the 1000 runs for precision, recall and F1 are listed in Table 1. The distribution of precision values is also depicted in the leftmost boxplot in Figure 1. The table indicates that the model on average performs rather poorly for a binary classification task: the F-score for the positive class is 0.40 and for the negative class 0.75. In addition, the plot as well as the standard deviations in the table show a large variance in quality between different random seeds. This indicates that the model is unstable.

	Class	
	Positive	Negative
Precision (sd)	0.69 (0.14)	0.65 (0.06)
Recall (sd)	0.30 (0.10)	0.90 (0.08)
F1-score (sd)	0.40 (0.09)	0.75 (0.05)

Table 1: The means and standard deviation for the evaluation statistics.

What also stands out is that the result by A&W (the horizontal red line in Figure 1) appears to be exceptionally high. Out of the 1000 runs, only 6 were able to match or outperform the precision presented in A&W (.867). The mean precision over 1000 runs is much lower than the precision reported by A&W. We argue that the mean precision over 1000 runs is more likely to be a realistic reflection of the actual model precision than the result for one random split.

From these results, we conclude that the model quality is relatively poor and unstable: changing the train–test split, an arbitrary alteration that should not make a big difference, leads to a wide range of outcomes. This has an effect on the generalizing power of the machine learning model: Although the reported results on the test set (with only one particular random seed) are good, they are not generalizable to other data splits.

That the model generalizes poorly is in fact confirmed by Figures 3 and 5 in Anastasopoulos and Whitford (2019, p. 503 and 506). These figures show that solely the occurrence of the word 'learn' or 'veteran' will make the model predict that a tweet is about moral reputation, regardless of any other words occurring in the tweet. This is an effect of these words being overrepresented in the data sample. This artefact effect is more likely to occur if a data sample is too small. This situation will lead to overfitting of the model, a likely effect that is not described by A&W. We explore the effects of the small data size in more detail in the next section.

4 Implications of small data sets on data quality

In the previous section we showed how the small amount of data leads to poor model stability. In this section we show how the small number of tweets negatively affects the quality of the data set that serves as input to the machine learning model. We also experiment with other preprocessing choices to investigate the effect on the model quality and stability.

A&W apply a number of common preprocessing steps to their data:

- Decapitalisation (e.g. 'Veteran' → 'veteran')
- Removal of all special characters, numbers, punctuation, and URLs
- Stop-word removal
- Removal of rare terms: all words that occur in fewer than 2% of the tweets are removed from the data.
- Stemming with the SnowballC stemmer (Bouchet-Valat, 2020)

The remaining unigrams are used as count features in the bag-of-words model.

In the next two subsections, we first analyze the effect of word removal (stop word and rare words), and then investigate the effect of changing the preprocessing steps on the quality of the model.

4.1 The effect of removing words

As introduced above, A&W remove both stop words and rare words from the data before the document–term matrix is created. Examples of

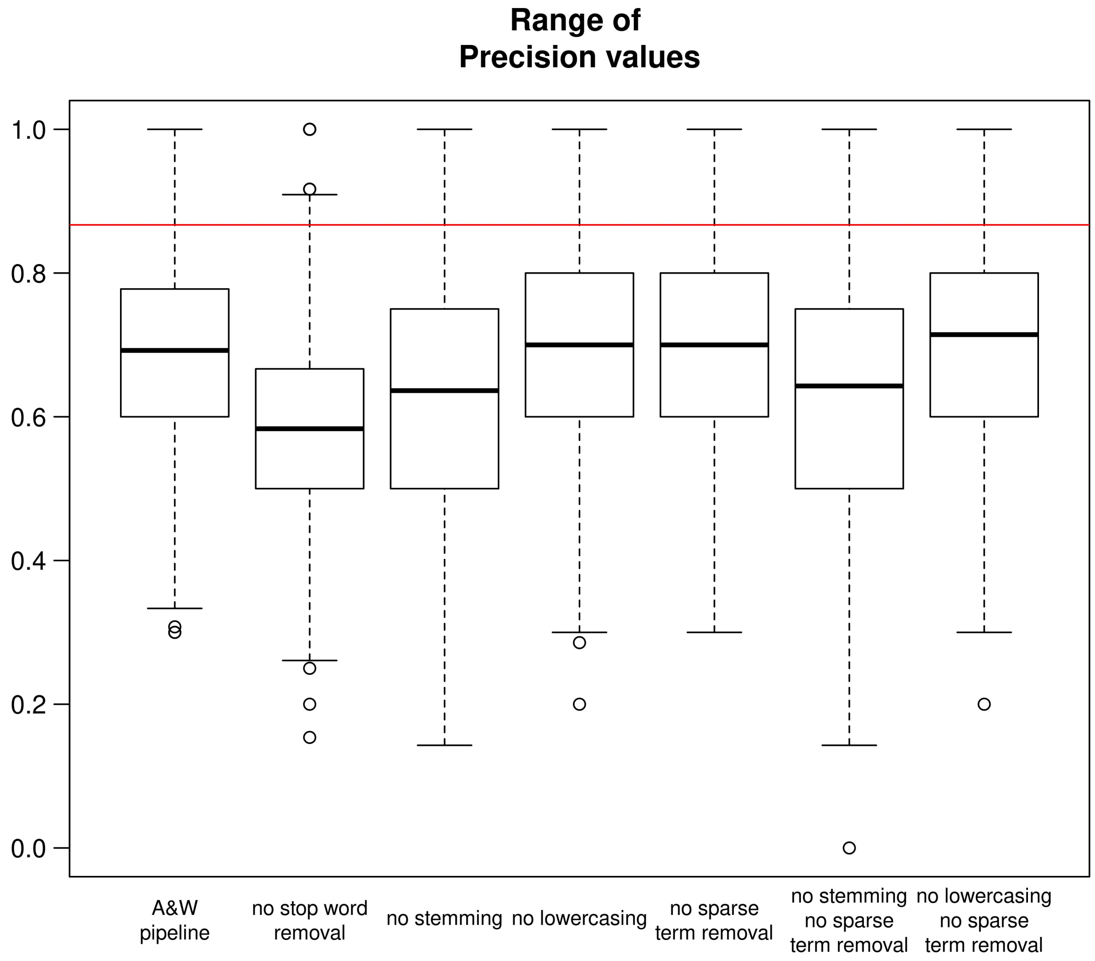


Figure 1: A visualization of the spread of results of the random seed variation experiment. The leftmost box summarizes the results of 1000 different runs with the same settings as A&W, except for the random seeds. The horizontal red line depicts the precision that is reported by A&W. The other box plots are the results of 1000 runs where each time one preprocessing step is omitted as described in section 4.2.

stop-words removed by A&W are ‘they’, ‘are’, ‘is’ and ‘and’. Removing such words prevents a model from learning that, for example, the word ‘the’ signals that a tweet is about moral reputation because the word ‘the’ occurs, by chance, more often in tweets about moral reputation.

Similarly, rare words are not considered to be a relevant signal. For example, the word ‘memorabilia’ occurs only one time in the tweet collection of A&W, and this happens to be in a tweet about moral reputation. A machine learning algorithm could, therefore, infer that ‘memorabilia’ contributes positively to a tweet being about moral reputation, which is not a generalizable rule. For this reason words that occur only rarely are commonly removed, as do A&W.

However in combination with the small data size, the effect is that almost every word is either a stopword or a rare word. Consequently, removing stop words and rare words leads to tweets from which almost every word is deleted. In fact, in the preprocessing setting of A&W, 95% of all the tokens in the collection were removed, reducing the dictionary size from 1473 to 70. As a result, many tweets have fewer than three non-zero features, making it difficult for the model to predict the label of those tweets.

This effect is further illustrated in Table 2, which lists the number of tweets from the data set with a given number of words. This table shows that after removing rare words and stop words, 15% of the tweets in the collection have no non-zero features at all, and 24% percent are represented by only one non-zero feature. As a result of this, the model tried to learn how to recognize whether a tweet is about moral reputation or not based on tweets with barely any words in them.

The situation is even more clear in the unlabeled collection. In this set, from 25% of the tweets every word was removed. By coincidence, the model in A&W learned that every tweet with no words left was about moral reputation. This means that 25% of the data set on which A&W based their conclusion, has received the label 'about moral reputation', while this is impossible to say based on zero words. This means that at least 25% of the tweets' labels can not be trusted.

The instability can be clarified further with a few examples. Example 1 (a tweet by @USTreasury with the label 'not about moral reputation') has only the words 'new' and 'provides' left after preprocessing. From example 2 (by @USDOT with the label 'not about moral reputation') only the word 'today' is left. Example 3 (by @CommerceGov) is 'about moral reputation' and only the word 'learn' is left.

1. **Before preprocessing**: "We have a new mobile website that provides a virtual tour of 1500 Penn <url><url>''
   After preprocesing: "new provides"

2. **Before preprocessing**: "RT @SenateCommerce TODAY AT 10AM @SenateCommerce to hold a hearing to examine #InfrastructureInAmerica with testimony from @SecElaineChao"
   After preprocessing: "today"

3. **Before preprocessing**: "RT @NASA: We've partnered with @American_Girl to share the excitement of space and inspire young girls to learn about science, technology,..."
   After preprocessing: "learn"

It is difficult – if not impossible – to train a reliable model on these very limited representations of tweets.

This could have been prevented if the number of tweets would have been larger. As a consequence of Heaps' law, the number of new unique terms becomes smaller with every new document that is added (Heaps, 1978). As a result of this, a document collection with more documents/tweets will have fewer rare terms.

4.2 The effect of preprocessing differences

We investigated what the effect on the quality of the model is of different preprocessing choices. We created variants of A&W's pipeline with one of the following adaptations:

- Not removing stopwords
- No stemming
- No lowercasing
- Not removing rare words
- No stemming and not removing rare words
- No lowercasing and not removing rare words

Like in Section 3 we ran each model 1000 times with different random seeds and show the range of precision values for each setting in Figure 1. This shows that there are differences between the preprocessing settings, but the model remains highly unstable and has relatively low median precision scores between 0.59 and 0.71 for the different preprocessing choices.

The different preprocessing steps naturally lead to different dictionary sizes (The number of variables in the document–term matrix). Not lowercasing, for example, increases the number of terms in the dictionary, as words like 'veteran' and 'Veteran' are now seen as diferent tokens. The effect of the different preprocessing steps on the dictionary sizes is listed in Table 3.

Table 3 shows that omitting any of the preprocessing steps (except rare term removal) increases the dictionary size. This makes sense, because all those steps are designed to reduce the dictionary size by collating different word forms to one feature or removing words. In the case of no stopword removal, the dictionary size after rare term removal is larger than if the pipeline of A&W is applied. This can be explained since the stopwords that remain, are never rare terms and thus are not removed. This also explains why there are almost no tweets with only 0 or 1 terms in this setting, because almost every tweet contains a stopword.

Omitting the stemming procedure leads to a larger dictionary size before, but a smaller dictionary size after rare term removal. Because terms are not collated, there will be more unique terms, but all those terms are more likely to be rare. The effect of more terms being removed also shows in the large amount of tweets with 0 or 1 term. The

N	0	1	2	3	4	5	6	7	8
Coded set	25 (15%)	47 (24%)	52 (26%)	37 (19%)	13 (7%)	11 (6%)	4 (2%)	4 (2%)	1 (0.05%)
Uncoded set	6519 (25%)	8099 (31%)	6295 (21%)	3558 (13%)	1349 (5%)	441 (1.7%)	108 (0.4%)	30 (0.1%)	–

Table 2: The amount and proportion of tweets from the human-labeled set and the uncoded set that contain N words.

	Dict size		% of tweets with n terms after rare term removal	
experiment	before rare term removal	after rare term removal	0 terms	1 term
A&W	1473	70	15 %	24 %
No stopword removal	1529	96	2 %	8 %
No stemming	1623	47	25 %	35 %
No lowercasing	1515	73	13 %	25 %
No rare term removal	1473	NA	NA	NA
No stemming and rare term removal	1623	NA	NA	NA
No lowercasing and rare term removal	1515	NA	NA	NA

Table 3: The size of the dictionary as the result of omitting different preprocessing steps before and after the removal of rare terms. Also the percentage of tweets with 0 and 1 terms after rare term removal is listed.

effect that 60% of the tweets only contains 0 or 1 words (25+35%) explains why the settings without stemming are the least stable settings of all (Figure 1).

Not lowercasing the tweets only seems to have a marginal effect. This is likely due to the fact that the number of (non rare) words starting with a capital letter is already small to begin with.

In conclusion, Figure 1 shows that the effect of preprocessing choices has on the precision is relatively small, if anything omitting the preprocessing steps made the models worse on average. This confirms that the data set size is detrimental to the model quality – even after lowercasing, stemming, removing stopwords and rare words, the model can not generalize between different data sampling splits.

5 Conclusions

In this paper, we replicated and analyzed a study that was published in JPART that explains and illustrates how to use machine learning for analyzing Twitter data. The data set used in the example study was too small to train a reliable model. We demonstrated this with a number of experiments: First, we replicated the example study exactly, then we studied the stability of the model by varying the train–test split. In the final experiment, we analyzed the effect of different preprocessing choices on the quality of the data and, subsequently, the quality of the model.

Answers to research questions We found that the results by A&W could be replicated, but only under very specific conditions; our experiment with 1000 random train–test splits showed that only 6 of those 1000 splits could meet or outperform the precision reported by A&W. We find a median precision of 69%, as opposed to the 86.7% reported by A&W. In response to RQ1, what the effect of small training data on the stability of a model for tweet classification is, we show that the small data size has caused the model to be highly unstable, with precision scores ranging from 30% to 100% depending on the train–test split used.

We analyzed the effect of choices in the preprocessing pipeline by varying them. In each setting, the range of precision scores obtained in 1000 train–test splits was large and none of the settings could improve upon the A&W setting. In response to RQ2, to what extent changes in the preprocessing pipeline influence the model quality and stability, we show that the effect of preprocessing choices is relatively small; we obtain median precision scores between 59% and 71% with large standard deviations. We conclude that the data set is too small to train a stable, high-quality model, largely irrespective of the preprocessing steps.

Overall, we showed that the small data issues reduce the validity of the results reported in A&W, especially as a machine learning example for the political research community.

Recommendations for future work As discussed in Section 2, there is no golden rule for how much training data is needed. In general; the shorter a document is, the more documents you need in the training set. In the case of tweets, one would need at least a few thousand hand-labeled training examples. Also, it is important to always report the size of the data set. Not only the number of documents/tweets but also the average number of words in each document.

Apart from recommendations on data set size, we also showed that validation of the model stability can be done by varying the random seed. This can indicate whether more training data is needed for a reliable classifier.

Any researchers seeking to follow up on A&W in designing a machine learning study could additionally consult Lones (2021), a concise overview of a multitude of points to consider to avoid machine learning pitfalls.

Finally, we would like to stress the importance of replication and reproducability. As is noted in Cohen et al. (2018) and Belz et al. (2021) replication studies in NLP are becoming more common in recent years. Belz et al. (2021) conclude that "worryingly small differences in code have been found to result in big differences in performance." (p. 5). This statement is reinforced by the findings in our paper.

A precondition for good debates in social and political sciences based on the outcomes of NLP experiments is that those outcomes are demonstrably reliable. If the results are not robust, a further debate based on the implications of the results is pointless.

References

Julio Cesar Amador Diaz Lopez, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo. 2017. Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data. *Statistics, Politics and Policy*, 8(1).

Jason L Anastasopoulos and Andrew B Whitford. 2019. Machine Learning for Public Administration Research, With Application to Organizational Reputation. *Journal of Public Administration Research and Theory*, 29(3):491–510.

Anja Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Milan Bouchet-Valat. 2020. *Snowball Stemmers Based on the C 'libstemmer' UTF-8 Librar*. R package version 0.6.0.

Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making: Machine Classification of Cyber Hate Speech. *Policy & Internet*, 7(2):223–242.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 785–794, San Francisco, California, USA. ACM Press.

K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, page 156. NIH Public Access.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tarek Elghazaly, Amal Mahmoud, and Hesham A. Hefny. 2016. Political Sentiment Analysis Using Twitter Data. In *Proceedings of the International Conference on Internet of things and Cloud Computing - ICC '16*, pages 1–5, Cambridge, United Kingdom. ACM Press.

Ingo Feinerer and Kurt Hornik. 2021. *tm: Text Mining Package*. R package version 0.7-8.

Bert Fraussen, Timothy Graham, and Darren R. Halpin. 2018. Assessing the prominence of interest groups in parliament: a supervised machine learning approach. *The Journal of Legislative Studies*, 24(4):450–474.

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Bettina Grün, Kurt Hornik, David Blei, John Lafferty, Xuan-Hieu Phan, Makoto Matsumoto, Takuji Nishimura, and Shawn Cokus. 2021. *topicmodels*. R package version 0.2-12.

Harold Stanley Heaps. 1978. *Information retrieval, computational and theoretical aspects*. Academic Press.

Michael A Lones. 2021. How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.

Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):205316801664334.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. SemEval-2017 Task 4: Sentiment Analysis in Twitter. *arXiv:1912.00741 [cs]*. ArXiv: 1912.00741.

Suzan Verberne, Eva D'hondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Frame Detection in German Political Discourses: How Far Can We Go Without Large-Scale Manual Corpus Annotation?

Qi Yu[1,3] and **Anselm Fliethmann**[2,3]

[1]Department of Linguistics, University of Konstanz, Germany
[2]Department of Politics and Public Administration, University of Konstanz, Germany
[3]Cluster of Excellence "The Politics of Inequality", University of Konstanz, Germany
firstname.lastname@uni-konstanz.de

Abstract

Automated detection of *frames* in political discourses has gained increasing attention in natural language processing (NLP). Earlier studies in this area however focus heavily on frame detection in *English* using *supervised* machine learning approaches. Addressing the difficulty of the lack of annotated data for training and/or evaluating supervised models for low-resource languages, we investigate the potential of two NLP approaches that do not require large-scale manual corpus annotation from scratch: 1) LDA-based topic modelling, and 2) a combination of word2vec embeddings and handcrafted framing keywords based on a novel, expert-curated framing schema. We test these approaches using a novel corpus consisting of German-language news articles on the "European Refugee Crisis" between 2014-2018. We show that while topic modelling is insufficient in detecting frames in a dataset with highly homogeneous vocabulary, our second approach yields intriguing and more humanly interpretable results. This approach offers a promising opportunity to incorporate domain knowledge from political science and NLP techniques for bottom-up, explorative political text analyses.

1 Introduction

Print media plays a substantial role in forming public opinion. *Framing*, defined by Entman (1993) as "select[ing] some aspects of a perceived reality and mak[ing] them more salient in a communicating text (...)", has been shown by political communication studies to have a regular influence on citizens' political opinions (Nelson and Oxley, 1999; Druckman, 2004; Slothuus, 2008). In the field of NLP, recent years have witnessed growing attention for the automated detection of frames in political discourse (e.g., Baumer et al., 2015, Card et al., 2016, Field et al., 2018, Khanehzar et al., 2019, Cabot et al., 2020).

Notwithstanding these developments, earlier studies comprise two major limitations. First, many of these studies apply supervised machine learning approaches and thus rely heavily on manually labeled data (a detailed review follows in Section 2). Second, as a consequence of this need of manually labeled data, the majority of the earlier studies utilize the English-language, human-annotated Media Frames Corpus (MFC; Card et al., 2015), thus neglecting framing in non-English language contexts, for which only few or no annotated data is available. Specifically, since the annotation of frames requires a deep understanding of both the text material itself and the background of the issue discussed in the text, creating large-scale annotated datasets in a high quality - such as the MFC - is time-consuming and labor intensive. This expensive enterprise would therefore be prohibitive for many low-resource languages.

To address these two limitations, this paper investigates the potential of unsupervised and knowledge-based NLP approaches for automated frame detection in cases where few to none labeled data is available. We use non-annotated German-language newspaper articles on the so-called "European Refugee Crisis" of 2014-2018 as data, and experiment with two approaches: 1) LDA-based topic modelling (Blei et al., 2003), and 2) a combination of word2vec (Mikolov et al., 2013) and handcrafted framing keywords. Our contributions are three-fold:

1) We show that topic modelling is insufficient in detecting frames in a dataset with highly homogeneous vocabulary;

2) We propose a novel framing schema, the *Refugees and Migration Framing Schema*, which is specifically designed to analyze frames in the context of refugees and migration;

3) We show that the combination of word2vec and handcrafted framing keywords based on our *Refugees and Migration Framing Schema* has a greater potential than topic modelling when conducting data-driven explorations of frame differences. Also, the results are more explainable.

2 Related Work

Owing to the public availability of the large-scale MFC, which includes manual annotations of frames based on the codebook of Boydstun et al. (2014), a large amount of previous work on frame detection focuses on the classification of the frame categories annotated in the MFC. The methods used vary from neural networks, such as Ji and Smith (2017) (RNN) and Naderi and Hirst (2016) (LSTM and GRU), to state-of-the-art language models as in Khanehzar et al. (2019) (XLNet, BERT and RoBERTa) and Cabot et al. (2020) (multi-task learning models combined with RoBERTa). Other studies that use a similar supervised or weakly supervised setting, but other manually annotated datasets than the MFC, include Baumer et al. (2015), Johnson et al. (2017), Liu et al. (2019) and Mendelsohn et al. (2021).

Frame detection in languages other than English remains so far greatly neglected. To the best of our knowledge, Field et al. (2018) and Akyürek et al. (2020) are the only two studies of this kind. Field et al. (2018) employ the annotations in MFC to extract a frame lexicon for each frame category. This English-language lexicon is then translated to Russian and used for identifying frames in Russian newspapers. Their work provides a transferable method for other languages lacking annotated data. Akyürek et al. (2020) use multilingual transfer learning to detect frames in low-resource languages by translating framing-keywords extracted from the MFC to the target language and then training classifiers on the code-switched texts. However, an application of this method on a low-resource target language still requires an available gold standard of that target language, in order to evaluate the meaningfulness of the trained model. In Akyürek et al. (2020), this is again achieved by manually annotating the texts of the target language.

3 Data Collection

To investigate the effectiveness of NLP approaches that do not require large-scale corpus annotation from scratch in the task of frame detection, our study uses a novel corpus of German newspaper articles on the ”European Refugee Crisis” between 2014-2018 as data, for which no prior annotation of frames is available. In order to build a wide representation of different styles (broadsheet vs. tabloid) and political orientations of the German press, while at the same time assuring comparability between newspapers, we selected the newspapers *BILD*, *Frankfurter Allgemeine Zeitung* (FAZ) and *Süddeutsche Zeitung* (SZ) for our study. All three are nation-wide daily newspapers. With FAZ, which is considered slightly right-leaning, and SZ, which is considered center-left-leaning (Pew Research Center, 2018), our sample is balanced and covers a range of the political spectrum within the media landscape in Germany. Moreover, by including BILD, we not only incorporate a tabloid, but also bring together the three most highly-circulated printed newspapers in Germany (Deutschland.de, 2020).

From each newspaper, articles containing at least one match of the following keywords (including all their inflected forms) were selected: {*Flüchtling*, *Geflüchtete*, *Migrant*, *Asylant*, *Asylwerber*, *Asylbewerber*, *Asylsuchende*}. We refer to this set of keywords as *refugee-keywords* in later sections. In a post-hoc cleaning phase, articles with a ratio of *refugee-keywords* smaller than 0.01 and articles from non-political sections such as *Sport* were excluded. We used the keyword-ratio as criterion instead of a keyword-count due to large differences in article length. After the cleaning phase, we obtained the dataset reported in Table 1. [1]

newspaper	category	#articles	#tokens
BILD	R, T	12,287	3,554,105
FAZ	R, B	6,832	3,526,323
SZ	L, B	4,770	1,893,868

Table 1: Dataset overview. (R = right-leaning; L = left-leaning; T = tabloid; B = broadsheet)

[1] The newspaper articles were purchased from the respective publisher. Due to their copyright regulations, the articles, and accordingly the resulting corpus reported above, cannot be distributed to third parties. However, we release the lexical resource resulting from this paper (see Section 5), which is available under: https://github.com/qi-yu/refugees-and-migration-framing-vocabulary

4 Experiment 1: Detecting Frames Using Topic Modelling

As the task of detecting frames strongly resembles the detection of sub-topics within the event under discussion, it is tempting to use topic modelling as a first bottom-up, data-driven exploration of differences in frames between the newspapers. In line with this consideration, we trained one LDA-based model per newspaper to explore frame differences between the publications.

4.1 Training

We used the Python library *Gensim* (Řehůřek and Sojka, 2010) to train the models. Monograms, bigrams and trigrams are used for training. The following preprocessing steps were done prior to the training:

1) All articles were tokenized and lemmatized using the *Stanza* NLP kit (Qi et al., 2020). All stop words, numbers, punctuation marks and URLs were removed;

2) For each newspaper, n-grams with a document frequency higher than 0.15 and n-grams occurring less than 5 times were excluded;[2]

3) Since the *refugee-keywords* appear in all articles, we masked them in order to eliminate their interference in the topic modelling algorithm. Note that not all of them can be excluded by step 2) since not all of them have a document frequency higher than 0.15.

Topic modelling requires the number of topics K to be pre-defined. As we do not have gold standard data available, we use the C_v coherence score as a measure to search for the optimal value of K, as well as to evaluate the model performance. The C_v coherence score is proposed by Röder et al. (2015) as the best performing coherence measure. C_v yields a value in the range of $[0, 1]$. The closer the value is to 1, the more coherent the topics are.

4.2 Results and Discussion

Figure 1 shows the C_v coherence scores of the LDA models trained respectively on BILD, FAZ and SZ for $K \in [2, 200]$, using 50 iterations. As indicated in the figure, C_v stops growing significantly after $K = 80$, $K = 90$ and $K = 78$ for BILD, FAZ and SZ, respectively. Thus, we chose 80, 90 and 78 as the optimal topic numbers for the final training, again using 50 iterations.

Yet, the results of the topic modelling approach post two major problems for our aim of detecting and comparing frame differences between the newspapers: First, the resulting C_v scores with the optimized K values are at a rather low level (BILD: $C_v = 0.544$, FAZ: $C_v = 0.471$, SZ: $C_v = 0.424$). A manual evaluation of the most dominant words in each resulting topic also suggests a high degree of overlap between topics, as illustrated in Table 2. Second, the high number of K also considerably complicates the human interpretation of the overall topic differences between the newspapers, making it hard to use the results to ultimately inform further political science studies on framing differences between the publications.

A possible explanation for the poor performance of topic modelling is that the degree of vocabulary homogeneity among the articles in our dataset is high, since all articles focus thematically on issues related to refugees and migration. In a closer manual check of the dataset and the topic modelling results, we found that many words appear in different sub-topics due to their high relevance to the overall topic of refugees and migration, e.g., the keywords *Syrien* ('Syria'), *Land* ('country') and *Zahl* ('number') can either appear in discussions of refugee allocation policies or in reports about security at the Eastern Mediterranean Route. This "stop word-resembling" behavior of such words may confuse the topic modelling algorithm. However, eliminating such words would lead to a loss of information in the results since they, unlike the real stop words, bear highly relevant information for the context of refugees and migration. A more elaborate inspection of the reasons for the poor performance of topic modelling and a comparison of model performance on corpora with different degrees of vocabulary homogeneity are yet beyond the scope of the current paper and we will leave them for future work.

5 Experiment 2: Detecting Frames Using *word2vec* and Framing Vocabulary

Facing the low-quality results of the bottom-up, data-driven topic modelling method, in our second experiment we investigate a top-down, theory-

[2] The threshold of document frequency as 0.15 was defined experimentally. With the threshold set as 0.15, most of the high frequency items with little discriminative power for the topic of refugees and migration, such as *Mensch* ('People') and *Jahr* ('year'), can be excluded.

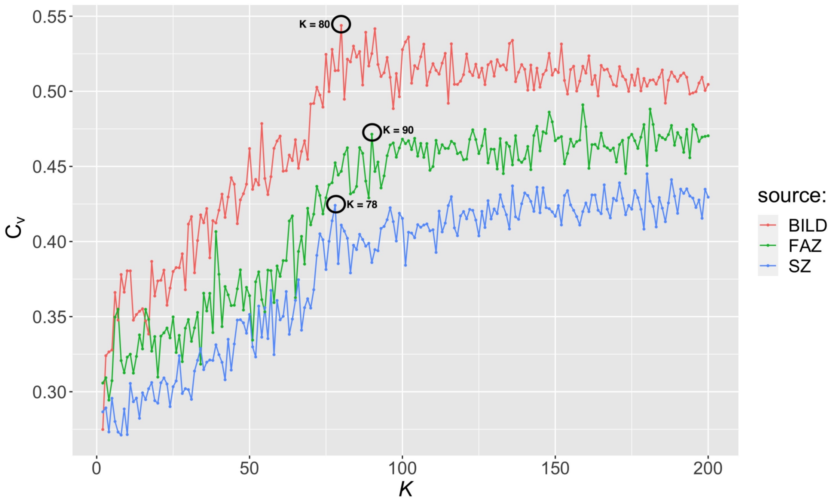


Figure 1: C_v coherence score of topic number $K \in [2, 200]$ in BILD, FAZ and SZ.

source	topic modelling results	remark
BILD	**Topic 21:** Vergewaltigung (rape), DNA (DNA), Abschiebepraxis (deportation practice), Feuerwehrmann (firefighter), Komplize (accomplice), Altena (Altena), Benzin (gasoline), Baden_Württemberg (Baden Württemberg), wegen_versuchtem_Mord (because of attempted murder), N. (N.) **Topic 23:** Jugendliche (youths), Mitarbeiterin (employee), Landkreistag (county council), Angreifer (attacker), Sexualdelikt (sexual offense), Schuss (shot), schwer_verletzt (heavily injured), Organisation_pro_Asyl (organization 'Pro Asyl'), Messer (knife), Polizei (police)	Both topics are about criminality and violence. Ideally, they should be aggregated to one topic.
FAZ	**Topic 77:** Griechenland (Greece), EU (EU), mehr (more), Million_Euro (million Euro), Land (country), Band (band), Europa (Europe), Türkei (Turkey), Integration (integration), Kreis (district) **Topic 80:** Türkei (Turkey), EU (EU), Griechenland (Greece), Ankara (Ankara), Europa (Europe), Brüssel (Brussels), türkisch (Turkish), EU_Staat (EU country), Flüchtlingskrise (refugee crisis), Erdoğan (Erdoğan)	Both topics are about the "refugee crisis" in term of the Eastern Mediterranean route of refugees and the EU.
SZ	**Topic 49:** Merkel (Merkel), Seehofer (Seehofer), Kanzlerin (chancellor), CDU (CDU), CSU (CSU), Flüchtlingspolitik (refugee policy), Partei (party), Union (union), AfD (AfD), Land (country) **Topic 61:** SPD (SPD), Bund (federation), Berlin (Berlin), Deutschland (Germany), Seehofer (Seehofer), Bundesregierung (federal parliament), Land (country), fordern (demand), mehr (more), neu (new)	Both topics are about domestic refugee policies and party competition.

Table 2: Overlapping topics in the results of topic modelling. The 10 most dominant items of each topic are listed.

driven method. We firstly deductively compiled a framing schema specifically tailored to the issue "refugees and migration" along which we can thematically classify and sort given frames in our data. Secondly, we created framing vocabulary lists for each category of our framing schema to further explore frame differences between newspapers that cannot be detected via topic modelling. This method is inspired by the observation and empirical verification in earlier studies that framing in news is to a large extent a keyword-driven phenomenon (Johnson et al., 2017; Field et al., 2018; Akyürek et al., 2020).

5.1 Creating the *Refugees and Migration Framing Schema*

Our Refugees and Migration Framing Schema is based on two theoretical works: 1) the general categorization of arguments by Habermas (1991), and 2) the extensive frame schema developed by Boydstun et al. (2014). We decided against creating a completely new framing schema in an inductive fashion (this is done by, amongst others, Helbling, 2014) for two reasons: First, the work of Habermas (1991), rooted in philosophical theory, exhaustively distinguishes all types of arguments that can justify actions (in our case these "actions" are attitudes towards refugees; see also Helbling, 2014 and Sjursen, 2002). He distinguishes between *identity-related*, *moral-universal* and *utilitarian* arguments. By applying his theory, we arrange for *all possible* kinds of arguments. Second, building on Boydstun et al. (2014) allows us to benefit off an already well-established and empirically verified frame schema. This schema is – unlike other published framing schemata such as Baumgartner et al. (2008) and Iyengar (1994) – designed to focus not only on a single issue, but includes very general, high-level issue dimensions of frames, beneath which more issue-specific categorizations can be specified. It therefore provides a comprehensive fit to a part of the general categorization by Habermas (1991). However, because the schema by Boydstun et al. (2014) is originally tailored towards coding and differentiating enacted $policies$, it can only provide a detailed and meaningful differentiation of frames in the category of *utilitarian* arguments in Habermas (1991). For our final Refugees and Migration Framing Schema, we therefore innovatively compiled the two theoretical works to incorporate the issue-related, scientifically evaluated breadth of the work by Boydstun et al. (2014), while providing for additional relevant categories presented by Habermas (1991). The resulting schema is elaborated in Table 3 (see columns *category* and *description*).

5.2 Creating the *Refugees and Migration Framing Vocabulary*

For each of the frame categories in our *Refugees and Migration Framing Schema*, we created one vocabulary list containing informative keywords for that category. The following two sources are utilized for constructing our *Refugees and Migration Framing Vocabulary*:

1) **Seed vocabularies by domain experts + GermaNet**: With an explorative reading of a small part of articles from our corpus, 5 domain experts (graduate students of political science) listed up words and phrases that they found highly relevant for each frame category in our schema. These seed vocabulary lists were then expanded by synonyms of each item, found using GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010).

2) ***DEbateNet-mig15* corpus**: The DEbateNet-mig15 corpus (Lapesa et al., 2020) is, to the best of our knowledge, the only annotated corpus of news on refugees and migration in German language. DEbateNet-mig15 contains 3,442 text passages from the German newspaper *Die Tageszeitung* (TAZ) in 2015 that are annotated as *claims* (i.e., statements made by political actors). The annotation was carried out using an ad-hoc annotation schema with eight high-level categories inductively developed by the authors.

We are aware that the *claims* annotated in DEbateNet-mig15 are by definition not equal to *frames*: While claims are strictly action-related, frames emphasize a certain aspect of an issue, whether action related or static. We also admit that a certain bias of word usage cannot be ruled out as DEbateNet-mig15 only contains data from the left-leaning TAZ. Nevertheless, DEbateNet-mig15 qualifies as an immediate base for the expansion of our *Refugees and Migration Framing Vocabulary* for two reasons: First, though claims per se differ from frames, the categorization of claims in DEbateNet-mig15 resembles frames to a large extent, i.e., claims are categorized based on the aspect(s) they emphasize. Second, the data of DEbateNet-mig15, as mentioned above, is in German language and arises from the same political issue as the one under investigation in our study. Considering these two reasons, we opted out of extracting vocabularies from corpora that are directly annotated with frames but are from different political backgrounds and/or in different languages, such as the MFC or the Gun Violence Frame Corpus (Liu et al., 2019).

For each of the eight high-level categories C in DEbateNet-mig15, we extracted the top 200 words w with the highest *pointwise mutual information* (PMI; Church and Hanks, 1990) to C:

$$PMI(C,w) \equiv log\frac{P(C,w)}{P(C)P(w)} = log\frac{P(w|C)}{P(w)} \tag{1}$$

Since the annotation schema of DEbateNet-mig15 diverges from our *Refugees and Migration Framing Schema* - although some of their categories are either identical to or are a subset of our categories - we re-sorted the extracted words into the suitable categories in our schema.

After merging the vocabulary lists obtained from the two sources above, a manual evaluation of the lists was conducted. In the evaluation, items that are too general and thus non-informative for detecting specific frame categories (e.g., *Einwanderung* ‘migration’, *wenigstens* ‘at least’) were omitted. Note that some items appear in more than one vocabulary list since they are highly relevant for multiple frame categories, e.g., *Fachkräfteeinwanderung* (‘skilled employee migration’) is a keyword for both economy frames and policy frames. Exemplary keywords for each frame category are given in Table 3 (see column *exemplary keywords*).

category	description: frames...	exemplary keywords
economy	... related to jobs, education, financial issues, etc., incl. *human resources frames*, *material resources frames*	Armutsflüchtling (poverty refugees), Arbeitskräftemangel (labor shortage) Ausbildung (training)
identity	... regarding group membership and individual senses of belonging, incl. *nationalism frames*, *cultural identity frames*	Herkunftsland (country of origin), Muslim (Muslim), rechtsextrem (right-wing extreme)
legal	... related to legal questions, incl. *jurisprudence frames*, *law frames*	Rechtsanspruch (legal entitlement), Bleibeperspektive (perspective to stay), Asylrecht (asylum right)
morality	... concerning ethics and moral concepts, incl. *humanitarianism frames*, *fairness and equality frames*	Menschenwürde (human dignity), Willkommenskultur (welcoming culture), solidarisch (showing solidarity)
policy	... related to concrete policies enacted by government, incl. *national policy frames*, *international policy frames*	Visum (visa), Richtlinie (guideline), Flüchtlingsquote (refugee quota)
politics	... regarding political proceedings and party competition	Asylstreit (Asylum-dispute), GroKo (grand coalition), Opposition (opposition)
public opinion	... on public attitudes and moods	Demonstration (demonstration), Meinungsmache (propaganda), Öffentliches Interesse (public interest)
security	... on violence and safety related issues, incl. *national security frames*, *terrorism frames* and *crime frames*	Anschlag (assault), Verbrechensrate (crime rate), Schlepperbande (human trafficking ring)
welfare	... on questions of benefit provision, incl. *health care frames*, *welfare benefit frames*	Sozialhilfe (social care), Hartz-IV (Hartz-IV), Versicherung (insurance)

Table 3: *Refugees and Migration Framing Schema* and corresponding example keywords to each category extracted with methods described in Section 5.2.

5.3 Mention Rate of Frames

As a first analysis using our *Refugees and Migration Framing Vocabulary*, we computed the *mention rate* of each frame in different newspapers. We represent a frame F as the list of extracted keywords $\{w_1, w_2, ..., w_k\}$ (as described in Section 5.2) of F, and the mention rate of F in a certain newspaper N as the cumulative frequency of $\{w_1, w_2, ..., w_k\}$:

$$mention\_rate_N(F) = \frac{\sum_{i=1}^{k} count_N(w_i)}{count_N(allwords)} \quad (2)$$

Figure 2 shows the mention rates of the frames in articles from all years between 2014-2018 in BILD, FAZ and SZ. To examine whether the mention rate differences between the newspapers are statistically significant, we applied a Kruskal-Wallis test to each frame. The Kruskal-Wallis test - a non-parametric variant of a variance analysis test (ANOVA) - is chosen because the mention rate values in single articles do not follow a normal distribution. A post-hoc Wilcoxon rank sum test was also conducted to understand pairwise differences between the newspapers.

Test results given in Table 4 indicate that the mention rate differences of all frames are statistically significant, except for the pairwise differences of the *Legal Frame*, *Politics Frame* and *Public Opinion Frame* occurrences between FAZ and SZ. As shown in Figure 2, the *Security Frame* shows the most striking difference, with the mention rate in BILD being considerably higher as compared to FAZ and SZ. Moreover, a large difference can be observed in *Economy Frame* occurrences, with FAZ showing the highest mention rate. The *Policy Frame* shows a higher mention rate in FAZ and SZ, which is expected given the tabloid-nature of BILD: BILD tends to produce sensational and shorter articles instead of in-depth discussions about intricacies of concrete refugee policies. These are instead more easily found in broadsheet newspapers. Finally, the *Morality Frame*, which includes mentions of moral ideas and concepts that tend to be more associated with a liberal, refugee-friendly discourse, is found to be mentioned more in FAZ and SZ.

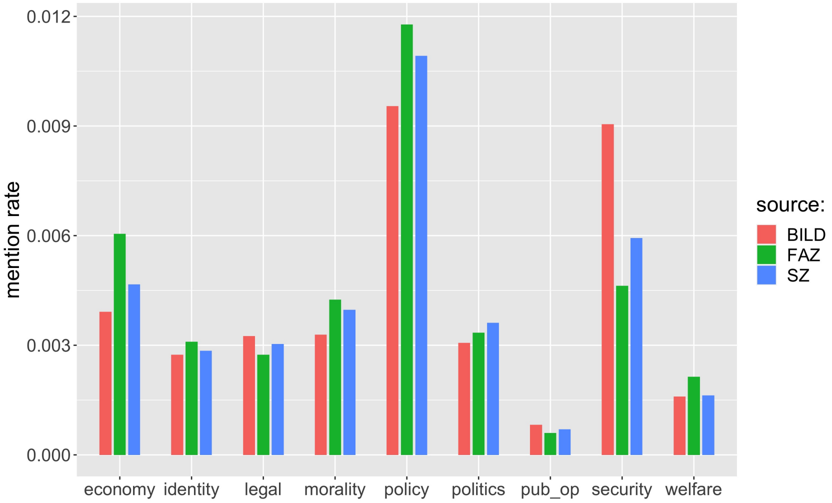


Figure 2: Mention rates of different frames in articles from 2014-2018 in BILD, FAZ and SZ.

	Kruskal-Wallis test		**Wilcoxon rank sum test (with Bonferroni adjusted *p*-values)**		
frame category	χ^2	p	BILD vs. FAZ	BILD vs. SZ	FAZ vs. SZ
economy	782.09	<2.2e-16	<2e-16	0.00016	<2e-16
identity	359.29	<2.2e-16	<2e-16	<2e-16	9.5e-08
legal	43.816	3.058e-10	3.3e-07	1.1e-07	1^{ns}
morality	775.02	<2.2e-16	<2e-16	<2e-16	<5.2e-14
policy	600.83	<2.2e-16	<2e-16	<2e-16	6.2e-09
politics	627.47	<2.2e-16	<2e-16	<2e-16	1^{ns}
public opinion	21.838	1.811e-05	5.9e-05	0.0031	1^{ns}
security	442.61	<2.2e-16	<2e-16	<2e-16	<2e-16
welfare	560.77	<2.2e-16	<2e-16	<4.3e-07	2e-16

Table 4: Kruskal-Wallis test and post-hoc Wilcoxon rank sum test of mention rate differences of each frame category in BILD, FAZ and SZ. (ns = not significant)

5.4 Semantic Similarity

Though some first intriguing frame differences can be observed by measuring the mention rate, this rather coarse metric is unable to distinguish the more subtle attitudinal differences associated to certain frames. For instance, the keywords *Fachkräftemangel* (‘shortage of skilled employees’) and *Wirtschaftsflüchtlinge* (‘economic refugees’) both belong to the *Economy Frame*. However, *Fachkräftemangel* in the context of refugees and migration conveys the migration-friendly attitude that skilled employees, and thus the migration of skilled employees, are sought after by the domestic economy. *Wirtschaftsflüchtlinge*, on the other hand, connotes a denunciation of refugees as exploiters of the social system and as (alleged) asylum abusers, because they did not flee for “real” political reasons (Bade, 2015; Wodak, 2015).

We apply word embedding to investigate such differences in greater depth. For each newspaper, we trained a 300-dimensional word2vec model. Before the training, all articles were tokenized and lemmatized using *Stanza*, and all stop words, numbers, punctuation marks and URLs were removed. To quantify how different newspapers portray refugees and the event “refugee crisis”, we use a `refugee_centroid`, which is computed as the average embedding of all *refugee-keywords* mentioned in Section 3. For each frame-specific vocabulary list, we rank items in the list by their cosine similarity to the `refugee_centroid`. Such a measurement allows us to find out which frame-specific keywords are collocated closer to the *refugee-keywords* in which newspaper, and thus gain insight on the fine-grained semantic differences in the discourse of the “refugee crisis” in different newspapers.

We inspect the top ten words with the highest cosine similarities to the `refugee_centroid` in the four frames we mentioned above that show the largest differences in mention rate, i.e., the *Security*, *Economy*, *Policy* and *Morality Frame*. Table 5 depicts the top ten keywords per frame category per newspaper. In all four frame categories intriguing differences can be observed:

Security Frame The highest semantic contrast is found in the keywords of the *Security Frame*. Whereas the item *Minderjährige* ('underage persons') has a high rank in all three newspapers - indicating an increased salience of reporting on the security of underage refugees - seven out of the top ten most similar items to the `refugee_centroid` in BILD are either related to criminality (e.g., *Delikt* 'offense', *Straftäter* 'perpetrator') or religious extremism (*Dschihad* 'Jihad', *Islamist* 'Islamist'). This implies a strong semantic association of refugees to threats to domestic security in BILD. For SZ, seven out of the top ten items are related to the security of refugees on the migration route or in their country of origin (i.e., *Rettungsmission* 'rescue mission', *Schlepper* 'human trafficker', *Bürgerkrieg* 'civil war'), rendering refugees as particularly threatened and thus in need of humanitarian aid. FAZ, finally, covers a middle ground between BILD and SZ with items both on crime (e.g., *Straftat* 'crime', *Kriminalitätsrate* 'crime rate') and on refugee related security issues, such as on the migration route (*Küstenwache* 'coast guard') or in the country of origin (*Bürgerkrieg* 'civil war').

Economy Frame Among the keywords of the *Economy Frame*, *Wirtschaftsflüchtling* ('economic refugee') is among the top ten similar words to `refugee_centroid` in the two right-leaning newspapers BILD and FAZ. For the left-leaning SZ, however, it only ranks as the 25th of all keywords of the *Economy Frame* (not shown in the table). Although the different ranks of keywords cannot be compared in absolute terms between newspapers, the lower rank of *Wirtschaftsflüchtling* in SZ indicates a reluctance to reduce refugees to having fled for economic reasons. Indeed, among the top ten most similar items for SZ, focus appears to lie on measures to support refugees to find jobs (i.e., *Berufsqualifikation* 'vocational qualification', *Ausbildung* 'training'). Also, *Wohnung* ('lodging') is one of the top ten items in this frame category only in SZ. Regarding the other two newspapers, items for BILD are related to integration (i.e., *Integrationskurs* 'integration course', *Deutschkurs* 'German course') and education (i.e., *Bildungsniveau* 'level of education', *Studium* 'academic studies'), opening up additional subject dimensions of cultural diversity and (educational) merit. Important items in FAZ, finally, are even more focused on merit with top ten items including *Fachkraft* ('skilled employee') and *Fachkräfteeinwanderung* ('skilled employee migration'). These results are not surprising because the FAZ is known for its economic focus.

Policy Frame Given that the mention rate of *Policy Frame* is the highest of all frames within each of the three newspapers, and given that within the top ten items of the *Policy Frame* in all three newspapers items related to the asylum procedure (i.e., *Aufenthaltserlaubnis* 'residence permit', *Asylverfahren* 'asylum procedure', *Abschiebung* 'deportation') feature prominently, this topic appears to play an outstanding role in the overall medial discourse on refugees and migration. Apart from this, however, some other semantic nuances among the top *Policy Frame* items can be observed: While SZ, again, is the only newspaper focusing on the issue of accommodation (*Wohnung* 'lodging') and has a humanitarian policy item within its top ten items (*Rettungsmission* 'rescue mission'), top items for BILD, once more, include references to integration policies (i.e., *Deutschkurs* 'German course') and the controversial issue of welfare benefits (*Sozialhilfe* 'social care' and *Sozialleistung* 'social benefit'). For FAZ, items related to education (*Studium* 'academic studies', *Schulausbildung* 'school education') again add economically focused nuance.

Morality Frame For the top ten items of the *Morality Frame*, the trends and focuses of the previously discussed frame categories are continued: Top items for BILD include once more *Integrationskurs* ('integration course') and impacts on the economy and the welfare system (i.e., *Wirtschaftflüchtling* 'economic migrant', *Arbeitslosengeld* 'unemployment benefit'), and top items for FAZ are again focused both on the economic impact of refugees (i.e., *Armut* 'poverty') and on their merit (i.e., *Fachkräfteeinwanderung* 'skilled employee migration' and *Punktesystem* 'point system', a system that aims to identify skilled migrants with better chances of receiving working permits). Though also partially featured in the top ten items

frame	BILD	FAZ	SZ
security	Minderjährige (underage persons)	Minderjährige (underage persons)	Rettungsmission (rescue mission)
	Delikt (offense)	illegal (illegal)	Minderjährige (underage persons)
	Straftäter (perpetrator)	Bürgerkrieg (civil war)	Krieg (war)
	Dschihad (Jihad)	Küstenwache (coast guard)	Bürgerkrieg (civil war)
	Gewaltkriminalität (violent crime)	Straftat (crime)	illegal (illegal)
	Islamist (Islamist)	Kriminalitätsrate (crime rate)	minderjährig (underage)
	Bürgerkrieg (civil war)	Schiffsunglück (shipwreck)	Schlepper (human trafficker)
	Tatverdächtiger (suspect)	Schlepper (human trafficker)	Straftat (crime)
	Schiffsunglück (shipwreck)	Gefängnis (prison)	Schutzstatus (protection status)
	inhaftieren (imprison)	Gefängnisstrafe (imprisonment)	Schiffsunglück (shipwreck)
economy	Kredit (credit)	Wirtschaftsflüchtling (economic refugee)	Kosten (costs)
	Arbeitsvertrag (working contract)	Fachkraft (skilled employee)	Wohnung (lodging)
	Bildungsniveau (level of education)	Studium (academic studies)	Berufsqualifikation (vocational qualification)
	Integrationskurs (integration course)	Schulausbildung (school education)	Ausbildung (training)
	Anstellung (employment)	Arbeitsstelle (workplace)	erwerbstätig (employed)
	Wirtschaftsflüchtling (economic refugee)	Arbeitsvertrag (working contract)	Arbeitslosenquote (unemployment rate)
	Studium (academic studies)	Berufsausbildung (vocational training)	zahlen (pay)
	Deutschkurs (German course)	erwerbslos (unemployed)	Bildungsniveau (level of education)
	Berufsausbildung (vocational training)	arbeitslos (unemployed)	Bleibeperspektive (prospect of staying)
	Hilfsmittel (aid)	Fachkräfteeinwanderung (skilled employee migration)	qualifiziert (qualified)
policy	Visum (visa)	Aufenthaltserlaubnis (residence permit)	Rettungsmission (rescue mission)
	Aufenthaltserlaubnis (residence permit)	Visum (visa)	Abschiebung (deportation)
	Ausreise (departure)	Asylverfahren (asylum procedure)	Asylverfahren (asylum procedure)
	Integrationskurs (integration course)	Abschiebung (deportation)	Herkunftsland (country of origin)
	Sozialhilfe (social care)	Balkanroute (Balkan route)	Wohnung (lodging)
	einstufen (classify)	Ausreise (departure)	Sozialleistung (social benefit)
	Studium (academic studies)	Studium (academic studies)	Ausreise (departure)
	Abschiebung (deportation)	Herkunftsland (country of origin)	Aufenthaltserlaubnis (residence permit)
	Deutschkurs (German course)	Schulausbildung (school education)	Balkanroute (Balkan route)
	Sozialleistung (social benefit)	Aufenthaltsrecht (right of residence)	Bleibeperspektive (prospect of staying)
morality	Integrationskurs (integration course)	Wirtschaftsflüchtling (economic refugee)	Rettungsmission (rescue mission)
	Wirtschaftsflüchtling (economic refugee)	Fachkräfteeinwanderung (skilled employee migration)	Flüchtlingsversorgung (provisioning for refugees)
	Hartz IV (Hartz IV)	Wirtschaftskrise (economic crisis)	Quote (quota)
	Hilfsmittel (aid)	Integrationskurs (integration course)	Armut (poverty)
	Flüchtlingsversorgung (provisioning for refugees)	Quote (quota)	Seenotrettungsprogramm (sea rescue program)
	Arbeitslosengeld (unemployment benefit)	Armut (poverty)	Leistung (merit)
	menschenwürdig (humane)	Wirtschaftsmigrant (economic migrant)	Kontingent (quota)
	Wirtschaftsmigrant (economic migrant)	Punktesystem (point system)	gemeinnützig (non-profit)
	Armut (poverty)	Hartz IV (Hartz IV)	Wirtschaftsflüchtling (economic refugee)
	Ungleichheit (inequality)	menschenwürdig (humane)	Versorgung (provisioning)

Table 5: Top ten most similar items to `refugee_centroid` within the *Security*, *Economy*, *Policy* and *Morality Frames* in BILD, FAZ and SZ.

for this frame category in BILD, SZ's focus on humanitarian issues (i.e., *Rettungsmission* 'rescue mission', *Flüchtlingsversorgung* 'provisioning for refugees' and *Seenotrettungsprogramm* 'sea rescue program') in the *Morality Frame* category is once more distinctive.

6 Conclusion and Outlook

Addressing the dilemma of many low-resource languages that there are no large-scale annotated datasets available for training and/or evaluating models of automated frame detection, we experimented with two NLP approaches for the data-driven exploration of frame differences which do not require building large-scale annotated corpora from scratch. Our first experiment with LDA-based topic modelling illustrated the difficulty of topic modelling for detecting topic preferences in a corpus where the vocabulary is highly homogeneous. Our second experiment with word2vec embeddings and the handcrafted *Refugees and Migration Framing Vocabulary* based on an expert-curated, comprehensive *Refugees and Migration Framing Schema*, however, yielded much more insightful and intelligible results.

Regarding the second experiment, it is worth mentioning that the quality of the handcrafted vocabulary lists has great impact on the quality of the results. In future work, we will therefor further improve the quality of our vocabulary lists by exploring the potential of more sophisticated keyword mining techniques, such as the method proposed by Jin et al. (2021) which ranks PMI-mined keywords by training interim classifiers.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) under Germany's Excellence Strategy - EXC-2035/1 - 390681379. We also greatly appreciate the valuable comments and suggestions from the anonymous reviewers of CPSS-2021.

References

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624.

Klaus J. Bade. 2015. Zur Karriere abschätziger Begriffe in der deutschen Asylpolitik. In *Aus Politik und Zeitgeschichte*, pages 3–8.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Frank R. Baumgartner, Suzanna de Boef, and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press, New York, Cambridge.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues. https://homes.cs.washington.edu/~nasmith/papers/boydstun+card+gross+resnik+smith.apsa14.pdf, last accessed 31 August 2021.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4479–4488.

Dallas Card, Amber Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of frames across issues. 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)*, pages 438–444.

Dallas Card, Justin H. Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1410–1420.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Deutschland.de. 2020. Überregionale Zeitungen in Deutschland. https://www.deutschland.de/de/topic/wissen/ueberregionale-zeitungen, last accessed 31 August 2021.

James N. Druckman. 2004. Political preference formation: Competition, deliberation, and the (ir)relevance of framing effects. *American Political Science Review*, pages 671–686.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Jürgen Habermas. 1991. *Erläuterungen zur Diskursethik*. Suhrkamp, Frankfurt am Main.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*, pages 9–15.

Marc Helbling. 2014. Framing immigration in Western Europe. *Journal of Ethnic and Migration Studies*, 40(1):21–41.

Verena Henrich and Erhard W. Hinrichs. 2010. GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.

Shanto Iyengar. 1994. *Is anyone responsible?: How television frames political issues*. University of Chicago Press, Chicago.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829*.

Yiping Jin, Akshay Bhatia, and Dittaya Wanvarie. 2021. Seed word selection for weakly-supervised text classification with unsupervised error estimation. *arXiv preprint arXiv:2104.09765*.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Modeling of political discourse framing on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 556–559.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66.

Gabriella Lapesa, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, and Sebastian Padó. 2020. DEbateNet-mig15: Tracing the 2015 immigration debate in Germany over time. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 919–927.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Nona Naderi and Graeme Hirst. 2016. Classifying frames at the sentence level in news articles. In *Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–9.

Thomas E. Nelson and Zoe M. Oxley. 1999. Issue framing effects on belief importance and opinion. *The journal of politics*, 61(4):1040–1067.

Pew Research Center. 2018. Datenblatt: Nachrichtenmedien und politische Haltungen in Deutschland. https://www.pewresearch.org/global/fact-sheet/datenblatt-nachrichtenmedien-und-politische-haltungen-in-deutschland/, last accessed 31 August 2021.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Helene Sjursen. 2002. Why expand?: The question of legitimacy and justification in the EU's enlargement policy. *Journal of Common Market Studies*, 40(3):491–513.

Rune Slothuus. 2008. More than weighting cognitive importance: A dual-process model of issue framing effects. *Political Psychology*, 29(1):1–28.

Ruth Wodak. 2015. *The politics of fear: What right-wing populist discourses mean*. Sage, London.

Share and Shout: Discovering Proto-Slogans in Online Political Communities

Irene Russo
ILC CNR
irene.russo@ilc.cnr.it

Gloria Comandini
Università degli Studi di Trento
gloria.comandini@unitn.it

Tommaso Caselli
University of Groningen, Groningen
t.caselli@rug.nl

Viviana Patti
Università degli Studi di Torino
patti@di.unito.it

Abstract

This paper proposes a methodology for investigating populism by analyzing proto-slogans, nominal utterances (NUs) typical of a political community on social media. We extracted more than 700.000 comments from the public Facebook pages of two Italian populist parties' leaders (Matteo Salvini and Luigi Di Maio) during the week preceding the 2019 European elections (i.e., from May 20 to May 26, 2019). These comments have been automatically clustered and manually annotated to find proto-slogans coming from the parties' supporters (bottom-up). We applied four layers of analysis: *Nominal Utterances* (NUs), a syntactic device widely used for slogans; *Slogans* for NUs with a slogan function; *Top-down/Bottom-up*, to recognize the slogans produced by the politicians and those produced by supporters; *Proto-slogans*, for NUs devoid of specific political content while expressing partisanship and support for the leaders.

1 Introduction

Social media have increasingly become arenas of mainstream political discourse. Platforms like Facebook and Twitter offer politicians venues to express their views, aggregate supporters and critics, and reinforce identities.

The vast amount of comments on political topics produced daily by users can be monitored and analyzed, using Natural Language Processing (NLP) tools to focus on relevant societal issues such as hate speech and fake news. However, apart from long comments that express more complex opinions, the majority of comments on social media are characterised by the synthetic expression of a point of view, often through the use of nominal utterances (NUs) (Comandini and Patti, 2019; Comandini et al., 2018). NUs, intended as syntactic declarative constructions built around a nonverbal head, can be part of a shared vocabulary used to express the in-group sense of cohesion and belonging on political pages and fora. For example, the NUs *Italia agli Italiani (Italy to Italians)* and *Porti chiusi (Closed harbors)* uniquely characterise one of the political communities investigated in this paper. Several of these recurrent NUs are slogans carefully created by the politicians' communication staff and used by supporters to reinforce the sense of belonging to a community. However, political slogans can also emerge from supporters' interactions on social media such as Facebook. They can become a trademark of a political community on other social media, such as Twitter. We define this process as proto-slogan generation.

Proto-slogans are semi-fixed linguistic expressions realised by NUs; they express a generic stance - positive or negative - toward a target. They emerge in online environments, in communities of people sharing the same perspectives or points of view.

In this paper, we study online political communities, extracting comments from the public Facebook pages of two populist Italian party leaders, Matteo Salvini for the Lega Nord (LN, Northern League) and Luigi Di Maio for the Movimento 5 Stelle (M5S, Five Stars Movement), during the week preceding the 2019 European elections (i.e., from May 20 to May 26, 2019). At that time, these two leaders were both covering the position of deputy prime minister in the so-called yellow-green government[1], and their parties were gaining consensus, with the LN winning the European elections. However, at the time of writing, both parties have lost consensus, and their leaders have changed their communication. This is mainly due to the new roles that these leaders are now covering (Salvini being part of the ruling majority with no position in

[1] https://en.wikipedia.org/wiki/Conte_I_Cabinet

the government and Di Maio being Foreign Minister). Besides this difference, the data analysed still represent a valuable tool to gain insights into the communication strategies of populist leaders and parties.

To do so, we propose a semi-supervised methodology that combines K-means clustering and manual annotation for the identification of proto-slogans. Additionally, we compare slogans extracted from Facebook with slogans retrieved on Twitter in different periods in order to distinguish between attested and emerging slogans. This comparison validates what has been extracted from Facebook's public pages on Twitter, where linguistic choices can be crucial in identifying communities if there are no other metadata available (such as the information that a user follows a politician).

Contributions The main contributions of this paper can be summarised as follows:

- an operational definition of proto-slogan as a key aspect in bottom-up populist communication on the web (Section 3);
- a methodology that combines unsupervised approaches (i.e., clustering) and manual annotation to identify political proto-slogans (Sections 5 and 6).

2 Populism on Social Media

Social media are fundamental to understand populist ideologies, which are mainly identified by their communication style (Kriesi, 2014; Aslanidis, 2016; Stanyer et al., 2016). In particular, Facebook seems to be the preferred social network of populist parties (Ernst et al., 2017). In this work, we will adopt a broad definition of populism, as a discourse based on the juxtaposition of two homogeneous and antagonistic groups: ”the good people” (the in-group) VS ”the bad elite/the foreigners” (the out-group) (Mudde, 2004; Rooduijn and Akkerman, 2017).

Charismatic leaders are particularly relevant for populist parties, and on Facebook, they are often more popular than the official party's page (Bobba, 2019). Thus, to study populist rhetoric, it is preferable to focus on the rhetoric of political party leaders, analysing how supporters react to it.

Populist leaders often adopt an emotional and straightforward communication style in order to be more persuasive and trigger a more emphatic response on social media (Oliver and Rahn, 2016). Indeed, it has been proved that emotionalized-style messages produced by Matteo Salvini on Facebook are more popular than his more neutral messages (Bobba, 2019).

Using these emotional messages and the direct connection with the public provided by social networks, populist leaders can forge close ties with their fan base, appearing more approachable (Jacobs and Spierings, 2016). Therefore, populist leaders can transform their Facebook pages into sheltered spaces for their fans, creating echo chambers in which aggressive tones can be cultivated (Engesser et al., 2017; Ernst et al., 2018). Together with the sense of belonging to the in-group due to the general resentment toward the out-group (Hameleers et al., 2019), this perceived intimacy with the leader creates a strong sense of being part of a homogeneous community, supportive to their leaders. In this way, populist leader's supporters may experience inter-group emotions, with each member experiencing emotions and taking action on behalf of the group (Smith and Mackie, 2008).

Previous computational linguistics analyzis of populism is scarce. Recently, Huguet Cabot et al. (2021) present a crowdsourcing annotated dataset for populist attitudes that collect comments about news on Reddit that mention a set of social groups (i.e., immigrants and Muslims) and classify attitudes toward them as supportive, critical, or discriminatory. In detecting the overall stance of comments, their analysis does not target exclusively populistic content. Instead, our work starts with the assumption that comments of the politicians' Facebook public pages are mainly supportive and sympathetic with the populist rhetoric. Thus, they constitute the ideal starting point for a stylistic investigation of populism.

3 Slogans and Proto-Slogans in Political Communication

Slogans are usually short, expressive, and assertive utterances, easy to memorize, and spread (Amălăncei et al., 2015). Slogans are defined linguistically by their pragmatic function: expressing an idea memorably and economically. They can have a broad range of syntactic forms and can be characterized by their use of figures of speech and rhetorical devices, such as metaphors (”Imagination at Work” from General Electric implies the metaphor ”General Electric is imagination”), parallel constructions (”Melts in your mouth, not in your hands” from M&M’s) or alliteration (”Don’t dream it. Drive it” from Jaguar) (Alnajjar and Toivonen, 2020).

The slogans that have received attention in previous work are those used in advertising. Political slogans are less studied, although they generally follow the same rules of advertising and have the same goal: change a person’s behavior (Ferrier, 2014). Furthermore, political slogans usually convey a strongly supportive or condemning message towards a person or a political program/action because voters are mainly influenced not by their conscious opinion on a politician’s program but by their feeling about a candidate or a party (Westen, 2007).

The procedure primarily used in studying slogans is top-down, concentrating on pre-existing slogans professionally crafted by politicians or companies. On the contrary, a bottom-up approach is much more complicated because it would require recognizing slogans in day-to-day communication focusing on their linguistic features. Top-down slogans clearly have a pragmatic function: they are created to persuade others. On the other hand, bottom-up slogans, emerging as linguistic devices shared by like-minded people, have a different function: they are used to structure and enhance the cohesion of online communities. Computational analyses of language used in online communities revealed that talking in a particular way on social media reinforces our networks and sense of belonging (McCulloch, 2019). For example, the use of written slang on Twitter depends on the number of times people saw the new word and if a member of their network uses it or not (Eisenstein et al., 2014). Also, adaptation at the stylistic level contributes to being well-received by a community. Tran and Ostendorf (2016) considers the reception of posted content focusing on the users with positive feedback, finding that stylistic features have discriminatory power for distinguishing between communities: the style is a better indicator of community identity than the topic. Interestingly, they found a positive correlation between the community reception to a contribution and the style similarity to that community. A correlation was not found for topic similarity.

Some of the linguistic devices used by the political electorate are slogan-like constructions that enhance cohesion inside the group. However, they are also simpler than real bottom-up political slogans. These slogan-like constructions do not convey a complex message but only the user’s stance (positive or negative) regarding a target, which is always explicitly mentioned. Thus, these slogan-like constructions usually appear as concise messages supporting or denigrating a politician or a group of people.

These short slogan-like bottom-up constructions that convey a basic message of support/denigration will be called proto-slogans, assuming that they are an embryonic form of a real slogan, since they convey a positive or negative stance, but not the more complex messages typical of slogans. In this paper, we elaborate on the notion of proto-slogan as a specific device to build cohesion in online communities, proposing a methodology to identify these items in social media.

Even if peculiar syntactic structures do not characterize slogans, proto-slogans are often realized syntactically as NUs (Comandini et al., 2018), also known as fragments without an overt antecedent (Merchant, 2005). NUs such as 1 and 2 are linguistic constructions without a verb in a finite form in their syntactic nucleus and are very common in informal spoken English (Merchant, 2005) and Italian (Cresti, 1998).

1. (After meeting Valentina at a social event, Katia says to her) Nice dress, by the way!

2. (When it begins to rain at the park, Monica says to her children) Presto, tutti a casa! [Quick, everybody home!]

NUs convey their content in a way that is expressive and informative, but also very economic (Ferrari, 2011b,a). It has been found that NUs are often used to convey hate speech against immigrants in the POP-HS-IT corpus, frequently taking the form of a verbless, hateful slogan (Comandini and Patti, 2019). Indeed, being without a verb in a finite form, NUs do not convey information about time, person, or aspect, creating messages similar to always valid maxims, mottoes, and, more importantly, slogans (Benveniste, 1990).

4 Data Extraction and Preprocessing

This paper focuses on the online audience of Matteo Salvini and Luigi Di Maio, the two leaders of widely recognized populist parties, LN and M5S.

In the selected period for data collection, between May 20 and May 26, 2019, covering the last week of the political campaign before the 2019 European elections, their communication was primarily conveyed through social media. We used Netvizz (Rieder, 2013), a tool that crawls data from Facebook, [2] to extract posts and comments from the Facebook public pages of the two politicians. We have excluded all posts written by the leaders and the replies to comments by other users, focusing our analysis on direct comments to the posts. Table 1 reports an overview of the extracted messages in terms of the average length of the posts published by the politicians (in tokens), the number of direct comments, and the average length of the comments (in tokens).

FB Page	**Avg. Post Length**	**Avg. Comm. Length**
Salvini	36.86±22.27	11.66±17.91
Di Maio	79.91±114.05	17.73±34.19

Table 1: Overview of the collected data from Salvini's and Di Maio Facebook pages between May 20th-26th, 2019.

As Table 1 shows, there are similarities and differences between the two groups. In both cases, we observe that the average length of the users' comments tends to be shorter than that of the posts published by the politicians. At the same time, it seems that the two groups of users (and ideally the level of the interactions with their leaders) tend to differ, with users on Salvini's page producing shorter messages than those on Di Maio's.

Since short comments are often verbless, we focus on NUs as syntactic declarative constructions built around a nonverbal head, framing them as the minimal unit of meaning in online communication. We developed a preprocessing procedure to find out all NUs contained in the comments.

Each comment has been preprocessed according to the following steps:

- its content has been sentence-splitted with NLTK (Bird et al., 2009);
- its content has been PoS-tagged with TreeTagger (Schmid, 1995);
- sentences that contain a verb in the finite form have been filtered out, to include in the final dataset only the potential nominal utterances;
- sentences containing proper nouns other than *Matteo* , *Salvini*, *Luigi*, and *Di Maio* have been filtered out to exclude comments mentioning Facebook users.

FB Page	**Comments**	**Eligible NUs**
Salvini	565,411	201,179
Di Maio	135,022	42,064

Table 2: Eligible NUs after preprocessing

The dimensions of the dataset before and after the preprocessing steps are reported in table 2. Preprocessed data have been the input for the clustering algorithm based on semantic similarity.

5 Aggregating Data Through Clustering

The amount of data after preprocessing, more than 240k comments (see Table 2 for details), is such that a manual exploration is not feasible. We thus decided to aggregate messages using clustering and perform manual annotation on aggregated data.

Our approach is based on K-means clustering. Such an approach has advantages and disadvantages for our task and, most importantly, our data. Results from K-means are easy to interpret and can be refined by manual inspection. At the same time, we are aware that K-means is not the best solution in an exploratory task, such as ours, where the number of clusters is not known, and it can hardly be assumed *a priori*. Using known estimate methods such as the Elbow curve does not represent a solution either in this case. We have addressed this issue by empirically validating the clusters of different sizes by using a sample of the data of 40k messages from Salvini's comments.

Each message representing an eligible NU has been converted as a 300-dimensions vector using FastText (Bojanowski et al., 2016). We then computed the pairwise cosine similarity scores between vectorized messages. This results in a N by N matrix of similarity scores. Similarity scores below 0.6 were excluded and replaced with zeros. The matrix has been used as input to the K-means algorithm. [3] We experimented with generating three groups of clusters of different sizes: 100, 150, and 200. Although none of them would correspond to an ideal amount of clusters for the aggregation of users' messages, their sizes allow for an easy and

[2] Netvizz is no longer available because, from September 4, 2019, it has no more Page Public Content access.

[3] We used the K-means implementation available in the sci-kit learn Python library (Pedregosa et al., 2011).

quick manual exploration of the data and may provide quite a fine-grained level of analysis. For each group of clusters, we plotted their centroids and observed their distributions. Quite interestingly, we could not find distinguishing differences or remarkable patterns. We finally selected 150 clusters as an appropriate level of aggregation to be subsequently manually annotated. Finally, we clustered the comments from Salvini's page a daily basis creating eight sets of comments, while we aggregated those for Di Maio in three blocks.

6 Manual annotation

The list of centroids (1,650 in total) obtained as a result of *K*-means clustering has been manually annotated by two annotators, with four annotation layers. These annotation layers are performed sequentially, and each of them is essential to understand the frequencies of NUs with different functions in each community. Furthermore, since centroids have been obtained by semantic similarity, focusing on them is a way to avoid annotating all the comments (a task that is not feasible) or annotating a not representative sample.

The first layer identifies NUs, which can be annotated following Comandini and Patti (2019) guidelines with a good agreement (0.96 in terms of Cohen's Kappa). We considered hashtags formed by two or more words as a single noun for this task, even if they contained a verb in a finite form. Most of these verbal hashtags are not used as VPs, but as nominal elements, linking the post to an "existing collective practice" (Zappavigna, 2015). The clause is excluded from the annotation when a NU has a coordinate clause with a verb in a finite form. Verbs in a non-finite form (infinitive, gerund, and participle) can be included in a NU. The list below provides several examples of NUs retrieved in our dataset:

3. <NU> bella intervista complimenti </NU> [Nice interview congrats]

4. <NU> forza salvini </NU> non pensare a sti dementi [go Salvini don't think about these idiots]

5. <NU> denunciare e sospendere il magistrato </NU> [to report and to suspend the magistrate]

The second annotation layer recognized particular NUs with a slogan-like form, with a binary value (yes-no). As noticed in section 3, an utterance is a slogan because of its purpose. Labeling an utterance produced by an anonymous user as a slogan is not a trivial or straightforward task, even if it is pretty simple to recognize political slogans created by politicians. Inter-annotator agreement for this level is 0.65 in terms of Cohen's Kappa, showing that recognizing slogans is not trivial and involves some form of subjective interpretation. Below we report examples of slogans in our dataset:

6. <NU> L'Italia agli Italiani </NU> [Italy to Italians]

7. <NU> Orgogliosi della propria identità </NU> [proud of our identity]

8. <NU> Forza Salvini </NU> [Go Salvini]

The third layer has been applied only to those items previously annotated as slogans by each annotator, distinguishing between top-down and bottom-up slogans. Top-down slogans are created by the political leader or party, while fans spontaneously produce bottom-up slogans. Annotators reached a better agreement on this distinction (0.74 Cohen's Kappa). One example for each category is reported below:

9. <NU> Porti chiusi </NU> [Closed harbors] [top-down]

10. <NU> Forza capitano </NU> [Go captain] [bottom-up]

As illustrated by example 10, bottom-up slogan-like NUs tend to be semantically close to encouragements and cheers that characterize sports competitions. They generally do not convey complex meanings but endorse the leader's message; they are phatic expressions with a clear social function (Jacobson, 1960).

As Table 3 reports, these NUs are predominant in the annotated dataset. Not surprisingly, the set of top-down slogans annotated is smaller than the set of bottom-up slogans: politicians' staff produce few slogans to communicate the politician's message. On the other hand, supporters use a broader set of NUs.

The fourth level of annotation explicitly targets proto-slogans, with an inter-annotator agreement of 0.63: several slogan-like NUs (*in alto i cuori (lift up your hearts)*, *sempre e per sempre (forever and ever)*) are not proto-slogans because they are hapax

		Slogan-like	
Facebook Page	NUs	Top-down	Bottom-up
Salvini	926	22	204
Di Maio	369	5	57

Table 3: Distribution of annotated NUs

Source	Bottom-Up NUs	Proto-slogans
Salvini	196	102
Di Maio	57	25

Table 4: Proto-slogans after annotation

in the list of centroids and lack of specific content. We recognize as proto-slogans the following NUs:

11. <NU> via i ladroni </NU> [away the robbers]

12. <NU> m5s tutta la vita </NU> [m5s for the rest of my life]

In Table 4 we report how many NUs can be labeled as proto-slogans. Bottom-up NUs are proto-slogans when they contain the reference to shared discourse targets for a community.

Comparing the bottom-up slogans and proto-slogans produced by the users to those produced by the politicians, it is clear that Salvini uses these kinds of slogans very frequently, while Di Maio generally uses only top-down slogans. Salvini often used bottom-up slogans such as (*avanti tutta (full steam ahead!)*), which appears three times a week, and it is also frequently used by Salvini's followers in the comments, often preceded by a proto-slogan such as (*forza Matteo (go Matteo)*).

However, the slogans most used by Salvini, appearing at least once a day, are two proto-slogan, both with a positive stance towards Italy or Italians: (*prima l'Italia (Italy first)*) and (*prima gli Italiani (Italians first*). These proto-slogans are not used in the comments by Salvini's followers, unlike top-down slogans such as (*porti chiusi [closed harbors*). Thus, (*prima l'Italia/gli Italiani*), while it conveys a political stance and is used by a political leader, does not act as a top-down slogan. Therefore, we may suppose that these proto-slogans act like a turn in an ongoing dialogue between Salvini and his followers, both of them expressing their support to each other through proto-slogans: Salvini expresses a positive stance towards his followers, who in return express their support to him through proto-slogans such as (*forza Matteo (go Matteo)*). Salvini refers to his followers as ”Italians” using a very common populist strategy that identifies populist voters with ”the people” and, in this case, with the Italian population as a whole. In this way, Salvini identifies his electorate with the whole Italian population, giving the impression of a much larger voter base and giving his electorate the perception that they are the real Italians, while their opponents are not.

7 Facebook and Twitter Data Comparison

Slogan-like NUs are specific linguistic items for a political community if supporters use them. However, they display different frequency patterns over time, i.e., they emerge as more frequent in a specific period. Therefore, the relationship between the frequencies of bottom-up slogans on social media and proto-slogans needs a more complex investigation based on more data. In this paper, we propose a qualitative classification of slogan-like NUs complementary to the characterization of proto-slogans.

In order to investigate this aspect, after the extraction and annotation of nominal utterances from Facebook public pages, the list was searched on Twitter with the help of GetOldTweets3 python library in three different one-week time spans across 3 years (2019, 2018, 2017) [4] to identify three types of slogan-like NUs:

- Generic slogan-like NUs: nominal utterances whose content does not directly concern populism or specifically related to the leaders. They can not be proto-slogans;
- Attested slogan-like NUs: specific to populist messages concerning Di Maio and Salvini, some attested slogan-like NUs are frequently used, but their presence varies through different periods. They tend to be proto-slogans, especially if they are bottom-up;
- Episodic slogan-like NUs: these NUs are linked to a specific event or period. However, they could still emerge as attested NUs if their use continues beyond a specific period. More data are needed to decide if they are proto-slogans or not.

Table 5 presents three examples with their frequencies in the different periods.

[4] The exact periods for each collection round are: 2019-11-20/2019-11-27, 2018-11-20/2018-11-27, and 2017-11-20/2017-11-27).

		Period		
Example	**NU Type**	**2017**	**2018**	**2019**
sempre avanti *(always ahead)*	Generic	115	108	104
avanti capitano *(come on captain)*	Attested	4	45	30
#26maggiovotolega *(#26mayIvoteLega)*	Episodic	1	0	0

Table 5: Types of NUs on Twitter

The presence of slogan-like NUs varies also depending on their bottom-up or top-down nature. Facebook slogan-like NUs are mostly bottom-up and generally composed by encouragements to the party or, more often, to the leader himself. They usually display a very familiar and affectionate tone, referring to the leader by his first name. This behavior is coherent to the perceived intimacy of Facebook communication, which makes leaders seem more approachable.

On Twitter, top-down slogans are more productive (see examples in Table 5) and with longer lifespans, primarily if they are not referred to a specific event, being instead relevant in a more general way. Thus, top-down slogans usually are attested slogan-like NUs or episodic slogan-like NUs.

A top-down episodic slogan made for the European election like *#domenicavotolega* (#sundayIvoteLega) is well-attested several months later, probably because it is still relevant for the next Italian Regional elections, planned on a Sunday too. Similarly, the generic, encouraging hashtag *#iostoconsalvini* (#Istaywithsalvini), an attested slogan, has been productive in every period considered. On the contrary, the more specific and episodic *#26maggiovotolega* (#26mayIvoteLega) is significantly less used after the European elections. Twitter displays some of LN's and M5S's main leitmotifs: the slogan-like NUs *porti chiusi* (closed harbors) and *tutti a casa* (everybody home). In 2018, *porti chiusi* had been used often in answers to Matteo Salvini's tweets, while in 2019 appeared more frequently in free-standing tweets. *Porti chiusi* is an example of attested slogan-like NUs that is distinctive for a political community and can be used to address this community to criticize its members.

Bottom-up slogan-like NUs are generally present on Twitter, but they show some peculiar differences from those on Facebook. Firstly, particularly familiar generic slogan-like encouragements like *forza matteo* (go matteo), very frequent on Facebook, are rare on Twitter, and they never appear in answers to Matteo Salvini's tweets. The less informal *forza salvini* (go salvini), *avanti capitano* (come on captain) and *forza capitano* (go captain) are far more frequent. Still, while on Facebook, they were placed inside the private echo chamber of the leader's page. They do not appear in answers to Matteo Salvini's tweets on Twitter, but they are characteristic of independent tweets. Most of the bottom-up generic slogan-like NUs, like *noi tutti con te* (all of us with you), are not attested on Twitter, but there are a few notable exceptions, such as *avanti tutta* (full steam ahead), *sempre avanti* (always forward) of *vergogna* (shame).

However, this investigation is still preliminary since it has not been possible to ensure that tweets with bottom-up generic slogan-like NUs, such as *forza capitano* (go captain), are unquestionably referred to LN. If the user explicitly mentions the politician, disambiguation is possible. Otherwise, the tweet could be used to support a football team.

8 Conclusion and Future Work

Political communication on social media can be analyzed with real data available on Twitter and Facebook public pages. This paper introduces the concept of proto-slogan as an economical device used to build and reinforce the in-group sense of belonging in online political communities.

We introduced a methodology for identifying NUs that are peculiar to a political community on social media. These NUs extracted from centroids, derived from the Facebook public page of Matteo Salvini and Luigi Di Maio, are often slogan-like. The political party or leader creates top-down slogans, and they are generally more linked to the party's program. Instead, the fans produce bottom-up slogans, which we define as proto-slogans, and they are usually less specific and more linked to informal encouragements.

Recognizing these slogan-like NUs makes it possible to recognize supporters of a specific populist political party, even when their messages are not otherwise contextually linked to it. Even if less specific, bottom-up slogan-like NUs are still recognizable on Twitter, they can uncover political support without explicit political content.

However, refining automatic recognition of NUs is still necessary since informal computer-mediated communication typically shows a substandard variety of Italian. For example, some verbs in the finite form may appear inside a NU, since they have a non-standard spelling.

Our analysis represents the first step toward identifying stylometric patterns in the populist electorate's informal writing on social media. We aim to characterize political affiliation in language even when explicit political themes are not mentioned. It would be advisable to remind that this kind of author profiling could have some ethical issues, but the final goal would not be monitoring opinions expressed on the web. Instead, we believe that public and open research on these topics would be helpful to show and make transparent for everyone what commercial systems - that often do not share their approaches with the scientific and the civil communities - can do with data.

References

K. Alnajjar and H. Toivonen. 2020. Computational generation of slogans. *Natural Language Engineering*, pages 1–33.

Brînduşa-Mariana Amălăncei, Cristina Cîrtiţă-Buzoianu, and Corina Daba-Buzioianu. 2015. Looking for the best slogan: an analysis of the slogans of the 2016 romanian parliament campaign. *Studies and Scientific Researches. Economics Edition*, 26:6–14.

Paris Aslanidis. 2016. Is populism an ideology? a refutation and a new perspective. *Political Studies*, 64:88–104.

Émile Benveniste. 1990. *Problemi di linguistica generale*. Mondadori, Milano, Italia.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Giuliano Bobba. 2019. Social media populism: features and 'likeability' of lega nrd communication on facebook. *European Political Science*, 18:11–23.

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

Gloria Comandini and Viviana Patti. 2019. An impossible dialogue! nominal utterances and populist rhetoric in an Italian twitter corpus of hate speech against immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, Florence, Italy. Association for Computational Linguistics.

Gloria Comandini, Manuela Speranza, and Bernardo Magnini. 2018. Effective communication without verbs? sure! identification of nominal utterances in italian social media texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Emanuela Cresti. 1998. Gli enunciati nominali. In M. T. Navarro, editor, *Italica matritensia: atti del IV convegno SILFI Società internazionale di linguistica e filologia italiana (Madrid, 27-29 giugno 1996)*, pages 171–191. Cesati, Firenze.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114.

Sven Engesser, Nayla Fawzy, and Anders Olof Larsson. 2017. Populist online communication: introduction to the special issue. *Information, Communication and Society*, 20:1279–1292.

Nicole Ernst, Sven Engesser, Florin Buchel, Sina Blassnig, and Frank Esser. 2017. Extreme parties and populism: an analysis of facebook and twitter across six countries. *Information Communication and Society*, 20:1347–1364.

Nicole Ernst, Frank Esser, Sina Blassnig, and Sven Engesser. 2018. Favorable opportunity structures for populist communication: Comparing different types of politicians and issues in social media, television and the press. *The International Journal of Press/Politics*, 24:165–188.

Angela Ferrari. 2011a. Enunciati nominali. *Enciclopedia dell'Italiano*. http://www.treccani.it/enciclopedia/enunciati-nominali_(Enciclopedia_dell'Italiano)/.

Angela Ferrari. 2011b. Stile nominale. *Enciclopedia dell'Italiano*. http://www.treccani.it/enciclopedia/stile-nominale_(Enciclopedia-dell'Italiano)/.

Adam Ferrier. 2014. *The Advertising Effect. How to Change Behaviour*. Oxford University Press University Press, South Melbourne.

Michael Hameleers, Carsten Reinemann, Desiree Schmuck, and Nayla Fawzi. 2019. The Persuasiveness of Populist Communication. Conceptualizing the Effects and Political Consequences of Populist Communication From a Social Identity Perspective. In C. Reinemann, J. Stanyer, T. Aalberg, F. Esser, and C. H. de Vreese, editors, *Communicating Populism. Comparing Actor Perceptions, Media Coverage, and Effects on Citiziens in Europe*, pages 143–167. Routledge, New York and London.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. Association for Computational Linguistics.

Kristof Jacobs and Niels Spierings. 2016. *Social media, parties, and political inequalities*. Palgrave Macmillan, Basingstoke.

Roman Jacobson. 1960. Linguistics and poetics. In T. Sebeok, editor, *Style in Language*, pages 350–377. M.I.T. Press.

Hanspeter Kriesi. 2014. The populist challenge. *West European Politics*, 37(2):361–378.

Gretchen McCulloch. 2019. *Because internet : understanding the new rules of language*. Riverhead Books, New York, NY.

Jason Merchant. 2005. Fragments and ellipsis. *Linguistics and Philosophy*, 27:661–738.

Cas Mudde. 2004. The Populist Zeitgeist. *Government and Opposition*, 39(4):541–563.

J. Eric Oliver and Wendy M. Rahn. 2016. Rise of the trumpenvolk: Populism in the 2016 election. *The ANNALS of the American Academy of Political and Social Science*, 667:189–206.

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

B. Rieder. 2013. Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 346–355, New York, NY, USA. ACM.

Matthijs Rooduijn and Tjitske Akkerman. 2017. Flank attacks: Populism and left-right radicalism in western europe. *Party Politics*, 23:193–204.

H. Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*.

Eliot R. Smith and Diane M. Mackie. 2008. Intergroup emotions. In L. F. Barrett M. Lewis, J. M. Haviland-Jones, editor, *Handbook of emotions*, pages 428–439. Guilford, New York.

James Stanyer, Susana Salgado, and Jaesper Strömbäck. 2016. Populist actors as communicators or political actors as populist communicators: Cross-national findings and perspectives. In T. Aalberg, F. Esser, C. Reinemann, J. Strömbäck, and C. H. de Vreese, editors, *Populist Political Communication in Europe*, pages 353–364. Routledge, New York.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Drew Westen. 2007. *The Political Brain. The Role of Emotion in Deciding the Fate of the Nation*. PublicAffairs, New York.

Michele Zappavigna. 2015. Searchable talk: the linguistic functions of hashtags. *Social Semiotics*, 25:274–291.

Twin Panel: The merits and challenges of interdisciplinary research

Nikoleta Yordanova, Goran Glavaš, Sebastian Haunss, Jonas Kuhn
University of Leiden, University of Mannheim, University of Bremen, University of Stuttgart
n.yordanova@fsw.leidenuniv.nl, goran@informatik.uni-mannheim.de
sebastian.haunss@uni-bremen.de, jonas.kuhn@ims.uni-stuttgart.de

Abstract

The workshop will host a panel with invited speakers from our different communities, with a focus on interdisciplinary methods and strategies supporting research at the intersection of Natural Language Processing and Political/Social science. In particular, we are interested in learning from previous experiences of interdisciplinary projects to gain a better understanding of what has made collaborations at the NLP/SocSci interface successful (e.g., strategies to optimize the conceptual/practical exchange between teams) as well as of the concrete problems encountered and about their solutions. Therefore, we will invite ”twin researchers” to the panel, i.e., pairs of researchers where each pair includes one researcher from political or social science and one with a background in CL/CS, who have worked together on a project. The twin researchers will talk about their collaboration and share their experiences with the audience. We hope that this will foster discussions and allow us to reflect on our different research practices, methods and tools, and will help to improve the communication between our fields.

Project I: Willingness and Capacity for EU Policy Action in Turbulent Times: Conflicts, Positions and Outcomes (EUINACTION)

Nikoleta Yordanova & Goran Glavaš
`https://www.euinaction.eu`

The European Union (EU) faces pressing demands to act in major policy areas amid public contestation of supranational governance. Our interdisciplinary project seeks to explain and facilitate responsive and effective policy reforms by increasing knowledge about the willingness and capacity for EU integration in specific policy areas. We study the conditions under which the EU institutions seek to increase or decrease EU policy competences, when their positions respond to public demands across and within member states, and under what conditions each institution manages to assert its position in the policy-making processes. We further investigate how the institutions' positions and capacity to steer the course of European integration across policy areas are reshaped by increased EU politicization and associated shifts in institutional identities, internal disunity and switch from formal political to informal technocratic procedures of policy-making. Public opinion surveys, party manifestos and speech data from European and national parliaments will serve to capture citizen, party and government preferences over the transfer of competences from the national to the EU level across policy areas. We will then examine under what conditions and to what extent these preferences determine the positions of EU institutions, policy proposals and adopted legislation with respect to the level of competence transfer to the EU, using cutting-edge methods for computational text analysis. These findings will serve to develop recommendations about innovation in policy and institutional design that can address pressing challenges and enjoy public acceptance in member states and among their citizens.

Project II: Modeling ARgumentation DYnamics in Political Discourse (MARDY)

Sebastian Haunss & Jonas Kuhn
`https://sites.google.com/view/mardy`

This interdisciplinary collaboration project involving Computational Linguistics, Machine Learning and Political Science has the aim of developing new computational models and methods for analyzing argumentation in political discourse – specifically capturing the dynamics of discursive exchanges on controversial issues over time. The goal is to develop tools to support analysis of the possible impact of arguments advanced by different political actors.

Twin Panel: The merits and challenges of interdisciplinary research

Nikoleta Yordanova, Goran Glavaš, Sebastian Haunss, Jonas Kuhn
University of Leiden, University of Mannheim, University of Bremen, University of Stuttgart
n.yordanova@fsw.leidenuniv.nl, goran@informatik.uni-mannheim.de
sebastian.haunss@uni-bremen.de, jonas.kuhn@ims.uni-stuttgart.de

Abstract

The workshop will host a panel with invited speakers from our different communities, with a focus on interdisciplinary methods and strategies supporting research at the intersection of Natural Language Processing and Political/Social science. In particular, we are interested in learning from previous experiences of interdisciplinary projects to gain a better understanding of what has made collaborations at the NLP/SocSci interface successful (e.g., strategies to optimize the conceptual/practical exchange between teams) as well as of the concrete problems encountered and about their solutions. Therefore, we will invite ”twin researchers” to the panel, i.e., pairs of researchers where each pair includes one researcher from political or social science and one with a background in CL/CS, who have worked together on a project. The twin researchers will talk about their collaboration and share their experiences with the audience. We hope that this will foster discussions and allow us to reflect on our different research practices, methods and tools, and will help to improve the communication between our fields.

Project I: Willingness and Capacity for EU Policy Action in Turbulent Times: Conflicts, Positions and Outcomes (EUINACTION)

Nikoleta Yordanova & Goran Glavaš
`https://www.euinaction.eu`

The European Union (EU) faces pressing demands to act in major policy areas amid public contestation of supranational governance. Our interdisciplinary project seeks to explain and facilitate responsive and effective policy reforms by increasing knowledge about the willingness and capacity for EU integration in specific policy areas. We study the conditions under which the EU institutions seek to increase or decrease EU policy competences, when their positions respond to public demands across and within member states, and under what conditions each institution manages to assert its position in the policy-making processes. We further investigate how the institutions' positions and capacity to steer the course of European integration across policy areas are reshaped by increased EU politicization and associated shifts in institutional identities, internal disunity and switch from formal political to informal technocratic procedures of policy-making. Public opinion surveys, party manifestos and speech data from European and national parliaments will serve to capture citizen, party and government preferences over the transfer of competences from the national to the EU level across policy areas. We will then examine under what conditions and to what extent these preferences determine the positions of EU institutions, policy proposals and adopted legislation with respect to the level of competence transfer to the EU, using cutting-edge methods for computational text analysis. These findings will serve to develop recommendations about innovation in policy and institutional design that can address pressing challenges and enjoy public acceptance in member states and among their citizens.

Project II: Modeling ARgumentation DYnamics in Political Discourse (MARDY)

Sebastian Haunss & Jonas Kuhn
`https://sites.google.com/view/mardy`

This interdisciplinary collaboration project involving Computational Linguistics, Machine Learning and Political Science has the aim of developing new computational models and methods for analyzing argumentation in political discourse – specifically capturing the dynamics of discursive exchanges on controversial issues over time. The goal is to develop tools to support analysis of the possible impact of arguments advanced by different political actors.

Application of the interactive Leipzig Corpus Miner as a generic research platform for the use in the social sciences

Christian Kahmann*, **Andreas Niekler***, **Gregor Wiedemann**†
*Natural Language Processing Department - University Leipzig
Augustusplatz 10, 04109 Leipzig
{kahmann, aniekler}@informatik.uni-leipzig.de

†Leibniz-Institut für Medienforschung - Hans-Bredow-Institut (HBI)
Rothenbaumchaussee 36, 20148 Hamburg
g.wiedemann@leibniz-hbi.de

Abstract

This article introduces to the *interactive Leipzig Corpus Miner* (iLCM) - a newly released, open-source software to perform automatic content analysis. Since the iLCM is based on the R-programming language, its generic text mining procedures provided via a user-friendly graphical user interface (GUI) can easily be extended using the integrated IDE RStudio-Server or numerous other interfaces in the tool. Furthermore, the iLCM offers various possibilities to use quantitative and qualitative research approaches in combination. Some of these possibilities will be presented in more detail in the following.

1 Introduction

The use of computational methods is becoming increasingly important in the social sciences and in its subdisciplines, such as communication science (van Atteveldt and Peng, 2018). This is mainly due to the rapidly growing amount of digital data. Especially for large textual datasets, automatic procedures are required since conventional content analysis involving steps of manual reading and interpretation is not feasible any longer. Automatic approaches to content analysis in textual data promise a much more efficient processing and allow for scaling with the constantly growing amount of available data. Furthermore, previous studies have shown that the application of automatic methods can lead to novel insights that would not have been possible to obtain with traditional methods alone (Wiedemann, 2013).

For researchers in the applying fields, coding their own analysis programs often is not a viable option due to lack of resources or expertise.[1] Instead, applied research usually relies on existing research software. On the one hand, this could be standalone software solutions designed for very specific analysis purposes (e.g. a word frequency analyzer). This restrains researchers to narrow study designs remaining within the limitations of the specific software. Any desired functionality deviating from this cannot be realized leading to a significant reduction in the method portfolio of the project. On the other hand, generic software solutions are available which provide a larger number of analysis tools and, thus, do not restrict users to a narrow methodological workflow. Instead, generic research tools enable the application of various methods which can be flexibly combined, and, therefore, be used in a wide range of projects. Moreover, this flexibility results from opportunities for adaptation and extension of the methodical approaches built into the generic software.

The *interactive Leipzig Corpus Miner* (iLCM) represents such a generic software solution for the use of text mining in the social sciences and humanities. In the following, we reflect on which features allow the tool to be customizable and extendable. Furthermore, Section 4 describes the possibilities of combining quantitative and qualitative research approaches in the iLCM. Based on this, Section 5 shows exemplarily which advantages result from the described features of the tool.

2 Related Work

Of course, the iLCM is not the only software solution available in the field of the application of text mining in the social sciences. Several other solutions exist. These include, for example, the various QDA software solutions such as *MAXQDA*[2], *Atlas.ti*[3] or *NVivo*[4] These have specialised in the

[1] Actually, more and more computational social scientists obtain coding skills. However, the development of complex research software remains a complicated process which usually requires trained software developers.

[2] https://www.maxqda.de/

[3] https://atlasti.com/

[4] https://www.nvivo.de/

process of qualitative data analysis of texts in the social sciences. These processes are excellently mapped in these software solutions. However, research questions that clearly deviate from this procedure cannot be mapped. Other more flexible tools include *RapidMiner* (Hofmann and Klinkenberg, 2013) and *KNIME* (Berthold et al., 2009). These offer a variety of different utilities, which can then be combined and put together to form a workflow. Both *RapidMiner* and *KNIME* offer a graphical user interface. In addition, there are command line-based libraries such as the R packages *quanteda* (Benoit et al., 2018) or *polmineR* (Blaette, 2020), which also provide numerous tools for the use of text mining in a social science context. However, these require a certain amount of prior knowledge in the use of command line editing tools. The iLCM, on the other hand, offers both a graphical user interface as well as the option of using a command line based environment. This makes it possible not only to adapt and expand the numerous analyses already available, but also to further optimise the entire application of the tool to suit one's own problem. Additionally, the iLCM offers numerous export options to support interoperability with other software solutions.

3 iLCM as a generic research tool

For a generic research software to support automatic content analysis, specific requirements must be met. In the following, we elaborate on four functional requirements together with the solutions as implemented in the iLCM.

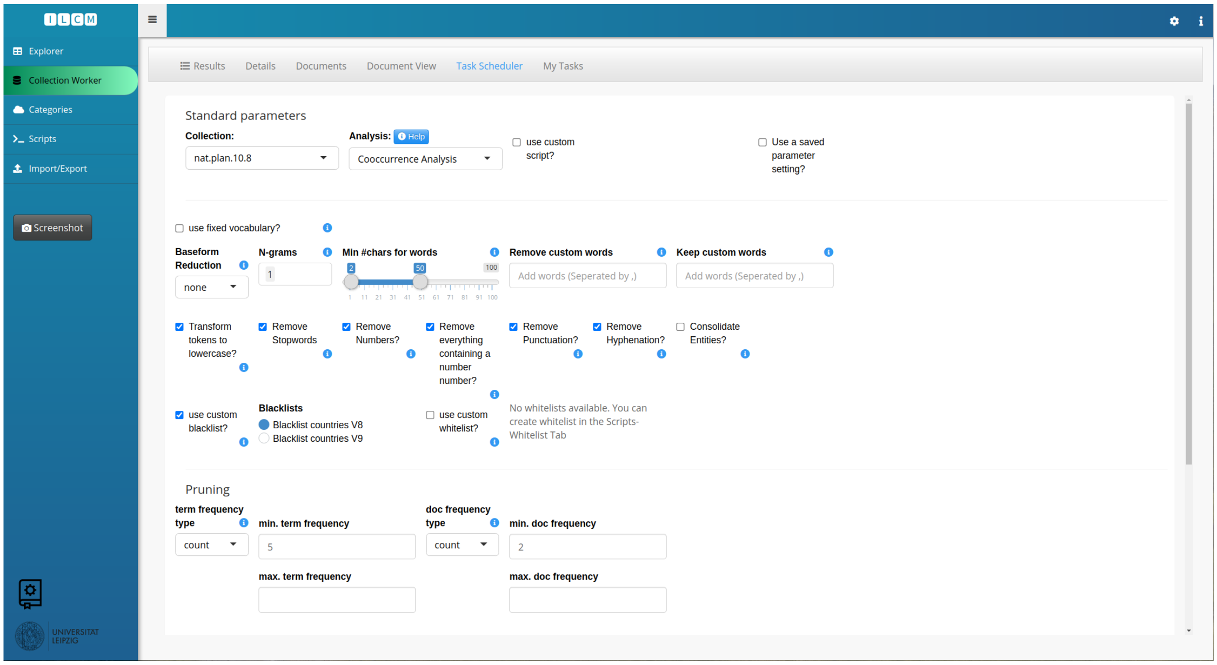
Analysis capabilities: Availability of a wide range of predefined functions: In order to be able to adequately address different kinds of research questions, it is necessary to provide as many predefined functions from the areas of text mining and machine learning as possible. In the iLCM, numerous procedures are implemented to this end. The iLCM supports multi-language document preprocessing, document retrieval and collection management, content deduplication, word frequency analysis, word co-occurrence analysis, time series analysis, topic models, category coding and annotation, supervised text classification (e.g. sentiment analysis), and more. Different methods can be combined with each other in flexible ways.

Adaptability: Predefined analysis capabilities often require research specific adaptions in both, either pre-processing or analysis of textual data. It is important to have a high tolerance for different text bases, so that various data sets, languages or metadata can be processed. In the iLCM a high degree of adaptability is guaranteed by the possibility to extensively parameterise each analysis step (see figure 1). Since internally every analysis method is implemented as a script written in the programming language R, it is possible to adapt the predefined methods directly within the tool, and to easily add the support for further languages.

Extensibility: If some of the required functions are not available in the tool, it should be possible to add these functions. In the iLCM, new scripts can be created within the iLCM script editor to add new or replace existing analysis capabilities. For instance, it is easy to add project-specific black- and whitelists of words for pre-processing steps. Further, it is also possible to implement additional analyses based on interim results in an associated IDE.[5] This design of the iLCM software allows researchers to apply the principles of Agile Development to carry out own implementations in a comparatively short time to realize an analysis workflow tailored to her/his individual requirements (Heyer et al., 2019). In detail, here Agile Development refers to the possibilities of building on the existing infrastructures of the iLCM to be able to answer one's own questions through independent implementations. These implementations can either be carried out separately from the iLCM or integrated into it. Due to the existence of numerous functions and an already existing IDE, which can fall back on various pre-installed key libraries, initial results can be achieved very quickly in the sense of a prototypical procedure. This in turn means that necessary adjustments to the code and the operationalisation of the problem can be recognised at an early stage and correspondingly implemented in an agile manner.

Data export: If it is not possible or desired to implement a research design fully within the framework provided by the iLCM, it may still be possible to map at least partial processes with the help of the tool. The result of these sub-processes can then be exported in standard formats such as CSV or the REFI-QDA standard (Evers et al., 2020) and

[5] Parallel to the iLCM GUI, an RStudio-Server instance is provided as an IDE that has access to the available data and results.



Standard parameters

Collection: nat.plan.10.8

Analysis: Cooccurrence Analysis

□ use custom script?

□ Use a saved parameter setting?

□ use fixed vocabulary?

Baseform Reduction: none

N-grams: 1

Min #chars for words: 2 – 50 (1–100)

Remove custom words: Add words (Seperated by ,)

Keep custom words: Add words (Seperated by ,)

☑ Transform tokens to lowercase?

☑ Remove Stopwords

☑ Remove Numbers?

☑ Remove everything containing a number number?

☑ Remove Punctuation?

☑ Remove Hyphenation?

□ Consolidate Entities?

☑ use custom blacklist?

Blacklists
● Blacklist countries V8
○ Blacklist countries V9

□ use custom whitelist?

No whitelists available. You can create whitelist in the Scripts-Whitelist Tab

Pruning

term frequency type: count

min. term frequency: 5

max. term frequency:

doc frequency type: count

min. doc frequency: 2

max. doc frequency:

Figure 1: Interface of the iLCM for parameterisation of an analysis; In the example shown, a cooccurrence analysis is carried out for which in detail unigrams are used, words consist of at least 2 and at most 50 characters, a specially created blacklist is used, stop words and numbers are removed and pruning is also carried out. Further settings are available. The flexible parameterisation of an analysis allows different text bases (in language and quantity) as well as research questions to be processed.

used in other software solutions like MAXQDA or Scripting environments like R and Python.

Validation: We have consistently paid attention to validation methods during implementation. In all methods for which a standard evaluation such as Precision, Recall, F1 or Sementic Coherence is described in the literature, an evaluation routine was built in to validate the results. This allows the specification and application of quality assurance processes for scientific publications and strengthens the confidence in automatic processes for textual content analysis.

The shown features, with a variety of already existing algorithms, which can be adapted to the respective research question, as well as the possibility to add further functions to the tool, allow the use of the tool in various fields of social sciences.

4 Combination of quantitative and qualitative approaches

The analysis capabilities of the iLCM allow for combinations of quantitative with rather qualitative (interpretive) research approaches to investigate large textual datasets. This combination of the two research paradigms can be achieved in different ways. In the following, we present three main hybrid research concepts and explain how they can be realized within the iLCM software.

Text classification: Qualitative Data Analysis (QDA) is a basic research method in the social sciences. Here, texts, resp. text segments are coded into different categories on the basis of a previously defined codebook. Then, distributions of codings in texts are used for further descriptive analysis. However, the manual coding of large amounts of text is time-consuming and costly. To support manual coding, computational methods such as supervised machine learning algorithms can be applied. These try to train a model which is able to predict the categories of the uncoded data based on a set of data manually coded by the researcher. To improve the performance of such a machine classifier efficiently, approaches of active learning (Settles, 2010) come into place. In a mutual interplay of manual and automatic procedures, it is thus possible to obtain a qualitatively based classification of data due to the researchers interpretive decisions during training data creation. At the same time, automatic coding allows for scaling the coding process to very large datasets ready for quantitative analysis of qualitative codings.

Topic model validation: The result of a topic model analysis (Blei et al., 2003) is described by a set of topics that represent probability distributions over the vocabulary. The documents in turn are reflected as distributions of these same topics. Topic models are based on the significant common occurrence of semantically related words. Ideally, this makes it possible to find word clusters that cover distinct semantic areas which can be interpreted as topics. If this is the case, further analyses can be carried out on the basis of these thematic clusters, for example in relation to existing metadata. However, the assessment of when a model captures interpretable topics sufficiently cannot be done automatically alone, but also requires qualitative steps. Knowing this, the iLCM has a built-in interface which allows the qualitative validation of topic model results. For this purpose the original documents are displayed. Depending on the selected topic, the words in these documents can be highlighted in colour to show which section of text is responsible for the assignment of a document to a certain topic (see Fig. 4). The researcher has the possibility to compare the word distributions found and to check their plausibility with his domain expert knowledge. In this way, possible sources of error during the import of the data, such as the presence of duplicate files or defective OCR, can also be detected and addressed in a subsequent step.

Thematic filtering: Simply put, the result of a topic model provides clusters of semantically coherent word probabilities and the association of these clusters with documents. To assign meaning to these clusters qualitative interpretation is required. This cannot be done by interpreting the word clusters alone (cp. the process of topic model validation above), but also by reading topic representative documents. For this purpose, the iLCM provides the possibility to select texts a topic model is based on according to its topic composition. Selected texts can be viewed with coloured highlightings to visualize the thematic affiliations of words. This makes it possible to understand which text passages are typical for a certain topic and, thus, enable a better understanding of what aspects a single topic is actually composed of. Once the topics have been interpreted, they can be linked to existing metadata in the tool in order to derive results and hypotheses from the model.

5 Exemplary study

An example application within the TRANSNORMS project[6] demonstrates some of the described analysis capabilities of the iLCM. The TRANSNORMS project seeks to examine the translation of global (political) norms to the local level. In this study, particularly the understanding of the norms on climate protection agreed on at the Climate Conference 2015 (COP 21) in Paris are examined. At COP 21, 195 countries concluded the first comprehensive and legally binding global climate protection agreement. As a result, all participating countries committed themselves to submit nationally determined contributions (NDC) in which they list their intentions to achieve the jointly agreed targets. These NDCs provide the data basis for the analysis.

For the analysis, the NDC texts were imported into the iLCM. At this point, the focus on a variable import interface was beneficial, which allows different document types such as .pdf, .doc or .csv and .xlsx to be uploaded and then interactively mapped to the data format of the iLCM in the graphical interface of the tool. The aim of the project was then to investigate to what extent unsupervised procedures are able to identify different thematic fields within the NDCs. Based on this, it was to be further investigated whether the distribution of the texts into the thematic fields would result in correlations with country-specific metadata. To answer these questions, the LDA topic model as provided in the iLCM was used. To not receive topics governed by geographic entities, it was necessary to remove the proper location names such as countries or cities from the documents. For this purpose a blacklist was created inside the iLCM building on the already existing named entity tags[7] as a list of candidates. Based on these pre-processing steps, a final model was then calculated. The topics found were then checked for validity (see Fig. 4) and subsequently interpreted. Subject domains such as: *Renewable Energy, Economic growth, Water-related vulnerability* and *UNFCCC collaboration* could be identified and labeled[8]. It was thus estab-

[6] https://www.transnorms.eu

[7] The named entity tags are assigned as part of the import of new data into the iLCM.

[8] The most relevant words of a topic as well as typical text passages (4) were used to define the labels. For the topic *international support*, for example, the most relevant words were: *financial, fund, funds, financing, required, support, needs, finance, capacity_building, investment, donors,...*

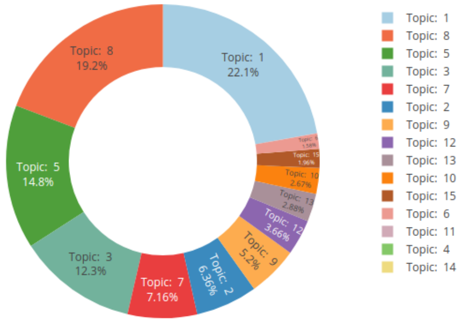


Figure 2: Topic distribution of the NDC of Gambia.

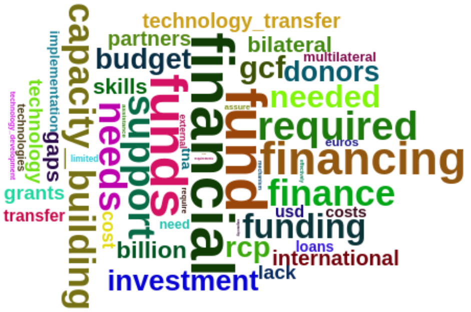
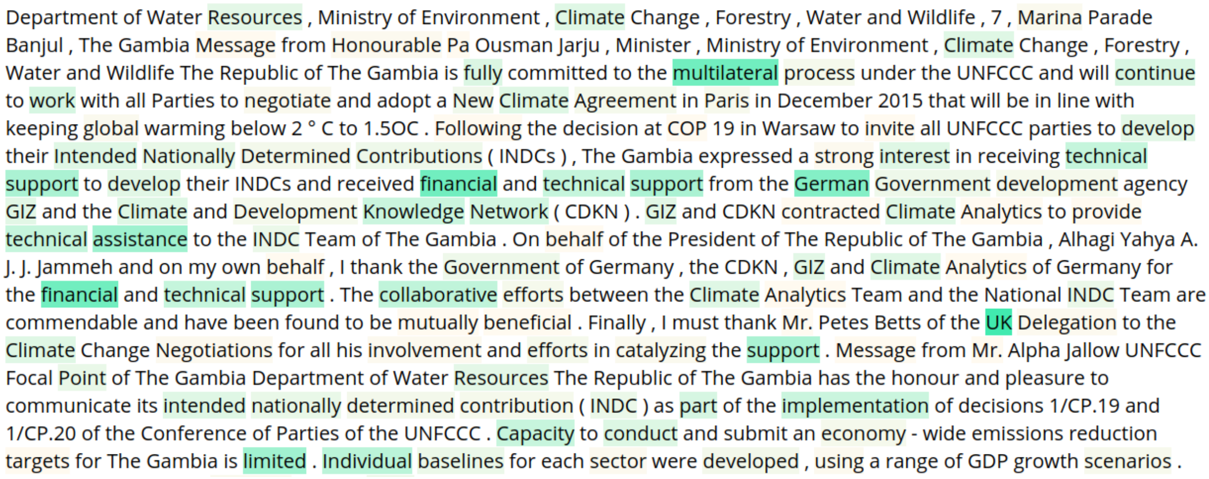


Figure 3: Wordcloud to display the most relevant words for Topic 8. This Topic has been labelled as *international support*.



Department of Water Resources , Ministry of Environment , Climate Change , Forestry , Water and Wildlife , 7 , Marina Parade Banjul , The Gambia Message from Honourable Pa Ousman Jarju , Minister , Ministry of Environment , Climate Change , Forestry , Water and Wildlife The Republic of The Gambia is fully committed to the multilateral process under the UNFCCC and will continue to work with all Parties to negotiate and adopt a New Climate Agreement in Paris in December 2015 that will be in line with keeping global warming below 2 ° C to 1.5OC . Following the decision at COP 19 in Warsaw to invite all UNFCCC parties to develop their Intended Nationally Determined Contributions (INDCs) , The Gambia expressed a strong interest in receiving technical support to develop their INDCs and received financial and technical support from the German Government development agency GIZ and the Climate and Development Knowledge Network (CDKN) . GIZ and CDKN contracted Climate Analytics to provide technical assistance to the INDC Team of The Gambia . On behalf of the President of The Republic of The Gambia , Alhagi Yahya A. J. J. Jammeh and on my own behalf , I thank the Government of Germany , the CDKN , GIZ and Climate Analytics of Germany for the financial and technical support . The collaborative efforts between the Climate Analytics Team and the National INDC Team are commendable and have been found to be mutually beneficial . Finally , I must thank Mr. Petes Betts of the UK Delegation to the Climate Change Negotiations for all his involvement and efforts in catalyzing the support . Message from Mr. Alpha Jallow UNFCCC Focal Point of The Gambia Department of Water Resources The Republic of The Gambia has the honour and pleasure to communicate its intended nationally determined contribution (INDC) as part of the implementation of decisions 1/CP.19 and 1/CP.20 of the Conference of Parties of the UNFCCC . Capacity to conduct and submit an economy - wide emissions reduction targets for The Gambia is limited . Individual baselines for each sector were developed , using a range of GDP growth scenarios .

Figure 4: Validation interface in which the text of the NDC of Gambia is displayed. In addition, the most relevant words for the selected Topic 8 are highlighted in colour. The visualisation makes it very easy to see that a) this section does indeed address a funding issue and b) the exact content focus here on the financial and technical support provided by GIZ and CDKN is evident.

lished that it is possible to identify thematic areas unsupervisedly in the NDCs. The issues found in this way were then examined using metadata analysis tools added to the iLCM specifically for this purpose. Among other things, it was found that there is a clear distinction between Annex-1 and Non-Annex countries. In some topics, however, more surprising results were also found. For example, for the topic titled *Water-related vulnerability*, high probability values were found for states such as island states, which are fairly obviously affected by their geographical location. Surprisingly, at the same time states that were considered a priori as rather sceptical with regard to measures for climate protection also showed high shares in this topic.

In summary, the iLCM was used here to uncover the initial questions regarding the various thematic priorities within the NDCs of the various countries and negotiating groups. A workflow was established, which will scale much better with increasing data volumes compared to purely qualitative approaches, which are rather commonly used in this field. For this, the possibilities of adaptability and expandability were essential to meet the specific requirements of the research question. The possibilities for qualitative assessment of the quantitative results were used. This allowed for a very efficient and in-depth evaluation, validation and interpretation of the found distribution of topics, in order to be able to make conclusive findings on the positions of the countries on the various aspects of climate change/climate protection.

References

Wouter van Atteveldt and Tai-Quan Peng. 2018. When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, 12(2-3):81–92.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1):26–31.

Andreas Blaette. 2020. *polmineR: Verbs and Nouns for Corpus Analysis*. R package version 0.8.2.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Jeanine Evers, Mauro Caprioli, Stefan Nöst, and Gregor Wiedemann. 2020. What is the refi-qda standard: Experimenting with the transfer of analyzed research projects between qda software. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 21(2).

Gerhard Heyer, Christian Kahmann, and Cathleen Kantner. 2019. Generic tools and individual research needs in the digital humanities - can agile development help? In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik - Informatik für Gesellschaft (Workshop-Beiträge)*, pages 175–180, Bonn. Gesellschaft für Informatik e.V.

Markus Hofmann and Ralf Klinkenberg. 2013. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman; Hall/CRC.

Burr Settles. 2010. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Gregor Wiedemann. 2013. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 14(2).

Textual Contexts for “Democracy”: Using Topic- and Word-Models for Exploring Swedish Government Official Reports

Magnus Ahltorp[1], **Luise Dürlich**[2], **Maria Skeppstedt**[1]
[1]The Institute for Language and Folklore, Sweden
{magnus.ahltorp,maria.skeppstedt}@isof.se
[2]Department of Linguistics and Philology, Uppsala University, Sweden
luise.durlich@lingfil.uu.se

Abstract

We here demonstrate how two types of NLP models – a topic model and a word2vec model – can be combined for exploring the content of a collection of Swedish Government Reports. We investigate if there are topics that frequently occur in paragraphs mentioning the word “democracy”. Using the word2vec model, 530 clusters of semantically similar words were created, which were then applied in the pre-processing step when creating a topic model. This model detected 15 reoccurring topics among the paragraphs containing “democracy”. Among these topics, 13 had closely associated paragraphs with a coherent content relating to some aspect of democracy.

1 Introduction and background

Methods developed within NLP have been useful additions to the computational social science and humanities toolbox. The classic NLP method of topic modelling is, for instance, widely used (Boyd-Graber et al., 2017). Examples of text genres analysed with topic modelling include news paper text (Blei, 2012), folk legends (Karsdorp and den Bosch, 2013), micro blogs (Surian et al., 2016), student essays (Ferrara et al., 2017) and open-ended survey questions (Baumer et al., 2017).

Topic models are used for discovering reoccurring topics in a collection of documents. The models are based on the co-occurrence of words. That is, words that frequently occur together indicate a recurring topic. Each topic detected is typically represented by (i) a ranked list of the words that have created the topic by frequently co-occurring, and (ii) a ranked list of the documents that are most typical for the topic, i.e., the documents in which the words frequently co-occur.

Since topic models are built on modelling the co-occurrence of words in the same texts, they model the syntagmatic relations between words. There are also NLP methods for building models based on paradigmatic relations of words. That is, models that can detect to what extent words typically occur in similar contexts, e.g., to what extent they are synonyms/near synonyms (Sahlgren, 2006). Although these models have existed for quite some time, interest has exploded in recent years with the re-emergence of neural networks as a popular method for machine learning (Vasilev et al., 2019). Also for these methods, there are many different types of use cases within social science and the humanities (Dahlberg et al., 2017; Loon et al., 2020).

We will here demonstrate how these two types of models can be applied to a text collection consisting of Swedish Government Official Reports, and how the output of the models can be combined for finding reoccurring content in the text collection. The aim for the demonstration task will be to investigate if there are topics which frequently occur in texts that mention the word “democracy”.

2 The Topics2Themes tool

Despite the shown usefulness of NLP models, previous research has also demonstrated the importance of performing a manual analysis of their output (Grimmer and Stewart, 2013; Baumer et al., 2017). For instance, to read topic-typical documents extracted by a topic modelling tool, in order to avoid misinterpreting the words representing the topics (Baumer et al., 2017; Lee et al., 2017). We, therefore, here use a tool for topic modelling, Topics2Themes (Skeppstedt et al., 2018), which has a graphical user interface meant to encourage the user to read and further analyse the documents extracted by the topic modelling algorithm.

This tool has previously been applied to other types of text collections (Skeppstedt et al., 2020a,b, 2021). Information from paradigmatic models has also been incorporated previously, in the form of word2vec models pre-trained on large corpora other than the text collection analysed. We here (i) apply the tool to the text genre of political texts, and (ii) use a word2vec model that has been trained on – and thereby is more specific to – the text type that is to be explored with topic modelling.

3 Swedish Government Official Reports

Committees or special investigators are often appointed by the Swedish Government to investigate a particular issue before a legislative proposal is presented. The results are compiled in reports, which are published in the official report series "the Swedish Government Official Reports" (or "Statens offentliga utredningar", SOU, in Swedish).[1]

The report series is made available as PDF documents and automatically extracted HTML pages at the open data site of the Swedish Parliament.[2] This HTML extraction has not preserved the logical structure of the PDF, e.g. headings and regular text are assigned the same HTML tags, and any distinction by font or type-face is encoded explicitly in style attributes which vary across different reports. Reports between the years 1994 – 2020 (3,558 reports) have, however, also recently been made available in a further processed version.[3] This version includes (i) a separation of summaries from the full texts of the reports, (ii) HTML markup that indicates titles, section headings and paragraphs in the body text, and (iii) removal of tables, lists, diagrams and non-Swedish texts. We here used the full texts of the reports from this further processed version, as well as title and heading markup.

4 Extracting "democracy" documents

Given the 2021 celebration of 100 years since the first general elections in Sweden, we decided to focus on the word "democracy", and the contexts in which it appears.

Another decision to make when constructing a topic model is how to define a document. When the collection, e.g., is made up of a compilation of short texts, the decision is easy. In this case, we instead have a collection of very long documents, which need to be split up to make them manageable (for a human as well as for the machine). We therefore decided to define a document as a paragraph.

Word	Occ.
demokratiska ("democratic", plur. & det.)	7,601
demokrati ("democracy")	6,965
demokratin ("the democracy")	6,222
demokratiskt ("democratic")	3,970
demokratisk ("democratic", common gender)	3,033
demokratins ("the democracy's")	2,613
demokratiutredningen ("the democracy-inquiry")	511
demokrativillkor ("democracy-conditioned subsidies")	449
demokratiutredningens ("the democracy-inquiry's")	396
demokrativillkoret ("the democracy-conditioned subsidies")	318

Table 1: Number of occurrences for the most common types containing the string *demokrati* ("democracy")

Following these two decisions, the collection to analyse with topic modelling was constructed as follows: We extracted all paragraphs containing the string *demokrati* ("democracy"). The string was allowed to occur as a sub-string of a word, which led to morphological derivations (e.g., democratic), as well as compound words (e.g., the democracy-inquiry) being captured. A total of 1,174 types were detected (top 10 are shown in Table 1). A manual inspection of the types showed occurrences of (different forms) of five political party names among the types detected, e.g., Social Democratic. Paragraphs containing the string *demokrati*, solely as a part of a party name were therefore excluded from the documents extracted. This resulted in a collection containing 25,988 documents (after making sure there were no exact duplicates), extracted from a total number of 2,965,751 paragraphs.

5 The word2vec model

Standard topic modelling does not take the meaning of the words into account. That is, the algorithm is agnostic to the semantic similarity of word-pairs such as "organisation"/ "organisations", and "states"/"countries". The semantic similarity of the first word-pair can be detected by, e.g., stemming, but for the second pair, there are no morphology-based solutions. Topics2Themes therefore provides a functionality for clustering the words occurring in the texts into groups of semantically close words. These words are then treated as a single

[1] https://www.riksdagen.se/sv/Dokument-Lagar/

[2] https://data.riksdagen.se

[3] At: github.com/UppsalaNLP/SOU-corpus.
In the project: datalabb.esv.se/esv-datalabb.html

concept by the tool, e.g., the combined concept "states/countries" is created. The clustering algorithm used is called DBSCAN (Ester et al., 1996). As input to the clustering algorithm, the tool needs to be provided with a suitable word2vec (Mikolov et al., 2013) model. That is, the clustering uses the semantic vector that represents each word included in the model. Since, e.g., "states" and "countries" are semantically close, they are likely to also have word2vec-vectors that are close to each other in vector space, and thereby be clustered together in the same concept cluster.

We have previously used pre-trained word2vec models. Here, we instead created a model specific to our collection, by training a word2vec model on the text type that is to be explored with topic modelling. We used the gensim library (Řehůřek and Sojka, 2010), and trained the model on 30% (48,313,487 tokens) of the report collection. We included tokens occurring at least 20 times, and used CBOW with a window size of three.

6 Applying the topic modelling tool and improving its configuration

The Topics2Themes tool was thereafter applied to the "democracy" paragraphs, using the newly created word2vec model for concept clustering.

Despite already having removed exact duplicates, the automatic duplicate detection of Topics2Themes detected 4,841 paragraphs with at least a 15-token overlap with another paragraph. These duplicates were removed, as duplicate content otherwise is interpreted as reoccurring topics.

We configured the Topics2Themes tool to use the topic modelling algorithm non-negative matrix factorization (Lee and Seung, 2001), and to extract 20 topics. However, we also configured the tool to automatically re-run the algorithm 100 times and only retain stably occurring topics. (Due to the algorithm's non-determinism, slightly different results are typically obtained each time it is run.)

Topics2Themes provides functionality for allowing the user to iteratively improve its output, both for improving the core topic modelling functionality and how the word2vec model is integrated. We therefore ran the algorithm 47 times (the first 13 times with another, pre-trained word2vec model), each time adding improvements to the model.

A basic configuration parameter is the maximum euclidean distance for two word2vec vectors to be allowed to be positioned in the same cluster. With a large distance, semantically distant words will be clustered together, whereas a small distance will lead to fewer relevant clusters being created. By manually inspecting the clusters created for different distances, we settled for a distance of 0.62.

Another important configuration improvement consists of adding additional content to four different lists. These lists contain the following (i) stop words (i.e., uninteresting words that are not to be included in the content sent to the topic modelling algorithm, e.g., "therefore", "mainly"), (ii) words to exclude from the automatic clustering since they are assigned to clusters to which they do not belong, e.g., clusters of antonyms and of semantically close words that might nevertheless be relevant to separate (e.g., party, place and person names), (iii) manually constructed clusters, i.e. groups of words that should be treated as the same concept but were not captured by the clustering (e.g. "the parliament"/ "the parliament's"), and (iv) a list of multi-word expressions that should be treated as one word by the algorithm (e.g. "political party").

For stop words, we extended the Swedish stop word list provided by NLTK (Bird, 2002) and a list from a previous Topics2Themes study (Skeppstedt et al., 2020b). By withholding the stop words from the algorithm, it is possible to prevent the creation of uninteresting topics based on these words. For instance, that two documents both contain the word "mainly" is a bad indicator of these two documents discussing the same topic. The clustering facilitates the stop word list expansion. That is, since uninteresting words are often clustered together, the entire cluster can be added as stop words.

In addition to removing stop words, we also removed low-frequency words/clusters, i.e., only the 5,000 most common words/clusters were retained.

7 Final configuration and topics detected

The final configuration resulted in 15 stable topics being detected by the topic modelling algorithm. For this configuration, there were 436 words in the list of words not to cluster, 73 manually constructed clusters, 20 multi-word expressions, and we had added 892 new words to the stop word list. A total of 530 word clusters were automatically detected by the word2vec-vector clustering.

We employed the graphical user interface of Topics2Themes for exploring the output produced by the topic modelling algorithm (Figure 1). The Topics-panel (in the center) contains one element

1: Democracy in municipalities and regions, e.g. its vitality, level of autonomy and responsibilities: municipality*, region council*, regions/responsible authority/regional councils/County Administrative Board, councils, regions, assignments, cooperation, consultation, tasks, municipal law, possibility*, activities*, municipal autonomy/the municipal autonomy

2: Internal school democracy for pupils, the school's commission to teach democracy: pupil*, school*, teacher*/teachers*, common values/values, education/teaching, influence, school, commission, knowledge, schools*, grade/primary school/gymnasium, Agency for Education, children*, curriculum*

3: Democracy-conditioned subsidies for organisations (many from SOU 2019:35): organisation*, conditions*, grants*/the support, activities*, support, civil, authority*, requirements*, ideas, Agency for Youth and Civil Society, fulfils, authority, submitted

4: A non-coherent topic: projects, perspectives, education*, universities, power

5: Political parties and their relations to voters and members (Many from SOU 2016:5): party*, voters/voters, members, internal, representative, candidates, elections, party members, shows/showed, election, role, the voters'

6: EU and democracy, e.g. how EU-democracy works, and its challenges: EC/EU, European*, national, level*, membership*, the Union/the community, Sweden, Swedish, member states*, Sweden's, countries*/states*, the deficit, the cooperation, the parliaments, power

7: Challenges, opportunities and interactions of local and regional democracy: local*/regional*, level*, national*, anchoring, county, strengthen*/improve, experimentation, work, development, development, local autonomy/municipal autonomy, responsibility

8: The importance of a broad political participation: political*, politics*, participation, the system, institutions*, system, engagement, representative, elections, equality, economic, power, forms, decision-making, social

9: Young people's political and societal participation and influence: young people*, children*, influence, commitment, participation, children and young people/children and adolescents, youth councils, to influence, engage, adults, increase, knowledge

10: About basic human and democratic rights: rights, fundamental, human rights*, limitation*/restriction*, law/ordinance rules*/provisions*, the Instrument of Government, protection, universal, common values/values, the right, society*, respect, requirements*, principles*

11: Democracy in municipalities e.g. how to strengthen it: municipal*, elected*, activity*, the audit, municipal autonomy, way of functioning, commission, strengthen*/improve, cooperation

12: Gender equality: women*, men*, gender equality, gender*, violence, power, organisation, equal, female*, women's movement, gender equality policy

13: Decision making in democracies and democratic organisations: decision*, take*, council, decision, opportunity*, influence, the board's/the council's, order, requirements*, majority, responsibility, views, legitimacy/credibility, level*, consultation

14: A wide topic, with texts containing mentions of "the Governmet". Some more specific texts about the relation between the Government and public authorities: the Government*, authority*, administration, state, the administration, commission, administrative policy, public, transparency, activities*, growth, the authority's/the agency's*/ the County Administrative Board, involvement, state administration, the work

15: The state of democracy in different countries, e.g. election participation (Many from SOU 2007:84): United States/China/Japan/Norway/Poles/India/Ireland/Canada/Hungary/Belgium/Denmark/Finland/Italy/Spain/Portugal/Romania/Czech Republic/Germany/Bulgaria/France/Slovakia/Slovenia/South Africa/Austria/Australian/the Netherlands/Great Britain, countries*/states*, Sweden, Estonia/Lithuania/Latvia, Russia, European*, elections, country, election participation, Iraq/Syria/Egypt, Lebanon/Somalia/Turkey/Jordan/Indonesia, Eastern European, Switzerland

Table 2: The 15 topics detected and their most closely associated words and concept clusters (translated into English). Concept clusters are indicated by "/" separating the words in the cluster. That the cluster contains different morphological versions of a word is indicated by a "*" following the shortest version.

for each one of the topics detected. To the left, the words/concept clusters associated with the topics detected are shown. Correspondingly, to the right, the documents (i.e., paragraphs in this case) associated with the topics are shown. It is possible to re-sort the texts according to their associated texts and words. The words associated with the topics are also highlighted in the texts. To further support the reading, each paragraph has labels that show (i) the title of the report in which it appears, and (ii) the nearest heading under which it appears.

We read a few of the most closely associated paragraphs (about five) for each one of the 15 topics detected, and added a description of the topic in its text area in the Topics panel. For the paragraphs associated with Topic 4, we were not able to find any common subject. For topic 14, the main connection between its associated paragraphs was that "the Government" was mentioned. However, for the other 13 topics, the associated texts all deal with some aspect of a common topic related to democracy. The topic descriptions and their most closely associated words can be found in Table 2

A potential effect of splitting up reports into paragraphs, and treating them as independent documents, is that the topics detected might correspond to the original reports. That was the case for topics 3, 5 and 15. For these topics, the algorithm had (more or less) detected what corresponded to the content of the reports "Democracy-conditioned subsidies for civil society organisations", "Let more people shape the future!" and "The importance of a high voter turnout", respectively. To avoid this, all paragraphs from the same source report could have been concatenated into one document. However, treating the paragraphs as independent documents also has potential advantages, e.g. making it easier to detect subtopics within a report.

8 Concluding words

It would have been a time-consuming task to manually search for reoccurring topics among the 25,988 paragraphs containing the word "democracy". With the Topics2Themes tool, in contrast, it was possible to very quickly gain an overview of reoccurring content. Additional relevant reoccurring topics might be found by analysing all paragraphs, as the tool is not likely to detect everything relevant to a human. However, in the absence of unlimited resources for manual text analysis, NLP models can be the next best option. For this particular collection – for which summaries and descriptive section headings are provided – there might be other means for gaining a quick overview of the collection. However, in cases where no such meta data exists, automatic models are even more important.

The curated concept clusters based on word2vec-vectors and the extensive stop word lists used here are by no means mandatory additions to the classic out-of-the box topic model. The classic topic model will still be able to produce topics based on the content of the texts. However, by providing the model with this additional data, it is possible for the user to transfer a part of their mental model of what they find relevant or irrelevant. E.g., we decided not to split up the automatically created clusters of foreign countries (see topic 15), as we were only interested in the concept "foreign country", and not *which* foreign country. In contrast, we decided to split up an automatically created cluster of political party names into individual concepts. Given another mental model of relevant/irrelevant, the opposite decision could have been made.

The full potential of Topics2Themes is not shown here, as the tool is also built to support a more thorough reading of the documents associated with the topics. The user interface provides a fourth panel (not shown in the figure), which makes it possible to manually add themes that the user identifies in the texts (Skeppstedt et al., 2020a, 2021). A possible continuation of the work described here would thus be to use this functionality to perform a manual search for fine-grained reoccurring themes in the documents closely associated with the topics. Another possible continuation would be to study the effect on the concept clusters created, when using a word2vec model trained on the entire report collection (instead of on a subset).

The source code for Topics2Themes is freely available[4] for use and expansion, and so are[5] the word lists, scripts, etc. used here. We hope that the demonstration provided here – of how NLP models can be used for finding reoccurring topics in large document collections – will form an inspiration for future work, e.g., work using Topics2Themes.

Acknowledgments

This study was funded by the Swedish Research Council (project number 2017-00626).

[4] github.com/mariask2/topics2themes

[5] github.com/mariask2/democracy-100-years

References

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Blei. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities*.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Stefan Dahlberg, Sofia Axelsson, and Sören Holmberg. 2017. The meaning of democracy. Using a distributional semantic model for collecting co-occurrence information from online data across languages. Technical report, Department of Political Science, University of Gothenburg.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Palo Alto, California, USA. AAAI Press.

Alfio Ferrara, Stefano Montanelli, and Georgios Petasis. 2017. Unsupervised detection of argumentative units though topic modeling techniques. In *Proceedings of the 4th Workshop on Argument Mining*, pages 97–107, Copenhagen, Denmark. Association for Computational Linguistics.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

F. Karsdorp and A. V. den Bosch. 2013. Identifying motifs in folktales using topic models. In *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556 – 562.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*.

Austin Van Loon, Sheridan Stewart, Brandon Waldon, Shrinidhi K Lakshmikanth, Ishan Shah, Sharath Chandra Guntuku, Garrick Sherman, James Zou, and Johannes Eichstaedt. 2020. Explaining the Trump gap in social distancing using COVID discourse. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Paris, France. European Language Resources Association (ELRA).

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, and Rickard Domeij. 2021. A pipeline for manual annotations of risk factor mentions in the covid-19 open research dataset. In *Selected Papers from the CLARIN Annual Conference 2020*, Linköping Electronic Conference Proceedings 180.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Andreas Kerren, Rafal Rzepka, and Kenji Araki. 2020a. Topic modelling applied to a second language: A language adaptation and tool evaluation study. In *Selected Papers from the CLARIN Annual Conference 2019*, volume 172:17, pages 145–156. Linköping Electronic Conference Proceedings.

Maria Skeppstedt, Rickard Domeij, and Fredrik Skott. 2020b. Adapting a topic modelling tool to the task of finding recurring themes in folk legends. In *Proceedings of the Digital Humanities in the Nordic Countries*, pages 388–392. CEUR Workshop Proceedings.

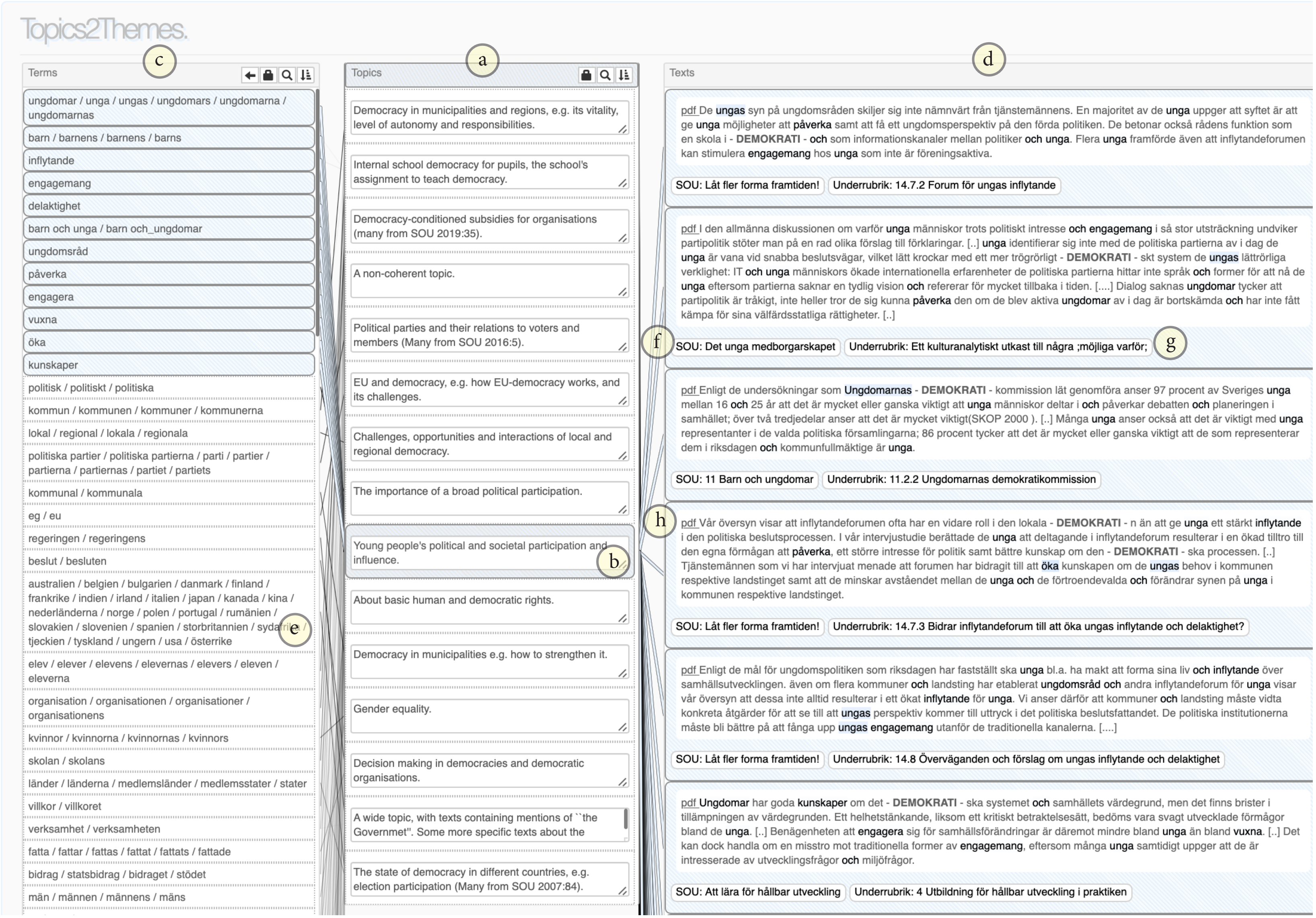
Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G Dunn. 2016. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *J Med Internet Res*, 18(8):e232.

Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, and Valentino Zocca. 2019. Python deep learning : Exploring deep learning techniques and neural network architectures with PyTorch, Keras and TensorFlow. Birmingham.

A Supplemental Material

Figure 1 shows the graphical user interface of the Topics2Themes tool after 15 topics have been automatically detected, and thereafter provided with a manually authored description.



Terms

ungdomar / unga / ungas / ungdomars / ungdomarna / ungdomarnas
barn / barnens / barnens / barns
inflytande
engagemang
delaktighet
barn och unga / barn och_ungdomar
ungdomsråd
påverka
engagera
vuxna
öka
kunskaper
politisk / politiskt / politiska
kommun / kommunen / kommuner / kommunerna
lokal / regional / lokala / regionala
politiska partier / politiska partierna / parti / partier / partierna / partiernas / partiet / partiets
kommunal / kommunala
eg / eu
regeringen / regeringens
beslut / besluten
australien / belgien / bulgarien / danmark / finland / frankrike / indien / irland / italien / japan / kanada / kina / nederländerna / norge / polen / portugal / rumänien / slovakien / slovenien / spanien / storbritannien / sydafrika / tjeckien / tyskland / ungern / usa / österrike
elev / elever / elevens / elevernas / elevers / eleven / eleverna
organisation / organisationen / organisationer / organisationens
kvinnor / kvinnorna / kvinnornas / kvinnors
skolan / skolans
länder / länderna / medlemsländer / medlemsstater / stater
villkor / villkoret
verksamhet / verksamheten
fatta / fattar / fattas / fattat / fattats / fattade
bidrag / statsbidrag / bidraget / stödet
män / männen / männens / mäns

Topics

Democracy in municipalities and regions, e.g. its vitality, level of autonomy and responsibilities.

Internal school democracy for pupils, the school's assignment to teach democracy.

Democracy-conditioned subsidies for organisations (many from SOU 2019:35).

A non-coherent topic.

Political parties and their relations to voters and members (Many from SOU 2016:5).

EU and democracy, e.g. how EU-democracy works, and its challenges.

Challenges, opportunities and interactions of local and regional democracy.

The importance of a broad political participation.

Young people's political and societal participation and influence.

About basic human and democratic rights.

Democracy in municipalities e.g. how to strengthen it.

Gender equality.

Decision making in democracies and democratic organisations.

A wide topic, with texts containing mentions of ``the Governmet''. Some more specific texts about the

The state of democracy in different countries, e.g. election participation (Many from SOU 2007:84).

Texts

pdf De ungas syn på ungdomsråden skiljer sig inte nämnvärt från tjänstemännens. En majoritet av de unga uppger att syftet är att ge unga möjligheter att påverka samt att få ett ungdomsperspektiv på den förda politiken. De betonar också rådens funktion som en skola i - DEMOKRATI - och som informationskanaler mellan politiker och unga. Flera unga framförde även att inflytandeforumen kan stimulera engagemang hos unga som inte är föreningsaktiva.

SOU: Låt fler forma framtiden! | Underrubrik: 14.7.2 Forum för ungas inflytande

pdf I den allmänna diskussionen om varför unga människor trots politiskt intresse och engagemang i så stor utsträckning undviker partipolitik stöter man på en rad olika förslag till förklaringar. [..] unga identifierar sig inte med de politiska partierna av i dag de unga är vana vid snabba beslutsvägar, vilket lätt krockar med ett mer trögrörligt - DEMOKRATI - skt system de ungas lättrörliga verklighet: IT och unga människors ökade internationella erfarenheter de politiska partierna hittar inte språk och former för att nå de unga eftersom partierna saknar en tydlig vision och refererar för mycket tillbaka i tiden. [....] Dialog saknas ungdomar tycker att partipolitik är tråkigt, inte heller tror de sig kunna påverka den om de blev aktiva ungdomar av i dag är bortskämda och har inte fått kämpa för sina välfärdsstatliga rättigheter. [..]

SOU: Det unga medborgarskapet | Underrubrik: Ett kulturanalytiskt utkast till några ;möjliga varför;

pdf Enligt de undersökningar som Ungdomarnas - DEMOKRATI - kommission lät genomföra anser 97 procent av Sveriges unga mellan 16 och 25 år att det är mycket eller ganska viktigt att unga människor deltar i och påverkar debatten och planeringen i samhället; över två tredjedelar anser att det är mycket viktigt(SKOP 2000). [..] Många unga anser också att det är viktigt med unga representanter i de valda politiska församlingarna; 86 procent tycker att det är mycket eller ganska viktigt att de som representerar dem i riksdagen och kommunfullmäktige är unga.

SOU: 11 Barn och ungdomar | Underrubrik: 11.2.2 Ungdomarnas demokratikommission

pdf Vår översyn visar att inflytandeforumen ofta har en vidare roll i den lokala - DEMOKRATI - n än att ge unga ett stärkt inflytande i den politiska beslutsprocessen. I vår intervjustudie berättade de unga att deltagande i inflytandeforum resulterar i en ökad tilltro till den egna förmågan att påverka, ett större intresse för politik samt bättre kunskap om den - DEMOKRATI - ska processen. [..] Tjänstemännen som vi har intervjuat menade att forumen har bidragit till att öka kunskapen om de ungas behov i kommunen respektive landstinget samt att de minskar avståendet mellan de unga och de förtroendevalda och förändrar synen på unga i kommunen respektive landstinget.

SOU: Låt fler forma framtiden! | Underrubrik: 14.7.3 Bidrar inflytandeforum till att öka ungas inflytande och delaktighet?

pdf Enligt de mål för ungdomspolitiken som riksdagen har fastställt ska unga bl.a. ha makt att forma sina liv och inflytande över samhällsutvecklingen. även om flera kommuner och landsting har etablerat ungdomsråd och andra inflytandeforum för unga visar vår översyn att dessa inte alltid resulterar i ett ökat inflytande för unga. Vi anser därför att kommuner och landsting måste vidta konkreta åtgärder för att se till att ungas perspektiv kommer till uttryck i det politiska beslutsfattandet. De politiska institutionerna måste bli bättre på att fånga upp ungas engagemang utanför de traditionella kanalerna. [....]

SOU: Låt fler forma framtiden! | Underrubrik: 14.8 Överväganden och förslag om ungas inflytande och delaktighet

pdf Ungdomar har goda kunskaper om det - DEMOKRATI - ska systemet och samhällets värdegrund, men det finns brister i tillämpningen av värdegrunden. Ett helhetstänkande, liksom ett kritiskt betraktelsesätt, bedöms vara svagt utvecklade förmågor bland de unga. [..] Benägenheten att engagera sig för samhällsförändringar är däremot mindre bland unga än bland vuxna. [..] Det kan dock handla om en misstro mot traditionella former av engagemang, eftersom många unga samtidigt uppger att de är intresserade av utvecklingsfrågor och miljöfrågor.

SOU: Att lära för hållbar utveckling | Underrubrik: 4 Utbildning för hållbar utveckling i praktiken

Figure 1: Topics2Themes applied to texts containing the string "demokrati" ("democracy"). The Topics panel (a) shows the 15 topics detected. The topic selected by the user (b) is shown with a blue background, and this topic's most closely associated words (in the Terms panel, c) and most closely associated texts (in the Texts panel, d) have been positioned as the top-ranked elements in their panels. The Terms panel shows examples of concept clusters, e.g., a large cluster of country names (e). To each text, two labels are attached: The name of the report in which the text appears (f), and the name of the nearest heading under which the text appears (g). There is also a link to the full PDF version of the report in which the text appears (h).

Detecting policy fields in German parliamentary materials with Heterogeneous Information Networks and Node Embeddings

Alexander Brand
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
alexander.brand@uni-hildesheim.de

Wolf J. Schünemann
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
wolf.schuenemann@uni-hildesheim.de

Tim König
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
tim.koenig@uni-hildesheim.de

Tanja Preböck
Chair of Business Education
University of Bamberg
Bamberg, Germany
tanja.preboeck@uni-bamberg.de

Abstract

This short paper illustrates a combination of Heterogeneous Information Networks (HINs) and Node Embeddings for application in policy research. HIN-based methodology is useful to cope with the dynamic and permeable ontology of policy fields, as it allows to integrate various entities and political interactions within a single network, prepared for diverse analytical operations. The paper presents the theoretical foundations, basic network construction and potential use cases for applied policy studies.

1 Introduction

In this short paper, we present a combined methodology of Heterogeneous Information Networks (HINs) and Node Embeddings for application in policy research. Our work is rooted in an ongoing project in which we develop a mapping and retrieval architecture for German policy debates at large scale, based on various data sources. Here, HIN-based methodology is useful to cope with the dynamic and permeable ontology of policy fields. It allows to integrate various entities and interactions identified in political discourse as constitutive elements of a policy field within a single network. Starting with theoretical reflections on policy fields as networks, the following sections explain the basics of network construction, the selection of metapaths and introduce some empirical applications beyond our central goal of field detection. While our methodology is designed for the integration of various data sources, this paper focuses on a dataset of parliamentary activity for the German Bundestag. For illustrative purposes, we limit our analyses to a single legislative period.

2 Theoretical foundations and the state of research

For the last three decades, policy research has moved from its traditional preoccupation with state-led political steering to studying more flexible and heterogeneous governance constellations. Correspondingly, leading scholars in the field widened their scope from institutional to ideational foundations of policy and policy development (Sabatier, 1988), thereby meeting the growing interest in dynamic processes of policy development, like policy learning, instability and the emergence of policy fields. Today, there is broad consensus about the dynamic and permeable ontology of policy fields. These theoretical re-orientations have been reflected in empirical research by at least two coterminous trends. First, network approaches emerged as particularly well suited for studying policy fields. While there had been early accounts that introduced the network as basic concept (and metaphor) for understanding policy development (Heclo, 1978), network-based policy research became more systematic only when taking up theoretical and methodological developments in Social Network Analysis (SNA) (Freeman, 2004; Heinz et al., 1990; Kenis and Schneider, 1991; Laumann and Knoke, 1987). Currently, we are observing a new wave of network-based policy research driven by methodological developments, in particular the rise of computational methods, and the drastic increase of data availability, allowing for the study of interdependent relations in complex social systems (Lazer and Wojcik, 2018).

Second, there is a rich theoretical tradition of interpretative policy analysis rooted in discursive approaches, centred around qualitative and critical research methods (Fischer, 2003; Hajer, 2002; Yanow, 2009). The same holds true for the research strand applying Bourdieuan field theory to policy analysis (Fligstein and McAdam, 2015). While, conceptually, we can draw on both strands' rich theory when defining the discursive entities and relations relevant in field detection, there is potential for advancement with regard to research methods. Most importantly, Computational Social Science has offered new trajectories for policy research in recent years, mostly relying on 'text as data'-approaches. Novel Natural Language Processing tools include Named Entity Recognition as well as Word Embeddings, which are recently gaining popularity in political science research (Rheault and Cochrane, 2020; Rodriguez and Spirling, 2021). Beyond these text-based methods, innovative Node Embedding approaches aim to represent the network-structural level of political discourse (Won and Fernandes, 2021). While combinations of network and discourse analysis are not new to policy studies with - most notably - the Discourse Network Analysis (DNA) (Leifeld, 2018), existing tools mostly do not meet the requirements of scalability and automated analysis. By using a mixed methodology based on HINs and Node Embeddings, we address this gap.

3 Methodological Considerations

In the following, we describe the formal properties and notations of the two techniques combined in this paper: HINs and Node Embeddings. HINs can be formalised as $G = (V, E)$ with a node type mapping $\phi : V \to A$, which is a graph defined over the node types A. This ability to incorporate multiple node types (heterogenous or multipartite) versus networks with only one type of nodes (homogenous or monopartite) enables the connection of multiple fragments of meaning in one coherent framework. Relating such different node types to each other inherently complicates relations within the network (e.g. meaningful relations can now consist of more complex connection patterns or paths). To map these complex relations we can utilise metapaths P, which describe paths (sets of connections between nodes of different or same kind of A) on the graph and have the form of $A_1 \longrightarrow A_2 \longrightarrow \ldots \longrightarrow A_{l+1}$. Such metapaths are usable in a wide variety of cases to map complex relationships between different objects in networks, e.g. two scientists presenting papers at the same venue ($Author_1 \longrightarrow Paper_1 \longrightarrow Venue \longrightarrow Paper_2 \longrightarrow Author_2$) (Sun and Han, 2013). Furthermore, metapaths have been used to enhance Node Embedding techniques, resulting in improved vector space representations (Dong et al., 2017). Node Embedding procedures can be understood as structurally similar to established Word Embedding methods (Rheault and Cochrane, 2020), treating paths in networks as equivalent to word sequences (Grover and Leskovec, 2016), with the goal to find an appropriate vector representation of each node. To construct such paths, random walks are performed on the network with a biased random walker, travelling only on specific metapaths. Such a learned representation can then be used in tasks like link prediction or, in our case, to find meaningful clusters.

4 Exemplary Analysis

For our analysis we used the "every single word" dataset (Remschel and Kroeber, 2020), which contains all written communication (reports, petitions, etc.) published by the German Bundestag between 1949 and 2017.

For our showcase, we restricted the sample to the 10th election period (1983-1987) for two reasons. First, this allows us to map node separation on a more granular level than the whole dataset. Second, given our overarching interest in major discursive shifts in relation to policy development, we focus on an election period with a new issue-oriented party in the Bundestag, i.e. the German Green Party ("Bündnis90/Die Grünen") achieving representation in the federal parliament for the first time in its history. After restricting the dataset in this way, our sample consists of N = 6534 documents. We utilised a multistep design to construct a clustered representation of the network. In the first step, we constructed a Heterogenous Information Network with the node types A of Committee (Com) (e.g. parliamentary committee), Fraktion/Bundes* (F/B*) (parliamentary faction, federal institution), Keyword, Named Entity and Document. We use documents as the seed nodes with every other node type being connected to a document via mentions or authorship. To determine keywords for each doc-

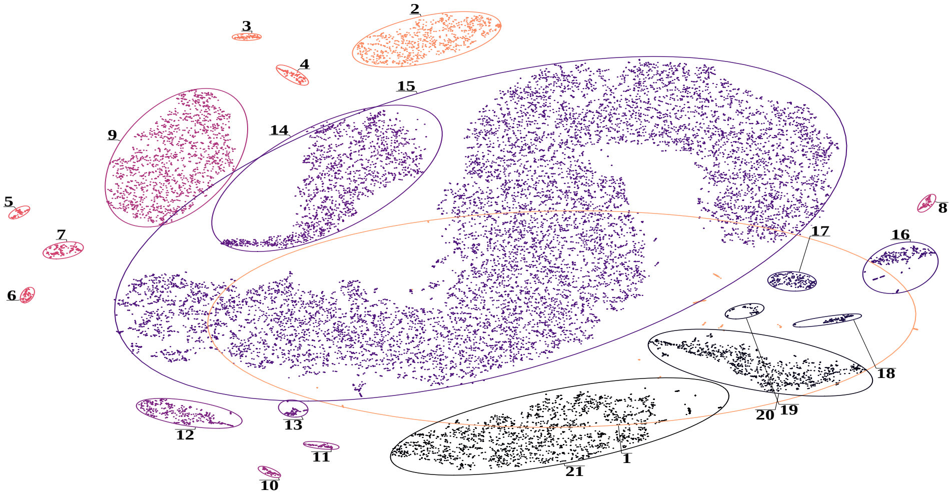


Figure 1: Visualisation of the Node Embeddings via Metapath2Vec. Colors symbolize assigned clusters via dbscan on the embedded model. X and Y coordinates are according to a two-dimensional projection via T-SNE. Cluster 1 (light orange) spans across multiple regions in the plot (label next to cluster 21).

ument we used a simple ranking method of nouns in the text via tf-idf (Aizawa, 2003) and cut terms below the sixth decile from the network. To identify Named Entities, we used a matching approach on the Named Entities already annotated in the GermaParl Corpus (Falk and Meisinger, 2020). To obtain a 2d representation of the network, we used a metapath-based Node Embedding as proposed by Dong et al. (2017) with the metapaths P of

1. $Document \longrightarrow Com \longrightarrow Document$
2. $Document \longrightarrow F/B^* \longrightarrow Document$
3. $Document \longrightarrow Keyword \longrightarrow Document$
4. $Document \longrightarrow Entity \longrightarrow Document$

to bias a random walker on the network with a walk length of 100. The application of these metapaths mirrors theoretical assumptions grounded in discourse studies and field theory. Metapaths 1 and 2 represent the fact that authors and other entities like parliamentary factions or specialised committees may become central nodes in policy fields due to their expertise, their dedicated roles or prolonged activity. Utilising these metapaths leverages the structuring roles specialised actors exhibit in their domains of expertise for the identification of policy fields. Metapaths 3 and 4 aim to identify documents connected through their content. By generalising Node Embeddings over textual and non-textual node types, we can leverage the structuring power of both content and actors in the identification of policy debates. As our main interest is in identifying documents that can be grouped and assigned to a certain policy, our meta-path construction is document-centred. Thus, there are other relations disregarded, e.g. between committees and/or factions via membership overlap. However, it is important to keep in mind that for the scalable architecture of our analysis, there are no 'real world' relations between committees or factions that could directly be drawn from the material. Therefore, these relations have to be constructed via meta-paths which are theoretically meaningful in the context of our analysis.

To display the 128-dimensional network embedding in 2d, we utilised a T-SNE-based dimension reduction (van der Maaten and Hinton, 2008). We then followed the approach by Stellargraph (2020) to estimate approximative clusters on the embedded network via dbscan. Minimal cluster size was set to 50 nodes for two reasons: First, we obtained rather good separation as quantitatively indicated by the average silhouette width $\overline{S_i}$ when compared to minimal cluster sizes like 10 or 20 (Table 1). Second, we found the interpretation of clusters under this parametrisation relatively easy. A representation of the clustered, embedded Heterogeneous Information Network ($V = 22763, E = 792127$) can be obtained from figure 1. Following the separation via dbscan (Ester et al., 1996), we can identify 21 clusters. Their relative position on the 2d representation indicates discursive proximity. The nomination of the clusters was conducted via Qualitative Content Analysis. Therefore, in line with the evaluation of structural topic models and the power-law distribution in heterogeneous complex networks (Sarshar and Roychowdhury, 2005), the 10 documents with the highest degree per cluster were qualitatively coded. In the second step, we selected the best fitting category from a well-established policy taxonomy for each of the clusters by comparing our qualitative results with the description in the respective codebook. Among relevant projects, we chose the codebook developed for the UK Policy Agenda Project (Jennings and Bevan, 2012) as our main orientation. While at the level of overarching policy codes, there is much overlap between the codebooks presented in the literature, we deliberately selected a codebook for a national polity (not the EU itself) and a parliamentary system. However, for interpretative analysis, we adjusted for German peculiarities, e.g. with respect to the topic of CMIC and the immigrants from GDR or the late repatriates ("Spätaussiedler"). The largest cluster (15), containing roughly half of all nodes, deals with Government Operations of various kinds. Clusters 3 and 4 are relatively distinct, yet close together, which is also represented in their topics (Agriculture and Environment). Cluster 2 is relatively big and handles a diverse topic: Social Welfare in all its shades, such as Child Raising Allowance and Single Parents.

The shape of cluster 1 is very distinct: Although it has few nodes, it takes up a lot of space. Law, Crime, and Family Issues is a transverse dimension that runs through the different sections (especially Government Operations, Civil Rights, Macroeconomics and Energy).

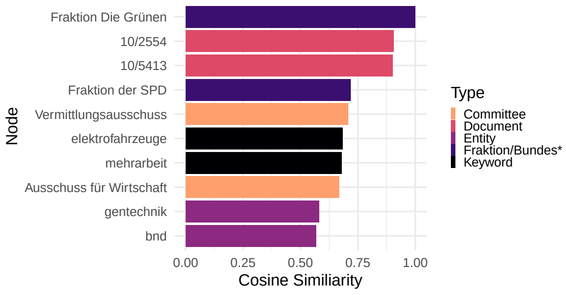


Figure 2: Visualisation of cosine similarity for "Fraktion Die Grünen" from the Node Embeddings in Figure 1. The graphic shows the two most similar nodes with regard to their connection pattern in the network for each node type.

Cluster	N	$\overline{S_i}$
1 LCFI	184	-0.61
2 Social Welfare	853	0.59
3 Agriculture	75	0.89
4 Environment	105	0.84
5 BFD; Social Welfare	68	0.9
6 Social Welfare	77	0.9
7 EU Affairs	132	0.82
8 Labour and Employment	77	0.92
9 Government Operations	1736	0.34
10 Environment; BFD; Mixed	67	0.89
11 CMIC	76	0.85
12 Macroeconomics	355	0.66
13 Government Operations	72	0.86
14 Infrastructure	1854	0.37
15 Government Operations	12846	-0.33
16 Mixed	292	0.61
17 CMIC	195	0.74
18 Macroeconomics	103	0.75
19 Government Operations	66	0.8
20 Energy	902	0.27
21 Government Operations	2007	0.36

Table 1: Average silhouette width by cluster. Higher values indicate better separation. BFD = Banking, Finance, and Domestic Commerce; CMIC = Civil Rights, Minority Issues, Immigration and Civil Liberties; LCFI = Law, Crime, and Family Issues

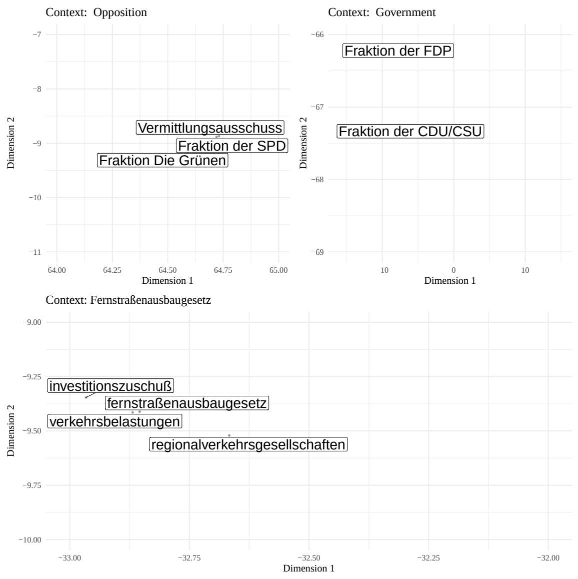


Figure 3: Visualisation of near neighbours for Opposition / Government and Fernstraßenausbaugesetz from the Node Embeddings in Figure 1. Documents are removed for readability.

Another advantage of node embedded representations is the possibility to exploit their structural properties and evaluate near neighbours of nodes with arbitrary types. For example, one can evaluate the distance between certain actors and different keywords, entities or documents via simple metrics like the cosine distance. A showcase of such evaluation for the ”Fraktion Die Grünen” (Green faction) is presented in figure 2. One can observe that the node has a high similarity with nodes like ”elektrofahrzeug” (electro mobility) or ”gentechnik” (genetic engineering) which means such words are more frequent in texts by the Green faction. Some evaluations can be even simpler. As the representation of the network itself through Node Embeddings allows an evaluation of the structure via their dimensions, one can simply zoom into the 2d representation to evaluate structurally similar nodes. Such a visualisation can be found in figure 3. The 2d representation clearly distinguishes between opposition and government (top panels) and can show similarly used words for specific keywords like ”fernstraßenausbaugesetz” (Highway Expansion Act) in a fashion comparable to Word Embedding techniques (bottom panel).

5 Conclusion

By combining HINs and Node Embeddings, this paper demonstrated how the integration of heterogeneous node types into a singular network can be used to identify policy fields in large text corpora. In selecting metapaths based on assumptions grounded in discourse- and field theory, we leverage the structuring power of different node types and contribute to closing the gap between network- and discourse-based approaches in policy analysis. Not only does our method allow for the scalable identification of policy fields in large corpora. We also demonstrated how it provides applications in the analysis of these policy fields through the nodes' and clusters' cosine similarity or relative proximity in their 2d representation. Furthermore, our generalised approach can be applied to various types of text, with the potential to combine heterogeneous data sources into a single network. Finally, the networked structure is easily sliced into time frames, allowing for the analysis of dynamic changes within policy fields as well.

Acknowledgements

This research has been funded by the Klaus Tschira Foundation in the framework of the EPINetz project.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Neele Falk and Nico Meisinger. 2020. Lapps-clarin germaparl metadata enrichment.

Frank Fischer. 2003. *Reframing public policy: Discursive politics and deliberative practices*. Oxford University Press, Oxford.

Neil Fligstein and Doug McAdam. 2015. *A theory of fields*, first issued as an oxford university press paperback edition. Oxford University Press, Oxford and New York and Auckland.

Linton C. Freeman. 2004. *The development of social network analysis: A study in the sociology of science*. Empirical Press, Vancouver, BC.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Maarten A. Hajer. 2002. Discourse analysis and the study of policy making. *European Political Science*, 2(1).

Hugh Heclo. 1978. Issue networks and the executive establishment. In Anthony King, editor, *The New American political system*, AEI studies, pages 46–57. American Enterprise Institute for Public Policy Research, Washington.

John P. Heinz, Edward O. Laumann, Robert H. Salisbury, and Robert L. Nelson. 1990. Inner circles or hollow cores? elite networks in national policy systems. *The Journal of Politics*, 52(2):356–390.

Will Jennings and Shaun Bevan. 2012. Uk policy agendas codebook.

Patrick Kenis and Volker Schneider. 1991. Policy networks and policy analysis: Scrutinizing a new analytical toolbox. In Bernd Marin and Renate Mayntz, editors, *Policy networks*, pages 25–62. Campus-Verl., Frankfurt am Main.

Edward O. Laumann and David Knoke. 1987. *The organizational state: Social choice in national policy domains*. WIS-Edition. The Univ. of Wisconsin Press, Madison, Wis.

David Lazer and Stefan Wojcik. 2018. Political networks and computational social science. In Jennifer Nicoll Victor, Alexander H. Montgomery, and Mark Lubell, editors, *The Oxford handbook of political networks*, pages 115–130. Oxford University Press, New York, NY.

Philip Leifeld. 2018. Discourse network analysis. In Jennifer Nicoll Victor, Alexander H. Montgomery, and Mark Lubell, editors, *The Oxford handbook of political networks*, pages 301–325. Oxford University Press, New York, NY.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Tobias Remschel and Corinna Kroeber. 2020. Every single word: A new data set including all parliamentary materials published in germany. *Government and Opposition*, pages 1–20.

Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*.

Pedro Rodriguez and Arthur Spirling. 2021. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. 2021. Publisher: The University of Chicago Press.

Paul A. Sabatier. 1988. An advocacy coalition framework of policy change and the role of policy-oriented learning therin. *Policy Sciences*, (23):129–168.

Nima Sarshar and Vwani Roychowdhury. 2005. Multiple power-law structures in heterogeneous complex networks. *Physical Review E*, 72(2):026114.

Stellargraph. 2020. Comparison of clustering of node embeddings with a traditional community detection method — StellarGraph 1.0.0rc1 documentation.

Yizhou Sun and Jiawei Han. 2013. Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology*, 18(4):329–338.

Miguel Won and Jorge M Fernandes. 2021. Analyzing twitter networks using graph embeddings: an application to the british case. *Journal of Computational Social Science*, pages 1–11.

Dvora Yanow. 2009. *Conducting interpretive policy analysis*, [nachdr.] edition, volume 47 of *Qualitative research methods series*. Sage Publ, Thousand Oaks, Calif.

A semi-supervised approach to classifying political agenda issues

Tim Kreutz and **Walter Daelemans**
CLiPS - Computational Linguistics Group
Department of Linguistics
University of Antwerp
{tim.kreutz,walter.daelemans}@uantwerpen.be

Abstract

This paper presents a semi-supervised approach to classifying political texts with the Comparative Agendas Project coding scheme. Starting with limited domain knowledge in the form of ten seed words that are central to the meaning of a topic, new candidate textual indicators are found using a graph propagation algorithm over a semantic network of words and phrases. We show that there is a balance between precision and recall when it comes to the number of candidates to add to a lexicon for each topic, and optimize this balance on the basis of a development dataset. The automatically generated lexica substantially outperform the handmade CAP-lexicon in four tested genres: political party manifestos, news articles, parliamentary documents and social media texts. Besides having better discriminatory qualities, these lexica require less resources to generate and are more genre-independent than their handmade counterparts.

1 Introduction

Political experts can analyze newspapers or television channel and summarize the attention given to certain political issues in the media. The Comparative Agendas Project (CAP) (CAP)[1] provides coding schemes in many languages to aid such analyses and make them comparative. But even with clear guidelines, manual coding of political texts becomes prohibitively time consuming.

Recently, more research has focused on automatic content analysis to help expert annotation, especially in the social and political sciences. Two main methods have been proposed: dictionary-based approaches and supervised learning approaches. In dictionary-based approaches an expert-made lexicon is constructed of high precision indicators that are linked to political topics. These indicators are words or other language units. Despite the insight and expertise contained in these topic lexicons, they usually suffer from low coverage over the instances; a dictionary-based system cannot make decisions about documents that contain none of the dictionary words.

In a supervised learning approach, annotators assign labels to a large collection of documents and a machine learning algorithm learns weights between features in the documents and the labels. For example, Purpura and Hillard (2006) classified US congressional legislation using support vector machines and score 88.7% accuracy on major topics and 81.0% on subtopics. This is close to human agreement on the task. There are drawbacks to supervised learning too. A system trained on congressional legislation will perform differently on newspaper articles or social media messages. To achieve consistent results across genres, the classifier would have to be retrained on additional annotated in-genre documents.

We introduce a hybrid solution in this paper; a semi-supervised approach to classifying political texts according the CAP coding scheme for political agendas. Our contribution does not require expert annotation yet improves an existing lexicon-based approach with regards to recall. We obtain these results for two languages (Dutch and English) and for four different genres of political texts namely party manifestos, news articles, parliamentary reports and tweets.

2 Related work

Most previous work on automatic content analysis of political texts with regards to political issues used the coding scheme developed by the Policy Agendas Project (PAP) (John, 2006) and the successive Comparative Agendas Project. The codebook developed by CAP discerns 20 major political top-

[1] https://www.comparativeagendas.net/

ics.

2.1 Dictionary-based

Sevenans et al. (2014) created a Dutch and an English dictionary by taking topic indicators from the respective CAP coding schemes and adding synonyms and related terms by hand. The topic indicators that identify a certain topic can be words or partial words (suffixes, infixes or affixes). In the English lexicon, for example, the topic macroeconomics contains “econom” which will match “economy”, “economist”, “noneconomic” and many others.

The classification performance of the lexicons differed greatly between the topics and the languages. For the English lexicon, a considerable number of the parliamentary questions did not contain any of the dictionary words (22%) and could not be classified. Interestingly, the parliamentary questions in Dutch did not receive a class in only 5% of the cases. The authors go into detail on the quality of the lexicon for specific topics, but on average, performance was low compared to human annotations: 0.43 recall, 0.52 precision and 0.61 recall, 0.60 precision for English and Dutch, respectively.

Praet et al. (2018) apply the Dutch CAP-lexicon to tweets by Flemish politicians. More than half (54%) of the tweets did not match with any dictionary word, leading to very low classification accuracy.

2.2 Supervised learning

Supervised classification with the CAP coding scheme has been applied to US congressional legislation (Purpura and Hillard, 2006), Norwegian news texts (Hagen, 2012), Kroatian news headlines (Karan et al., 2016), and tweets by US state legislators and governmental bodies (Li, 2016; Qi et al., 2017). There is also a body of work on the supervised classification of party manifestos, which draws its labels from the separate but similar coding-scheme in the Comparative Manifesto Project (CMP) (Zirn et al., 2016; Glavaš et al., 2017).

These standalone applications of machine learning architectures work well in general. Congressional documents are assigned the right major topics in almost 90% of cases while performance drops with shorter texts such as media headlines (0.77% accuracy) and tweets (around 65% accuracy).

As far as we know there has not been an extensive study on cross-domain portability of the supervised classification systems. Rihiu Li (2016) trains a CNN on tweets from state legislators from Iowa and Nebraska. They note that there is a drop in performance when training on data from one state and testing on the other, but this drop could also be due to a smaller training set compared to training on both. Grimmer and Steward (2013) note that supervised machine learning systems are inherently domain- and problem-specific but see this as an advantage over multi-purpose dictionary-based systems.

2.3 Semi-supervised learning

Semi-supervised approaches bootstrap minimal domain knowledge to learn about a problem. As such, the invested expert knowledge and effort are far less than in supervised learning.

Semi-supervised approaches have seen frequent use in sentiment analysis. Rao and Ravichandran (2009) use synonym and hypernym relationships from WordNet to deduce sentiment information. Starting with positive and negative seed terms obtained from the General Inquirer[2] lexicon, polarities are propagated over the word graph using a label propagation algorithm (Zhu and Ghahramani, 2002). Even with as few as ten seeds terms, word polarity scores could accurately be predicted using semi-supervised learning: “*label propagation is especially suited when annotation data is extremely sparse*” (Rao and Ravichandran, 2009).

Kreutz and Daelemans (2018) induce polarity scores without using handmade knowledge graphs. Instead, they take seed words from an existing sentiment lexicon and propagate sentiments to candidate words that appear in similar contexts. The additions made to the existing lexicons improved sentiment analysis for two different domains.

3 Data

3.1 CAP lexicon

We take the CAP-lexicons, which were developed for the Flemish and United States contexts by (Sevenans et al., 2014), as a starting point for semi-supervised learning. The number of indicators linked to a topic ranges from 35 to 102 and 28 to 143 in Dutch and English respectively.

[2] http://www.wjh.harvard.edu/~inquirer/

3.2 Text genres

The datasets for testing were obtained from the Comparative Agendas Project [3] and were created for the Flemish and U.S. CAP-subprojects [4]. We selected the datasets in Table 1 to reflect the diversity in genres while having comparable types of documents across Flemish and U.S. contexts. All documents have major topic codes from the CAP codebook.

4 Methods

4.1 Seed selection

To demonstrate that our approach requires only limited domain knowledge the initial dictionary is restricted to only ten indicators per topic. These seed terms needs to both precisely and frequently denote a topic. We calculate degree centrality between topic indicators in our data sets and select seeds based on this score since it is found to be an effective measure for quantifying such keyword like qualities (Boudin, 2013).

4.2 Extending seed terms

We use a network of words to extend seed terms with other candidates. Edges between words are based on distributional semantics, in which words that occur in similar contexts are more strongly connected. We use the well-established Word2Vec (Mikolov et al., 2013) algorithm to calculate cosine similarities between candidates (words or phrases) to use as edge weights. Our Word2Vec models were trained on the Google News dataset for English[5] and an unpublished corpus of Dutch newspaper and online news data with 300-dimensional vectors and negative sampling.

Suitable candidates are found by doing random walks over the network of words starting from the selected seeds. This method is adapted from SentProp (Hamilton et al., 2016), a package originally intended for inducing sentiment lexicons. We adopt its default settings of connecting words to their ten nearest neighbors. A word or phrase which is found often by a random walk from a seed get a higher score for that topic, while a penalty is applied if the word or phrase is found from a seed term that is linked to another topic. A ranking of candidates by scores then determines in which order they be added to the lexicon.

4.3 Determining a cut-off

We split the annotated data in a stratified development and test set (50% each). The development set is used to determine a cut-off for candidate words. As seed terms and candidates become more dissimilar, adding more candidates can harm the discriminative performance of the dictionary. Figure 1 shows precision, recall and F-score for the Civil Rights topic on development data at different numbers of added candidates. Although recall improves when more indicators are being added for this topic (more documents are classified as belonging to Civil Rights), precision suffers. The cut-off is determined as the highest harmonic mean of precision and recall (the F-score optimum).

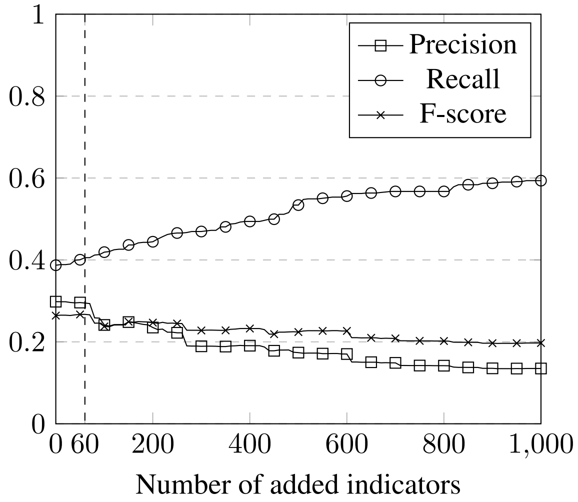


Figure 1: As the number of words added to *Civil Rights* increases, recall improves and precision decreases. The optimal number of added words (60) lies at the F-score optimum.

4.4 Classification algorithm

Classifying a text is done by simply checking if any of its words appear in the lexicon for one of the topics. We assign the document a label based on the topic that occurred most often. Although a more refined algorithm could be used to take into account class distribution or to assign different weights to words, using a simple classification algorithm ensures that each entry precisely denotes a topic and does not greatly offset the decision boundaries.

[3] https://www.comparativeagendas.net/datasets_codebooks

[4] Full citations for the Belgian datasets, the U.S Democratic and Republican party platform datasets, the U.S. bills dataset and NYT dataset in acknowledgements 1, 2, 3 and 4 respectively. (excluded here for anonymous review)

[5] Available at https://code.google.com/archive/p/word2vec/

	Belgium		U.S.	
Domain	Type	# Documents	Type	# Documents
Manifestos	Party manifesto excerpts	5,147	Party manifesto excerpts	7,296
News media	De Standaard abstracts	17,981	New York Times abstracts	17,216
Parliament	Bills	4,868	Congressional bills	52,366
Social media	Tweets by politicians	6,027	Tweets by state legislators	16,988

Table 1: Data spanning four genres and two contexts is used to tune and evaluate the semi-supervised approach.

	U.S.				Belgium			
Method	Precision	Recall	F1	Entries	Precision	Recall	F1	Entries
Original lexicon	0.308	0.207	0.210	3,610	0.446	0.408	0.378	8,353
Seed selection	0.311	0.262	0.246	200	0.428	0.367	0.346	200
Induced lexicon	**0.321**	**0.286**	**0.278**	9,976	**0.448**	**0.422**	**0.387**	18,650

Table 2: The macro-averaged results of the lexica created with SentProp compared to the original hand-made lexicon and the lexica containing only seed terms.

5 Results

Adding the optimal number of candidate indicators to the seeds results in the final induced lexica for testing. The lexica contain 9,976 and 18,650 words for the U.S. and Belgian context respectively. Table 2 lists their results compared to using the original lexicon of hand-picked words and the seed lexicon on the test data.

The induced lexica outperform the original lexica both in the U.S. and Belgian context and not only with regards to recall. Surprisingly, the added candidates also more precisely denote a topic compared to a handmade lexicon. We regard this latter result mainly as a demonstration of how difficult it is, even for experts, to construct a dictionary that can distinguish between topics in a real-world setting by hand.

In absolute terms, results are still rather poor. This is to be expected considering that a lexicon distinguishes 20 different major political topics, and that some genres, social media texts in particular, contain very little information. Another problem is the genre independence that a classifier needs to work on party manifestos, news, bills as well as social media texts. We believe a supervised classifier trained on other political texts will face the same difficulty in deciding on the right label, although future work will have to compare these approaches in detail.

6 Conclusion

We introduced an easy to use semi-supervised approach for inducing dictionaries suitable for classifying diverse political texts. The induced dictionaries outperform a handmade lexicon from the Comparative Agendas Project across contexts (U.S. and Belgium) and genres (political party manifestos, news articles, bills and social media texts).

Creating lexica in an automatic way is less time consuming while remaining as interpretable and easily adaptable as existing dictionary-based approaches; the words or phrases can be inspected and changed by experts when necessary. Future work should compare the semi-supervised method with supervised models, both in terms of overall performance and in diverse cross-genre settings.

6.1 Data and code availability

All datasets used in this paper, except for the tweets which cannot be freely shared due to GDPR[6] restrictions, are available from the CAP website.

To enable replicability and direct comparison in future work, we publish our method in a public code repository. Alongside the code we present the induced lexica for both U.S. and Belgian contexts here: `https://github.com/clips/lextension`.

References

Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the sixth international joint conference on natural language processing*, pages 834–838.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of top-

[6] `https://gdpr.eu`

ics in political texts. Association for Computational Linguistics (ACL).

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Thomas M. Hagen. 2012. Automatic topic classification of a large newspaper corpus. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, chapter 6, pages 111–130. John Benjamins Publishing, Oxford.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 595. NIH Public Access.

Peter John. 2006. The policy agendas project: a review. *Journal of European Public Policy*, 13(7):975–986.

Mladen Karan, Jan Šnajder, Daniela Širinić, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 12–21.

Tim Kreutz and Walter Daelemans. 2018. Enhancing general sentiment lexicons for domain-specific use. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1056–1064.

Rihui Li. 2016. Classification of tweets into policy agenda topics.

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

Stiene Praet, Walter Daelemans, Tim Kreutz, Peter Van Aelst, Stefaan Walgrave, and David Martens. 2018. Issue communication by political parties on twitter. In *Data Science, Journalism & Media 2018, August 20, 2018, London, UK*, pages 1–8.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225.

Lei Qi, Rihui Li, Johnny Wong, Wallapak Tavanapong, and David AM Peterson. 2017. Social media in state politics: Mining policy agendas topics. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 274–277.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Julie Sevenans, Quinn Albaugh, Tal Shahaf, Stuart Soroka, and Stefaan Walgrave. 2014. The automated coding of policy agendas: A dictionary based approach (v. 2.0.). In *CAP Conference*, pages 12–14.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *CMU-CALD-02-107*.

Căcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos.

INVITED TALK II
Coding and Mining Arguments for a Better Democracy

Katharina Esau
University of Düsseldorf
Düsseldorf, Germany
katharina.esau@hhu.de

Abstract

The increased desire of citizens to participate in political processes has prompted numerous state-organized participation procedures in the last two decades (e.g., citizens' assemblies, deliberative forums). Such attempts seem particularly promising at the local level of politics where, if successful and well designed, publicly expressed opinions and local knowledge of citizens can be incorporated into decision-making. Against the theoretical background of deliberative democracy, asynchronous online discussions offer a desirable infrastructure for a reasoned public sphere. As such these platforms are of great interest to investigate. However, successful platforms attract large numbers of participants and produce large amounts of text data and that quickly becomes difficult to manage manually. Therefore, automated techniques are invaluable when it comes to analysing citizens' contributions and informing decision makers. The talk presents a method mix combining both manual content analysis and automated techniques. It shows that this can be a fruitful approach for the extraction of argument components and other discussion elements, such as emotions and narratives, from user content. To illustrate this, the methodology and results of a semi-automated content analysis that examined one participation platform (Tempelhofer Feld, Berlin) will be presented. The annotation tool BRAT will be briefly demonstrated and the possibilities for relational coding and analysis of text data explained. Throughout, the challenges and opportunities of interdisciplinary collaboration between social sciences and computer science will be addressed.

Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks

Allison Koh
Centre for International Security
Hertie School
koh@hertie-school.org

Daniel Kai Sheng Boey
School of International and Public Affairs
Columbia University
daniel.boey@columbia.edu

Hannah Bechara
Data Science Lab
Hertie School
bechara@hertie-school.org

Abstract

Hand-labeled political texts are often required in empirical studies on party systems, coalition building, agenda setting, and many other areas of political science research. While hand-labeling remains the standard procedure for analyzing political texts, it can be slow, expensive, and subject to human error. Recent studies in the field have leveraged supervised machine learning techniques to automate the labeling process of political texts. We build on current approaches to label shorter texts and phrases in party manifestos using a pre-existing coding scheme developed by political scientists for classifying texts by policy domain and preference. Using labels and data compiled by the Manifesto Project, we make use of the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) with Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) to seek the best model architecture to supplant manual coding of political texts. We find that our proposed BERT-CNN model outperforms other approaches for the task of classifying political texts by policy domain.

1 Introduction

During campaigns, political actors communicate their position on a range of key issues to signal campaign promises and gain favor with constituents. Identifying the political positions of political actors is essential to understanding their intended political actions. This is why policy preferences—or positions on specific policy issues expressed in speech or text—have been extensively analyzed within the relevant political science literature (Abercrombie et al., 2019; Budge et al., 2001; Lowe et al., 2011; Volkens et al., 2013). Methods employed to investigate the policy preferences of political actors include analysis of roll call voting, position extraction from elite studies or regular surveys, expert surveys, hand-coded analysis, and computerized text analysis (Debus, 2009). Studies that utilize political manifestos, electoral speeches, and debate motions often rely on the availability of machine-readable documents that are labeled by policy domain or policy preference.

Quantitative methods, especially in the field of natural language processing, have enabled the development of more scalable methods for predicting policy preferences. These advancements have enabled political scientists to analyze political texts and estimate their positions over time (Nanni et al., 2016; Zirn et al., 2016). To better understand the political positions of political actors, many social science researchers have turned to hand-labeling political documents, such as parliamentary debate motions and party manifestos. Much of the previous work on analyzing political texts relies on hand-labeling documents (Abercrombie and Batista-Navarro, 2018; Gilardi et al., 2009; Krause, 2011; Simmons and Elkins, 2004). Thus, the analysis of political documents in this field stands to benefit from automating the coding of texts using supervised machine learning. Most recently, neural networks and deep language representation models have been employed in state-of-the-art approaches to automatic labeling of political texts by policy preferences.

In this paper, we present a deep learning approach to classifying labeled texts and phrases in party manifestos, using the coding scheme and documents from the Manifesto Project (Volkens et al., 2019). We use English-language texts from the Manifesto Project Corpus, which divides party manifestos into statements—or *quasi-sentences*—that do not span more than one grammatical sentence. Based on the state-of-the-art deep learning methods for text classification, we propose using Bidirectional Encoder Representations from Transformers (BERT) combined with neural networks to automate the task of labeling political texts. We compare our models that combine BERT and neural networks against previous experiments with similar architectures to establish that our proposed method outperforms other approaches commonly used in natural language processing research to predict policy domains and policy preferences. We identify differences in performance across policy domains, paving the way for future work on improving deep learning models for classifying political texts. To the best of our knowledge, we offer the most comprehensive application of deep language representation models incorporated with neural networks for document classification of political manifesto statements.

The rest of this paper is structured as follows. In Section 2, we provide a brief overview of the current state-of-the-art methods in the classification of political texts, focusing mainly on detecting policy domains and preferences. Section 3 goes into detail about the Manifesto Project Corpus. Section 4 then introduces our classification approach and provides important details of our models and evaluation approach. In Sections 5 and 6, we present our results and address some limitations of our system. Finally, Section 7 concludes our findings and presents a roadmap for future improvements.

2 Related Work

For the task of classifying political texts, many studies have concentrated on building scaling models for identifying the political positions of documents (Laver et al., 2003; Nanni et al., 2019; Proksch and Slapin, 2010). However, most of this seminal work in this area failed to consider the task of classifying texts by topic or policy area prior to detecting policy preferences associated with the topic. Over the past couple of years, several studies have addressed this gap in *opinion-topic identification* by classifying text data from political speeches, manifestos, and other documents by topic before predicting policy preferences (Glavaš et al., 2017; Zirn et al., 2016). With regards to party manifestos, the coding of policy preferences after dividing documents into topics could be expansive, pointing to the necessity of more complex models for text classification to take on this task. This is why recent studies have begun to utilize neural networks (Subramanian et al., 2018) and deep language representation models (Devlin et al., 2018) to address the computationally intensive task of classifying political texts into over thirty categories.

Against this background, this project closely follows the methods proposed by Abercrombie et al. (2019), who worked to detect the policy positions of UK Members of Parliament through natural language processing methods. Using motions and manifestos as data sources, the authors employed a variety of methods to predict the policy and domain labels of texts. Thereafter, they compared the predicted labels with the gold standard labels to produce F1 scores. For their proposed BERT model, Abercrombie et al. (2019) used a final softmax model and added CNN and max-pooling layers. Furthermore, they fine-tuned the results of the aforementioned BERT Model by training it first on the manifestos and then on the motions. The authors evaluated the predicted labels of each experimental model against the gold standard labels (i.e., when two annotators agree on the same labels) produced during the annotation process. Ultimately, they found that the use of BERT demonstrated 'state-of-the-art performance' on both manifestos and motions via supervised pipelines, with a Macro-F1 score of 0.69 for their best performing model, pointing to the effectiveness of this model in predicting policy preferences from political texts.

3 The Manifesto Project Corpus

The Manifesto Project Corpus[1] (Volkens et al., 2019) provides information on policy preferences of political parties from seven different countries based on a coding scheme of seven policy domains, under which 57 policy preference codes are manually coded. The Manifesto Project offers data that divides party manifestos into quasi-sentences, or individual statements which do not span more than one grammatical sentence. Quasi-sentences are then individually assigned to categories pertaining to policy domain and preference. The 57 policy preference codes refer to the position—positive or negative—of a party regarding a particular policy area. The 57 policy preference codes fall into a macro-level coding scheme comprising of eight policy domain categories[2].Hereafter, we refer to the policy preferences and policy domains as 'minor' and 'major' categories, respectively. In political science research, the Manifesto Project Corpus is particularly useful for studying party competition,

[1] manifesto-project.wzb.eu

[2] Each topic classification scheme includes a distinction for "non-categorized" texts

the responsiveness of political parties to constituent preferences, and estimating the ideological position of political elites. While the official classification of manifestos in this dataset has primarily relied on human coders, the investigation of automatically detecting policy positions of the text data is valuable for scaling up the classification of large volumes of political texts available for analysis.

Our final subset of all English-language manifestos comprises of 99,681 quasi-sentences. Tables 1 and 2 illustrate the distribution of English-language manifestos across countries and policy domains. To ensure that the ratio between policy domains remains consistent across policy domains in running our models, we applied a 70/15/15 split between training, validation, and test sets separately for the eight major categories and the 57 minor categories. Test and validation sets were sampled to have the identical class distribution of the training data.

Table 1: English language manifestos by policy domain

Topic	QSs	%
External Relations	6580	6.7
Freedom and Democracy	4700	4.8
Political System	10557	10.7
Economy	24757	25.2
Welfare and Quality of Life	30750	31.3
Fabric of Society	11099	11.3
Social Groups	9910	10.1

Note: Excludes “non-categorized” statements.

Table 2: English language manifestos by country

Country	QSs	%
United States	10819	10.9
South Africa	6423	6.5
New Zealand	28561	28.7
Ireland	25352	25.5
Great Britain	14839	14.9
Canada	3047	3.1
Australia	10370	10.4

4 Experimental Setup

BERT has proven successful in prior attempts to classify phrases and short texts (Devlin et al., 2018). We test two variants of BERT—one incorporating a bidirectional GRU model, and another incorporating CNNs. Between these two variants, we propose that BERT-CNNs are the state-of-the-art application of deep learning for classifying statements from political texts.

4.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT’s key innovation lies in its ability to apply bidirectional training of transformers to language modeling. This state-of-the-art deep language representation model uses a “masked language model”, enabling it to overcome restrictions caused by the unidirectional constraint. Our experiments use the standard pre-trained BERT transformers as the embedding layer in our model. We make use of the BERT BASE uncased tokenizer, with the following parameters:

$$\text{BERT}_{\text{BASE}}\text{: (L=12, H=768, A=12,}\\ \text{TotalParameters=110M)}$$

Since BERT is trained on sequences with a maximum length of 512 tokens, inclusive of start and end of sentence tokens, all quasi-sentences with more than 510 words were trimmed to fit this requirement. Pre-trained embeddings of the entire transformer body were frozen and not trained for the base models. We utilized the Hugging Face `transformers` library to run our BERT and other deep language representation models[3]. Model specifications and training times for our neural networks and deep language representation models are shown in Tables 3 and 4.

4.2 RoBERTa

The RoBERTa model was proposed by Liu et al. (2019) in a replication study that evaluates several approaches to augmenting the process of pre-training BERT models. The adjustments made to improve upon BERT include training the model longer, removing the model’s objective of predicting the next sentence, training on longer sequences of text, and changing the pattern of masking texts applied in the This masked language model improves on the performance of BERT models in several downstream tasks. In this research, we fine-tune RoBERTa with a simple linear classifier on top, using the RoBERTa BASE tokenizer.

4.3 BERT with Gated Recurrent Units (GRU)

First proposed by Cho et al. (2014), Gated Recurrent Units use update gates and reset gates to solve

[3] https://huggingface.co/transformers/

Models	Text Representation	Layers	Epochs
CNN	GloVe Wikipedia w-emb	2 Convolutional Layers (1 per filter) 2 Max Pooling Layers 1 Dropout Layer 1 Linear Layer	100
BERT	Base BERT (uncased)	1 Linear Layer	10
RoBERTa	Base RoBERTa	1 Linear Layer	10
BERT-CNN	Base BERT (uncased)	2 Convolutional Layers (1 per filter) 2 Max Pooling Layers 1 Dropout Layer 1 Linear Layer	10
BERT-GRU	Base BERT (uncased)	1 Bidirectional GRU RNN Layer 1 Dropout Layer 1 Linear Layer	10

Table 3: Model specifications of neural networks and deep language representation models

Table 4: Training time (in seconds) for neural networks and deep language representation models for classifying political texts by *major* and *minor* policy domain

Model	8 topics	57 topics
CNN	559	672
BERT	4123	3883
RoBERTa	4120	4110
BERT-CNN	2177	2085
BERT-GRU	2564	4820

the vanishing gradient problems often encountered in applications of recurrent neural networks (Kanai et al., 2017). The update gate helps the model determine the extent to which past information is carried on in the model, whilst the reset gate determines the information to be removed from the model (Chung et al., 2014). It solves the aforementioned problem by not completely removing the new input, instead keeping relevant information to pass on to further subsequent computed states. In our analysis, we employ a multi-layer, bidirectional GRU model from PyTorch[4]. The results are subject to a dropout layer prior to classification via a linear layer.

4.4 BERT with Convolutional Neural Networks (CNN)

We incorporate CNNs with BERT using the same CNN architecture as our baselines (Table 3). The model utilizes the aforementioned BERT base, uncased tokenizer with convolutional filters of sizes 2 and 3 applied with a ReLu activation function. We use a 1D-max pooling layer, a dropout layer ($N = 0.5$) to prevent overfitting, and a Cross Entropy Loss function. We employ the model to classify policy domains ($N = 8$) and policy preferences ($N = 57$), each of which includes a category for quasi-sentences that do not fall into this classification scheme. A graphical representation of our model is shown in Figure 1.

4.5 Evaluation

We evaluate the performance of our proposed method against several baselines, which include:

- **Multinomial Naive Bayes** (Eyheramendy et al., 2003): This algorithm, commonly used in text classification, operates on the *Bag of Words assumption* and the assumption of *Conditional independence*.

- **Support Vector Machines (SVM)** (Tong and Koller, 2001): We used this traditional binary classifier to calculate baselines with the `SVC` package from `scikit-learn`[5], employing a "one-against-one" approach for multi-class classification.

- **Convolutional Neural Networks (CNN)** (Kim, 2014; LeCun et al., 1998): To run this deep learning model, originally designed for image classification, we first made use of pre-trained word vectors trained by GloVe, an unsupervised learning algorithm for obtaining vector representations for words (Pennington et al., 2014)[6].

[4] https://pytorch.org/

[5] https://scikit-learn.org/stable/

[6] See Table 8 in the appendix for detailed information on pre-trained word embeddings.

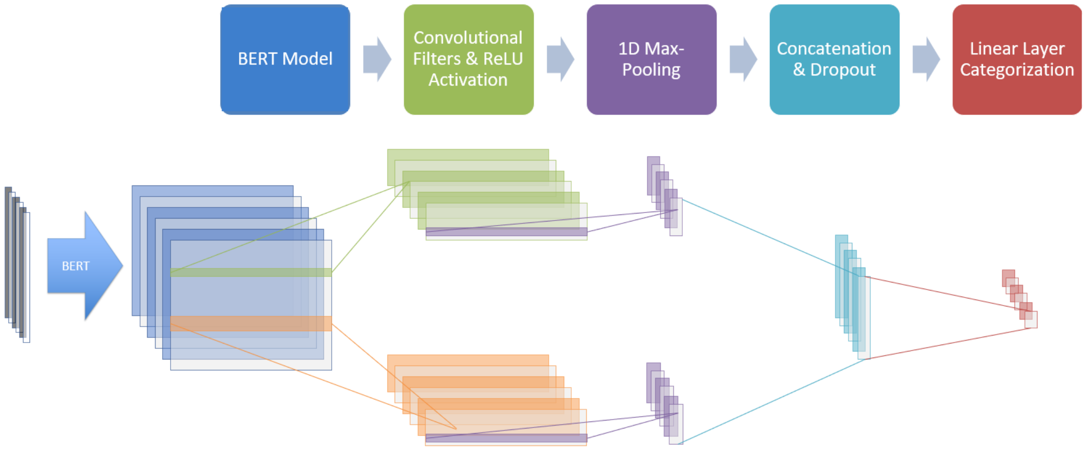

BERT Model
Convolutional Filters & ReLU Activation
1D Max-Pooling
Concatenation & Dropout
Linear Layer Categorization
BERT

Figure 1: Graphical representation of the base BERT-CNN model to predict major policy domains.

To evaluate model fit, we utilized *accuracy* and *loss* as key metrics to compare performance of our CNN and BERT-GRU baseline against the BERT-CNN model. We calculated the *F1-score* for each model that we ran. In our results, we present both the Macro-F1 and Micro-F1 scores[7].

4.6 Architecture fine tuning

We tested different modifications of the CNN and BERT models as a robustness check on the performance of our base model for the task of political text classification. For the CNN models, we compared our base model to the following modifications:

- **Stemming and Lemmatization**: We test whether stemming or lemmatizing text in the pre-processing steps improves predictions using quasi-sentences from the Manifesto Project Corpus.
- **Dropout rates**: We decreased the dropout rate from 0.5 to 0.25 to determine whether fine-tuning dropout rates yield differences in performance. This is because we initially found that our models were overfitting.
- **Additional linear layer**: An additional linear layer was added prior to the final categorization linear layer to establish whether "deeper" neural networks generate improved predictions.
- **Removal of uncategorized quasi-sentences**: The results from our base models yield lower Macro-F1 scores due to the difficulty of correctly categorizing quasi-sentences that do not fall into any of the eight policy domains or 57 policy preference codes. We are thus interested in whether predictions improve if the uncategorized quasi-sentences are taken out of the data used for analysis.

For the BERT models, we compared our base model to the following modifications:

- **Training Embeddings**: For our base BERT models, all training of embeddings were frozen. In this modification, we enable the training of the embeddings to establish how training embeddings contributes to the performance of deep language representation models with this classification task.
- **Training models based on recurrent runs**: We trialed training the BERT models sequentially with different learning rates (LR = 0.001, 0.0005 and 0.0001) of 10 epochs each for a total of 30 epochs in aims to improve the performance of our neural networks and deep language representation models.
- **Large, cased tokenizer**: The BERT Large cased tokenizer was used instead of the BERT BASE uncased tokenizer employed in our base models.

5 Results

As shown in Table 5, the BERT-CNN model performed best for predicting both major and minor categories compared to the BERT-GRU model and

[7] The micro score calculates metrics globally, whilst the macro score calculates metrics for each label and reports the unweighted mean.

Category	Model	Test Loss	Test Acc.	Micro-F1	Macro-F1
Major	MNB	—	0.553	0.553	0.398
	SVM	—	0.578	0.578	0.460
	CNN	1.177	0.589	0.589	0.466
	BERT	1.379	0.502	0.502	0.363
	RoBERTa	1.350	0.514	0.515	0.360
	BERT-GRU	1.166	0.594	0.593	0.479
	BERT-CNN	**1.152**	**0.591**	**0.591**	**0.473**
Minor	MNB	—	0.385	0.385	0.154
	SVM	—	**0.463**	**0.463**	**0.299**
	CNN	2.136	0.454	0.454	0.273
	BERT	2.457	0.376	0.376	0.177
	RoBERTa	2.621	0.354	0.354	0.136
	BERT-GRU	2.216	0.432	0.432	0.239
	BERT-CNN	**2.098**	0.448	0.448	0.260

Table 5: Baseline, CNN and masked language models run with base model specifications as detailed in Table 3

CNN baseline. However, our SVM baseline outperformed the neural network models for predicting minor categories. We believe that the shortcomings of our neural networks and deep language representation models for this text classification task are due to computational limitations in specifying the number of epochs in training. We also observed overfitting in our models. For instance, Figure 3 illustrates that training accuracy of our CNN model increased at the cost of validation accuracy. However, this was not the case for deep language representation models classifying texts by minor categories. Overall, our results demonstrate that, between the two BERT models, the BERT-CNN model demonstrates superior performance against bag-of-words approaches and other models that utilize neural networks.

CNN and BERT Modifications

Comparing modifications to our CNN models, our results suggest that the base model outperforms most alternative model specifications. As outlined in Table 6, reducing the dropout rate to 0.25 improved the model on some indicators marginally. As expected, the removal of uncategorized quasi-sentences yielded improvements in predictions, with a significantly higher Macro-F1 score compared to other model specifications. Based on these results, future work should focus on how model predictions of uncategorized quasi-sentences can be improved, given their random nature.

While we observed some improvements with modifications to the CNN model, we find that our base BERT models performed best compared to other fine-tuned modifications to model architecture. The results of our base BERT model and alternative model specifications are shown in Table 7. Even though it is possible that our base BERT model is best for this classification model, our results could also indicate the presence of over-fitting or the lack of sufficient training available given the low number of epochs.

6 Limitations and Analysis

As shown in Figure 2, we observed overfitting with our major policy domain classification models. Despite employing changes and modifications to our models, including varied dropout rates, architecture fine-tuning and different learning rates, we did not find any variants of the models employed in analysis that would yield significant improvements in performance. We posit that potential improvements on these issues could be resolved by employing transfer learning and appending our sample of English-language manifestos with other political documents, such as debate transcripts.

In contrast, as shown in Figure 3, we observed underfitting in some of our minor policy domain classification models. Our classifier could benefit from employing transfer learning and appending our sample of manifesto quasi-sentences with other political texts, especially for policy domains with relatively fewer quasi-sentences to train on. It is also important to note that, compared to the more computationally intensive neural networks and deep language representation models, our Multi-

Model	Change	Test Loss	Test Acc.	Micro-F1	Macro-F1	Epochs
CNN	Base model	1.177	**0.589**	**0.589**	0.466	100
	Lemmatized text	**1.174**	0.585	0.585	0.460	100
	Stemmed text	1.213	0.577	0.576	0.448	100
	Dropout = 0.25	1.177	**0.589**	0.588	**0.467**	100
	Additional layer	1.180	0.586	0.586	0.462	100
	Removing uncategorized QSs	**1.136**	**0.596**	**0.595**	**0.535**	100

Table 6: Comparing results of modifications to CNN base models for predicting major policy domains

Model	Change	Test Loss	Test Acc.	Micro-F1	Macro-F1	Epochs
BERT-GRU	Base model	**1.152**	**0.594**	**0.593**	**0.479**	10
	Training emb	1.163	0.592	0.592	**0.479**	10
	Recurrent runs, training	1.234	0.582	0.581	0.459	30
	Large, uncased	1.172	0.592	0.591	0.469	10
BERT-CNN	Base model	1.166	**0.591**	**0.591**	**0.473**	10
	Training emb	1.167	0.587	0.587	0.458	10
	Recurrent runs, training	**1.157**	0.589	0.589	0.468	30
	Large, uncased	1.192	0.580	0.580	0.450	10

Table 7: Comparing results of modifications to BERT base models for predicting major policy domains

nomial Bayes and SVM baselines did not perform significantly worse. In fact, for the minor categories, the SVM yielded superior performance in some metrics compared to that of the neural network models. Notwithstanding the lack of training of certain models, this may suggest that increasing the model complexity and consequently the computational power required may not necessarily lead to increased model performance.

Substantially lower Macro-F1 scores across all models point to mixed performance in classification by category. As shown in Figure 4, we observe high variation in the performance of our classifiers between categories. However, we observe poor performance in classifying quasi-sentences that do not belong to one of the seven policy domains. For our BERT-CNN model, the easiest categories to predict were "welfare and quality of life", "economy", and "external relations". The superior performance of predicting the first two categories is not particularly surprising, as a substantial number of quasi-sentences in our sample of English-language party manifestos are attributed to these topics. As shown in Table 1, 30,750 quasi-sentences are attributed to the "welfare and quality of life" category and 24,757 quasi-sentences are attributed to the "economy" domain.

In contrast, the relatively superior performance of predicting the "external relations" category is surprising. Out of our total sample of $n_{\text{sentences}} = 99,681$, only $6,580$ documents are attributed to this category[8]. The performance of our classifier with this underrepresented policy domain could be attributed to a variety of possible explanations. One possible explanation is the presence of distinct features, such as topic-unique terms, that do not exist in other categories. Future work on classification of political documents that fall under this category would benefit from looking into features that might establish which policy domains perform better than others with the BERT-CNN classifier.

7 Conclusion

In this paper, we trained two variants of BERT—one incorporating a bidirectional GRU model, and another incorporating CNNs. We demonstrate the superior performance of deep language representation models combined with neural networks to classify political domains and preferences in the Manifesto Project. Our proposed method of incorporating BERT with neural networks for classifying English language manifestos addresses issues of reproducibility and scalability in labeling large volumes of political texts. As far as we know, this is the most comprehensive application of deep language

[8] Some of the policy preferences coded under "External Relations" include foreign special relationships, anti-imperialism, peace, military, internationalism, and European community/union.

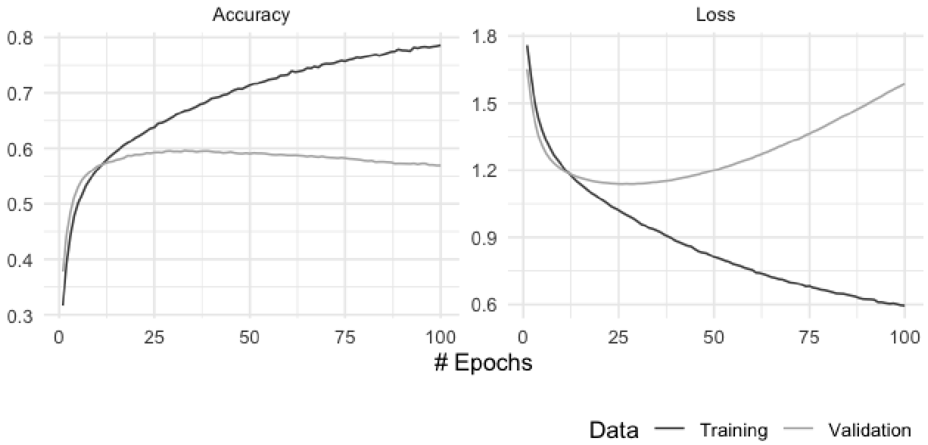


Figure 2: An illustration of overfitting in our CNN model for classifying manifesto quasi-sentences by major policy domain

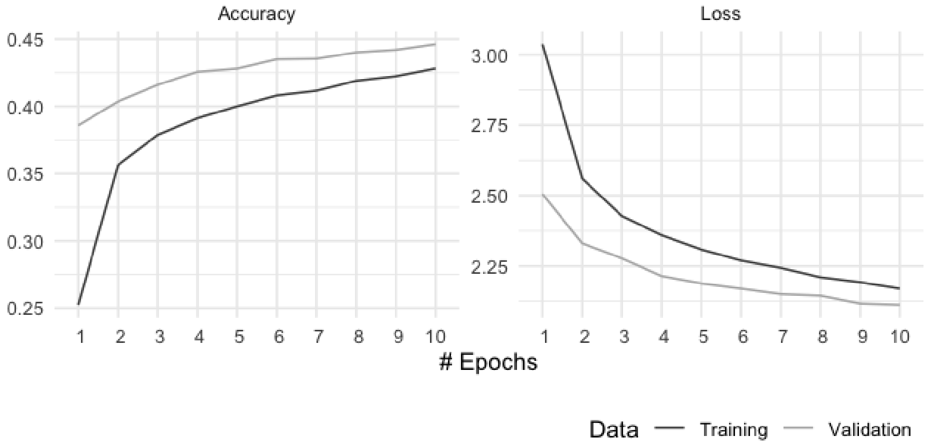


Figure 3: Training and validation metrics for the BERT-CNN model on English language manifestos on minor policy domains

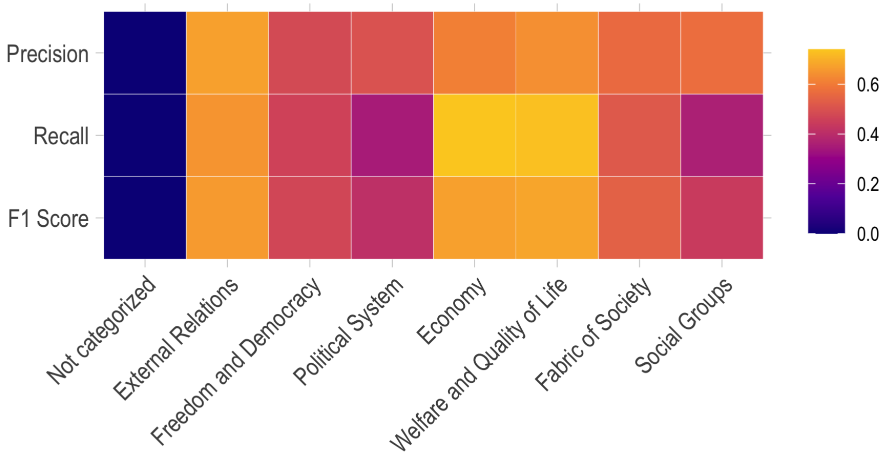


Figure 4: Average precision, recall, and Macro-F1 scores by major category across all models

representation models and neural networks for classifying statements from political manifestos.

We find that using BERT in conjunction with Convolutional Neural Networks yields the best predictions for classifying English language statements parsed from party manifestos. However, our proposed BERT-CNN model requires further finetuning to be effective in providing acceptable predictions to improve on less computationally intensive classifiers of fine-grained policy positions. As expected, our proposed approach and baselines perform better for classifying major policy domains over minor categories. We also observe differences in performance between categories. Among the major policy domains, the categories that performed best were "welfare and quality of life", "economy", and "external relations". The superior performance of the latter category is surprising because it makes up a relatively small proportion of quasi-sentences in the Manifesto Project Corpus.

There are several avenues for future work on neural networks and deep language representation models for the automatic labeling of political texts. For instance, investigating the features of individual categories that demonstrate superior performance could shed light on how we could incorporate additional features of texts to improve model performance. This area of research would also benefit from better understanding how we can filter out texts that do not fall into a particular classification scheme. Knowledge on how these issues could be resolved to improve model performance would allow for extensions in the application of deep learning models to the classification of political texts.

References

Gavin Abercrombie and Riza Batista-Navarro. 2018. A sentiment-labelled corpus of hansard parliamentary debate speeches. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Ed. by Darja Fišer, Maria Eskevich, and Franciska de Jong. Miyazaki, Japan: European Language Resources Association (ELRA).*

Gavin Abercrombie, Federico Nanni, Riza Theresa Batista-Navarro, and Simone Paolo Ponzetto. 2019. Policy preference detection in parliamentary debate motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259.

Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tanenbaum, et al. 2001. *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*, volume 1. Oxford University Press on Demand.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Marc Debus. 2009. *Estimating the Policy Preferences of Political Actors in Germany and Europe: Methodological Advances and Empirical Applications*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Susana Eyheramendy, David D Lewis, and David Madigan. 2003. On the naive bayes model for text categorization. In *International workshop on artificial intelligence and statistics*, pages 93–100. PMLR.

Fabrizio Gilardi, Katharina Füglister, and Stéphane Luyet. 2009. Learning from others: The diffusion of hospital financing reforms in oecd countries. *Comparative Political Studies*, 42(4):549–573.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. Association for Computational Linguistics (ACL).

Sekitoshi Kanai, Yasuhiro Fujiwara, and Sotetsu Iwamura. 2017. Preventing gradient explosions in gated recurrent units. In *Advances in neural information processing systems*, pages 435–444.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Rachel M Krause. 2011. Policy innovation, intergovernmental relations, and the adoption of climate protection initiatives by us cities. *Journal of urban affairs*, 33(1):45–60.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155.

Federico Nanni, Goran Glavas, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2019. Political text scaling meets computational semantics. *arXiv preprint arXiv:1904.06217*.

Federico Nanni, Cäcilia Zirn, Goran Glavaš, Jason Eichorst, and Simone Paolo Ponzetto. 2016. Topfish: topic-based analysis of political position in us electoral campaigns.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543.

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

Beth A Simmons and Zachary Elkins. 2004. The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review*, pages 171–189.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. *arXiv preprint arXiv:1805.02823*.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Andrea Volkens, Judith Bara, Ian Budge, Michael D McDonald, and Hans-Dieter Klingemann. 2013. *Mapping policy preferences from texts: statistical solutions for manifesto analysts*, volume 3. OUP Oxford.

Andrea Volkens, Onawa Lacewell, Pola Lehmann, Sven Regel, Henrike Schultze, and Annika Werner. 2019. The manifesto data collection. manifesto project (mrg/cmp/marpor). *Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB)*.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos.

A Additional Information on Baselines

For the first two methods, the Multinomial Naive Bayes Model and the Support Vector Machines, the `TfidfVectorizer` from `sklearn` was employed. This method makes use of term frequency - inverse document frequency weighting to remove terms that are present commonly but carry very little information (e.g. stopwords).

A.1 Multinomial Naive Bayes Model

As a baseline, we used a multinomial naive Bayes algorithm, commonly used in text classification. The assumptions of this model includes:

- Bag of Words: Position does not matter
- Conditional Independence: Feature probabilities are independent given the class.

The Naive Bayes Model is quick and provides a baseline for the other classification techniques.

A.2 Support Vector Machines

Support Vector Machines (SVM) seek the most optimal decision boundaries by creating hyperplanes that separate the training data (Tong and Koller, 2001). The aim of the separating hyperplane or the set of hyperplanes is to maximise the distance between the nearest training data points of any class (i.e. functional margin). Whilst SVMs are traditionally binary classifiers, `scikit-learn`'s package `SVC` employs a "one-against-one" approach for multi-class classification. Where a SVM is trained based on data from two classes and repeated for each relationship with each other class present. Since there are eight policy domains (including unclassified), there will be 28 distinct SVMs created.

We trained SVMs on both datasets with four different kernels:

- Linear kernel: $< x, x' >$
- Polynomial kernel: $(\gamma < x, x' > +r))^d$
- Radial basis function kernel: $(\gamma||x - x'||^2)$
- Sigmoid kernel: $(\tanh(\gamma < x, x' > +r))$

	GloVe 6B
Tokens	6 billion
Dimension	300
Vocabulary size	400 thousand
Cased?	No

Table 8: Details of GloVe pre-trained vectors utilized

A.3 Convolutional Neural Networks

Convolutional Neural Networks are neural networks that utilize layers that contain convolving filters that help to aggregate data into multiple layers (LeCun et al., 1998). Whilst it was originally designed for image classification, it has also been utilized for Natural Language Processing purposes - semantic parsing, sentence modeling, sentence classification, etc (Kim, 2014).

In our model, we first made use of pre-trained word vectors trained by GloVe, an unsupervised learning algorithm for obtaining vector representations of words(Pennington et al., 2014). Specifically, we chose pre-trained vectors trained on a corpus of 1.6 billion tokens from a 2014 Wikipedia dump.

Filter-sizes of 2 and 3 were used with 100 2D convolutional filters each. After a single convolutional layer per filter size, each of the layers are fed into the soft-max activation functions. Thereafter, a single max-pooling layer was utilized per filter. The outputs from the corresponding max-pooling layers were then concatenated and passed through a dropout layer. Lastly, the results were passe d through a linear layer to predict the different classifications.

In all models (the current and the following models), the Adam Optimizer was utilized with a Cross Entropy Loss function. The latter is a combination of a logistic softmax and a negative log likelihood loss functions, useful for classification problems with multiple classes.

UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus

Luis Glaser
University of Potsdam
Potsdam, Germany
luis.glaser@uni-potsdam.de

Ronny Patz
Hertie School
Berlin, Germany
patz@hertie-school.org

Manfred Stede
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

We present the Named Entity (NE) add-on to the previously published United Nations Security Council (UNSC) Debates corpus (Schoenfeld et al., 2019). Starting from the argument that the annotated classes in Named Entity Recognition (NER) pipelines offer a tagset that is too limited for relevant research questions in political science, we employ Named Entity Linking (NEL), using DBpedia-spotlight to produce the UNSC-NE corpus add-on. The validity of the tagging and the potential for future research are then discussed in the context of UNSC debates on Women, Peace and Security (WPS).

1 Introduction & Motivation

There is a growing interest in research questions at the intersection of political science, its subfield focused on international relations, and Natural Language Processing (NLP). New diplomatic speech corpora are being created to understand state preferences through correspondence analysis (Baturo et al., 2017), discursive landscapes through topic modeling (Schoenfeld et al., 2018) or inter-state agreement in international negotiations through linguistic style matching (Bayram and Ta, 2019).

Building on the long-established understanding that linguistic choices are central to the legitimising work of international institutions (Claude, 1966), and that states make deliberate choices about what they say—and what they do not say—in diplomatic fora to shape the global order (Schmitt, 2020), a central methodological question is how to make use of the growing NLP toolbox to study such choices on a large scale.

In this contribution, we start from the assumption that one important choice states make is what entities and concepts they mention—or ignore mentioning—in their diplomatic speeches. Mentioning one conflict location over another may hint at states' specific political attention. Pointing to a single conflict party instead of all of them in a speech could indicate a more partisan rather than a diplomatic approach. Failing to reference an international convention or a particular UN resolution, and choosing one concept from international law over another, can be speakers' deliberate attempts to frame a multilateral debate in one direction, for instance by shifting attention from human rights to states' rights for non-interference in their internal affairs.

However, automatically recognizing entities, including the correct entity classes, in diplomatic speech is non-trivial. Various out-of-the-box tools for NER exist but have not yet been extensively applied and validated for the existing diplomatic speech corpora. We therefore present the UNSC Debates Corpus NEL add-on, an entity-tagged extension to the UN Security Council debates corpus that was previously published by Schoenfeld et al. (2019).

After introducing recent research in political science using NER, and discussing why we choose NEL over NER, we explain the technical and conceptual basis for NEL and the Resource Description Framework (RDF), compare the quality of annotations of DBpedia-spotlight to `spaCy` (Honnibal et al., 2020), and then present the corpus format. We further demonstrate the potential of the corpus add-on in an experiment looking at what entities the five permanent members (P5) of the UNSC (China, France, Russia, the United Kingdom and the United States) mention in UNSC debates on the agenda item of Women, Peace and Security. This is discussed in relation to previous political science research that has identified important differences between the P5 on this agenda item. The resulting is corpus publicly available under CC0 license.[1]

2 Background: NER and NEL

Both NER and NEL try to find NEs in natural language text, but differ in the way these NEs are extracted and represented. NEs are words or phrases that refer to an entity in the real world, roughly equivalent to a proper noun (Jurafsky and Martin, 2018). NER tries to detect NEs in natural language and assigns a class from a predefined set of classes.[2] NER can also disambiguate between different NEs, e.g. “Washington” could refer to a person, a location or a global political entity.

NEL on the other hand tries to detect NEs in natural language that refer to an entity within a knowledge graph. These entities are represented by unique identifiers that describe real world entities or abstract concepts. Within these knowledge graphs, additional information is linked to the unique entities, e.g. a node with the label “Washington” may be an instance of a city, while another distinct node with the label “Washington” might be an instance of a state.

2.1 NEs in Political Science

NER is a recent addition to the toolbox of political science research, with political scientists increasingly turning towards deep learning (Chatsiou and Mikhaylov, 2020).

However, applications of NER published in political science journals are still rare. Most existing contributions focus on geographical locations (Nardulli et al., 2015), demonstrating how geolocated event data using NER can be used to identify places of conflict or protest (Lee et al., 2019). Geolocation is also applied by Fernandes et al. (2020) to understand how policy makers in Portugal reference their own or distant constituencies in their speeches. A more recent application uses NER to identify the appearance of interest groups in a UK news corpus of $3,000$ stories, and finds that the off-the-shelf tool *analyzeEntities* was able to find 54% of entities identified by expert human coders (Aizenberg and Binderkrantz, 2021). An additional novel contribution comes from the NLP community: Kerkvliet et al. (2020) use spaCy to identify political actors in a Dutch speech corpus by combining the off-the-shelf model with additional training material.

Peer-reviewed applications of NER to diplomatic speech and documents are so far mainly limited to the UN General Debate corpus (Baturo et al., 2017). Gray and Baturo (2021) study the specificity of different speakers in these debates by calculating shares of recognised named entities over all terms in a speech. However, there are indications that NER-tagged corpora will become more frequent: the recently presented PeaceKeeping Operations Corpus (PKOC) comes with an additional tagged version (tPKOC), using the Stanford CoreNLP Toolkit for NER (Amicarelli and Di Salvatore, 2021). Understanding the accuracy (resp. precision and recall) and relevance of different NER tools will therefore become increasingly important for political science and international relations research. There is also an increasing need to discuss the diverse fields of potential application of NER: from measuring conflict between speakers by the difference in NE references in their speeches to speakers’ geographical or topic focus based on NEs, from shifts in attention or meaning over time to the different use of NEs or NE classes. Many different research questions at the intersection of NLP and political science can be asked but also require further exploration.

2.2 Named Entity Linking

This section explains what NEL provides and why we consider it to be a powerful alternative to NER for use in political science. As previously outlined, researchers have turned to NER when examining NEs in their work. We argue that NER systems can have a strong limitation depending on the intended use. Due to the limited number of potential annotation classes in NER, concepts are conflated, where political scientists would demand a finer disambiguation. For example “United Nations Security Council”, “European Union” and “Bundestag” are all tagged as Organization (ORG) by the spaCy NER-pipeline. This may be an acceptable limitation in some use cases, e.g. review classification or identifying locations, but for using NEs in political science, more fine-grained NE annotations are required to broaden the scope of possible analyses. We therefore suggest to use NEL instead of NER as a potential improvement. Instead of tagging an

[1] Accessible at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OV1FLX

[2] E.g. the latest NER tagset spaCy uses has the following entity labels: Person, Nationalities or religious or political groups, Organization, Global Political Entity, Location, Product, Event, Work of Art, Language, Date Time Percent, Money, Quantity, Ordinal, Cardinal.

NE with a class it belongs to, e.g. “United Nations” as an ORG, each NE is referenced by a specific Unique Resource Identifier (URI) that denotes a singular entity represented in a knowledge graph. It still allows researchers to summarize the United Nations as an instance of the class organization, as an NER tagger would. But because the annotation is not a shallow tagging but a linking to a URI, the granularity of an analysis can be altered as needed.

An NEL pipeline may annotate any entity that exists in the knowledge graph it is trained on. Thus, choosing a different knowledge graph as the foundation of an NEL tagger will lead to different annotations. In many cases however entities in different knowledge graphs are linked between each other in order to make them inter-operable. In the case of the two knowledge graphs we used for this work, DBpedia and Wikidata, URIs that refer to the same NE in both graphs are linked via the `owl:sameAs`[3] property.

2.3 Representing NEs in Knowledge Graphs

RDF provides a formalism to represent data as statements called *triples*. These triples are comparable to natural language statements, as they consist of a subject, a predicate and an object. We can group a number of triples to form a *knowledge graph*, also called a *document*. Each part of a triple (subject, predicate and object) may be a URI (Cimiano et al., 2020). These URIs can represent entities that are only defined within the knowledge graph it is a part of. However, they may also refer to external resources, e.g. an entry in Wikidata. That way, information can be stored in a distributed way. Also, information that once was linked to a URI can be enhanced and brought into context by querying the external resources that refer to this URI.

Consider the statement “The UNSC is a council”. We can represent this in form of a triple `ex:unsc ex:is-instance-of ex:council`. Using a second triple, we can link the first to an external resources, in this case Wikidata: `ex:unsc owl:sameAs wd:Q37470`. Now, we can query Wikidata for information on `wd:Q37470`. That way, partial information that is available locally can be enhanced by information that is available externally.

2.4 Comparing DBpedia to Wikidata

DBpedia and Wikidata are both publicly available knowledge graphs. They differ in their conceptual basis, scope and aim. The DBpedia project uses Wikipedia as its data foundation and extracts the contained links, info boxes and texts in order to create a knowledge graph. The Wikidata project on the other hand contains systematically created entities in its knowledge graph, which may be linked and annotated automatically or by a human. Wikidata can be understood as a *top-down* approach, while DBpedia works *bottom-up*. Because entries in DBpedia contain a larger amount of natural language data by design, it is better suited to train an automatic classifier on its basis, namely DBpedia-spotlight. Wikidata however offers a more fine-grained ontology. Thus, we decided to use the DBpedia-spotlight service as an annotation basis and then automatically link the correspondent Wikidata entries to each annotation. We also considered alternatives to DBpedia-spotlight. `spaCy` offers NEL integration, but does not offer pretrained models yet. Thus, using DBpedia-spotlight directly was preferred. TAGME (Ferragina and Scaiella, 2010) resp. WAT (Piccinno and Ferragina, 2014) solve a similar problem, however the ability to run DBpedia-spotlight on a local machine without ratelimits allowed us to prototype faster and speedup the annotation process itself. Also neural approaches like Kolitsas et al. (2018) could improve the corpus quality. This would have required to procure our own knowledge base, which can be considered in future release but was beyond the scope of the first corpus add-on.

3 Creating the UNSC-NE Add-on

3.1 The UNSC Corpus

The data set this work is based on is the UNSC Debates corpus published by Schoenfeld et al. (2019).[4] It contains all meeting transcripts of the UNSC from 1995 to 2020. The corpus consists of $82,165$ speeches extracted from $5,748$ meeting protocols. Speeches are annotated with their speakers, country affiliations and other information, such as the agenda item. This information is transferred

[3] This paper uses the turtle format to represent triples, which allows abbreviations of URIs. In this document `http://example.org/` is abbreviated as `ex:`, `http://www.w3.org/2002/07/owl#` as `owl:` and `http://www.wikidata.org/entity/` as `wd:`

[4] In this paper we refer to version 5 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KGVSYH

to the UNSC-NE add-on and can be used as a link between both the corpus and its add-on.

3.2 Cleaning, Annotating & Linking

In order to annotate the UNSC corpus with named entities, we did the following: we first removed process descriptions, that did not contain actual speech but described events during the speech itself (e.g. "(The speaker spoke in Spanish)") from documents using regular expressions. Using a locally running DBpedia-spotlight instance, we then extracted all linked entities with the default confidence of $> .5$. To increase the available context, each call to DBpedia-spotlight contained an entire paragraph. The sentences were split up again afterwards and the offsets were fixed accordingly. In order to link these DBpedia entities to Wikidata, we used the `owl:sameAs` property of the DBpedia entry, if available. If not, we queried the GlobalFactSync (Hellmann et al., 2020) service in order to retrieve the corresponding Wikidata URL. This approach can lead to errors, because a DBpedia entry might be linked to multiple Wikidata entries if the term is rather broad or if the links are false themselves. In order to arrive at a 1-1 mapping between DBpedia and Wikidata, we compared the labels of both DBpedia and Wikidata to select the one that matched exactly. After that, for each entity linked to Wikidata, we retrieved the class linked with the relation *is instance of* (`wd:P31`). Furthermore, we extracted all superclasses via the relation *subclass of* (`wd:P279`).

Note that the labels instance, class and superclass which we use are not inherent to a node in Wikidata, but depend on the relation it has to others. E.g. in an utterance, we might find the entities "Syria" and "country". Within the knowledge graph, "Syria" is an instance of "country". Either may occur in text. The relations simply allow users to combine different entities together in their research. Thus, the UNSC-NE corpus add-on makes no distinction between them in their representation, they are all referred to as `WDConcepts` in the corpus. In order to keep UNSC-NE in sync with the underlying UN Security Council debates corpus, we provide build scripts online with which one may recreate the NEL annotations with minimal manual work.[5]

	spaCy	DBpedia
Accuracy	.478	.405
Precision	.503	.444
Recall	.904	.821
F1	.647	.576

Table 1: Comparison of annotation quality metrics between `spaCy` and `DBpedia`

3.3 Quality comparison of NER and NEL

We validated the quality of the DBpedia-spotlight NEL pipeline for our use-case compared to the most-prominent off-the-shelf solution that has seen previous usage in the field: `spaCy`.[6] We randomly sampled 20 speeches from the UNSC corpus and marked each span that we considered an entity relevant to the field manually. Then, we ran the sample through the `spaCy` NER and DBpedia-spotlight NEL pipeline. Because both approaches differ in what they annotate, we were only able to compare NE recognition, not whether the annotated classes or linked entities were correct themselves. The computed quality metrics are presented in Table 1. DBpedia-spotlight performs significantly worse compared to `spaCy` in all categories. This can be explained by the relatively harder task that NEL tries to solve, as it is not limited to a small number of classes but all entities present in a knowledge graph. However, depending on the usage scenario, this can be remedied by filtering for distinct classes, as will be shown in the experiments. Also, the gain of having Wikidata entities directly annotated in a more fine-grained manner may justify the cost in many cases.

4 The UNSC-NE Addon

4.1 Descriptives

After cleaning, the corpus contains $1,921,352$ sentences. Performing NEL on the UNSC corpus yielded $2,377,371$ entities in total, with $29,897$ distinct entities. Of these distinct entities, $28,776$ were linkable to wikidata either directly via the `owl:sameAs` property or via the GlobalFactSync project. These Wikidata entities are instances of $4,907$ distinct classes which in turn are subclasses of $10,989$ superclasses.

[5] Available at https://github.com/glaserL/unsc-ne.

[6] We used `spaCy` version 3.0 with the `en_core_web_sm` language model.

4.2 Format

The UNSC-NE corpus add-on is distributed in jsonlines format online. jsonlines (.jsonl) is a file format that contains a valid json value on each line. That makes it more easily streamable. We also distribute the corpus as a simple neo4j dump, that can be loaded into a neo4j graph database using the admin tool. Conceptually, UNSC-NE is a graph consisting of nodes and relationships between them. Each json object either represents a node or a relationship between two nodes. Nodes are identified with an id, have one or multiple labels and may have properties in form of a dictionary. Relationships are identified with their own id and the ids of the two nodes that are connected. Relationships may also contain properties in form of a dictionary.

4.3 Nodes

The following list shows the different node types the UN Security Council debates NEL add-on contains. We also provide a small explanation of each property that a node has. The two node types *Meta* and *Speaker* can be used as links to the foundational corpus.

- AgendaItem
    - name: the name of the agenda item
- Country
    - name: the name of the country
- DBConcept
    - uri: the DBpedia uri this node represents
- Institution
    - name: the name of the institution
- Meta *Represents an entry in* `meta.tsv` *of the fundamental UN Security Council debates corpus*
- Paragraph
    - index: the index within the speech it's contained in
- Sentence
    - index_in_speech: the index within the speech it's contained in
    - index: the index within the paragraph it's contained in
    - text: the text of the sentence itself

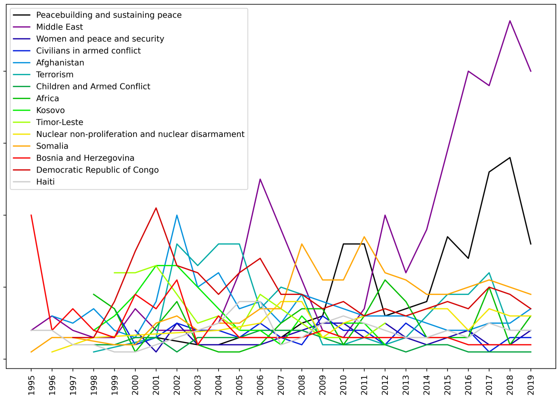


Figure 1: Top 15 agenda items of UNSC meetings

- Speaker
  Represents an entry in `speaker.tsv` *of the fundamental UN Security Council debates corpus*
- Speech
- WDConcept
    - uri: the Wikidata URI this node represents
    - label: the English string label of this node (taken from property `rdfs:label`)

4.4 Relationships

The following list contains all relationship that link the nodes above with each other. If a relationship has properties, these are also enumerated and explained shortly.

- AGENDA
    - Speech → AgendaItem
- CONTAINS:
    - Speech → Paragraph
    - Speech → Sentence
    - Paragraph → Sentence
- HAS_METADATA
    - Speech → Meta
- MENTIONS
    - Sentence → DBConcept
- NEXT
    - Sentence → Sentence

  - Speech → Speech
  - Paragraph → Paragraph

- owl_sameAs: links a URI in the DBpedia knowledge graph to a URI in the wikidata knowledge graph it corresponds to
  - DBConcept ↔ WDConcept

- wd_P279: points from a class to a superclass
  - WDConcept → WDConcept

- wd_P31: points from an instance to a class
  - WDConcept → WDConcept
  - surfaceForm: the string that has been annotated
  - offset: the character offset within the sentence

- REPRESENTS
  - Speaker → Institution
  - Speaker → Country

- SPOKE
  - Speaker → Speech
  - Speaker → Paragraph
  - Speaker → Sentence

5 Experiment: The WPS debates in the UNSC

To show the potential usages of the UNSC-NE corpus add-on, we performed an exemplary experiment on the data. While not an extensive exploration of the corpus, this experiment points to potential use cases for the corpus extension and confirms the substantive validity of the entity tagging in the context of existing political science research on the UNSC. We demonstrate in particular that NEL has the potential to detect meaningful similarities and differences in what kinds of entities, or classes of entities, representatives of the UNSC members address or fail to address.

Each meeting (and thus each speech) in the original corpus is linked to a single agenda item. Figure 1 shows the 15 agenda items that are most prominent in the UN Security Council debates corpus. This information is provided by the UN Security Council Debates corpus metadata. For this experiment, we focus on speeches of the P5 members in debates on the WPS agenda item, which emerged out of UNSC Resolution 1325 on Women, Peace and Security adopted in 2000. While not the most frequent agenda item, we select WPS for its relevance in political science research.

This research has focused on various questions, for example how the WPS agenda has evolved over time and how Resolution 1325 has been mainstreamed into other UNSC agenda items (Schoenfeld et al., 2018) or into UN peacekeeping practices (Kreft, 2017). Accurately identifying relevant NEs under the WPS agenda item could be a starting point for understanding mainstreaming across the corpus and in further UNSC agenda items.

To focus on the most relevant speeches, and to make the visualization of NEs more readable, we only consider NEs in the interventions by representatives of the P5, ignoring speeches of the UNSC presidency even when the presidency is held by one of the P5.

Figure 2 shows the distribution of the top 25 entities used most frequently by the P5 in their speeches during meetings with the WPS agenda item. The entity labels are drawn from Wikidata via DBpedia. The y-axis represents the shares of the respective NE references relative to all entities mentioned by each P5 country during those debates.

A first observation is that some very frequent NEs such as the more conceptual "sexual violence" or the more organizational references to the "United Nations" and "United Nations Secretary-General" have relatively similar shares among the P5. These terms are therefore not indicative of strategic NE use where the P5 differ.

In contrast, China and Russia refer more frequently to other UN entities such as the "United Nations Security Council" and the "United Nations General Assembly" than France, UK, or the US. This is in line with existing research on the WPS debates (True and Wiener, 2019) showing that China and Russia want to limit the policy scope of what is discussed in the UNSC debates on WPS. This is why they like to point to the competencies of the "General Assembly" and other bodies for issues that they do not consider covered in UNSC Resolution 1325. This is also likely why Russia refers most frequently to the NE identifying this particular resolution. China talks most frequently about the conceptual NEs "peacebuilding", "conflict resolution", "peacekeeping" or "terrorism", indicating that it sees the WPS agenda most relevant in these

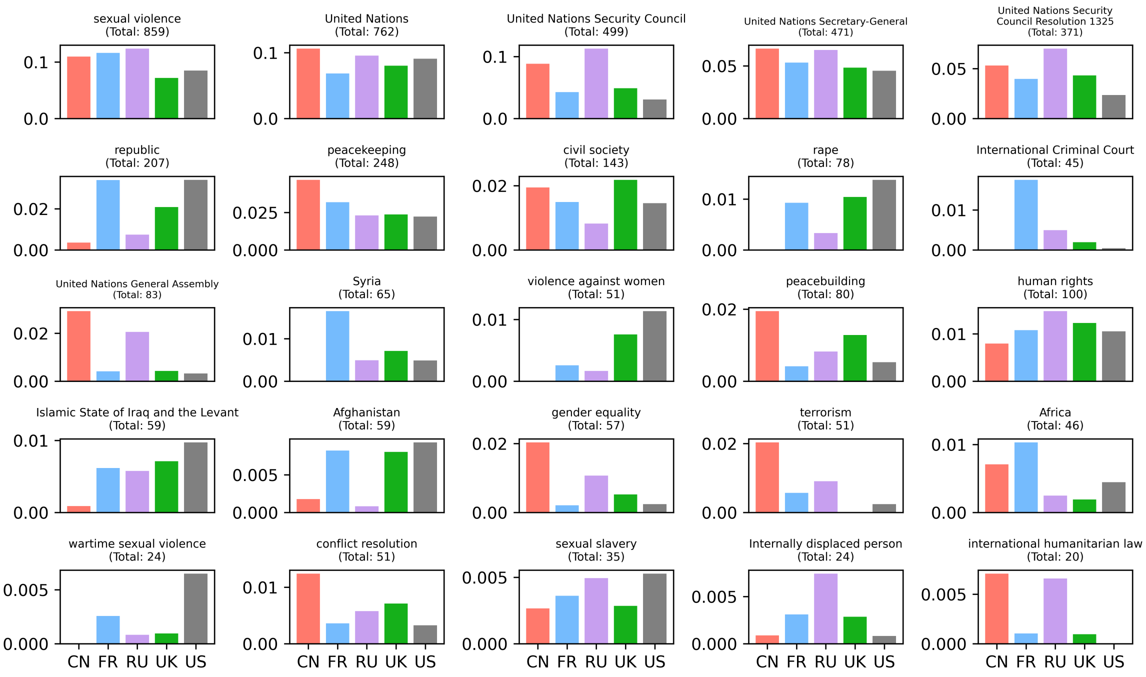


Figure 2: Most frequently used NEs by P5 countries in WPS debates

contexts, i.e. areas that are narrowly in the UNSC's realm. In contrast to the other P5 members, France highlights the (potential) role of the "International Criminial Court" in the context of crimes related to conflict-related sexual violence.

Using DBpedia for NEL allows the detection of more conceptual or policy-related entities, which provides insights into differences in legal and political framing of WPS debates by the P5. As discussed in international law (Macfarlane, 2021), there is a difference between the concepts of conflict-related "sexual violence" (the most frequent NE used by all P5) or terms such as "wartime sexual violence" (used mainly by the US but not China) or the more narrow but more concrete crime of "rape" (used more frequently by the US, UK and France than by Russia and not used by China). Detecting similarities and differences in such conceptual or policy NEs can be indicative of how consensual or contested certain legal or political terms are.

Finally, the NEL tagger also recognizes politico-geographic entities. In the WPS debates, the most frequently NEs of this class are countries (e.g. "Syria") or continents ("Africa") mentioned at different frequencies by different speakers. This is relevant because the WPS debates are not linked to any particular country or region, so P5 speakers reveal their particular geographical attention by making the choice to highlight some conflict zones and ignoring others. While China rarely speaks about concrete countries, it highlights "Africa", a continent it has focused its foreign and development policy on, France highlights "Syria" and the "DR of the Congo", two countries where it has been present militarily, but also "Africa", where, due to its colonial past, France is involved in diverse military and post-conflict operations. The three western P5 members mentioning "Afghanistan" in the context of WPS debates mirrors insights by Schoenfeld et al. (2018) who found, through topic modeling, that mainly western countries would mention the topic "women and human rights" during UNSC debates on the UNSC agenda item "The Situation in Afghanistan".

Finally, using NEL also allows us to make use of the underlying knowledge graph. To do so, we selected those entities from the top 25 NEs shown in fig. 2 that relate to legal or political terms. From the knowledge graph, we added all NEs that are directly related via a subclass or an instance-of relation to the selected NEs (e.g. "sexual assault" or "reproductive rights") and that are also mentioned by P5 speakers in WPS debates. Figure 3 depicts a network of weighted directed edges (normalized) between the P5 members and all entities in the knowledge graph that they mention. We then added undirected edges (in green) between

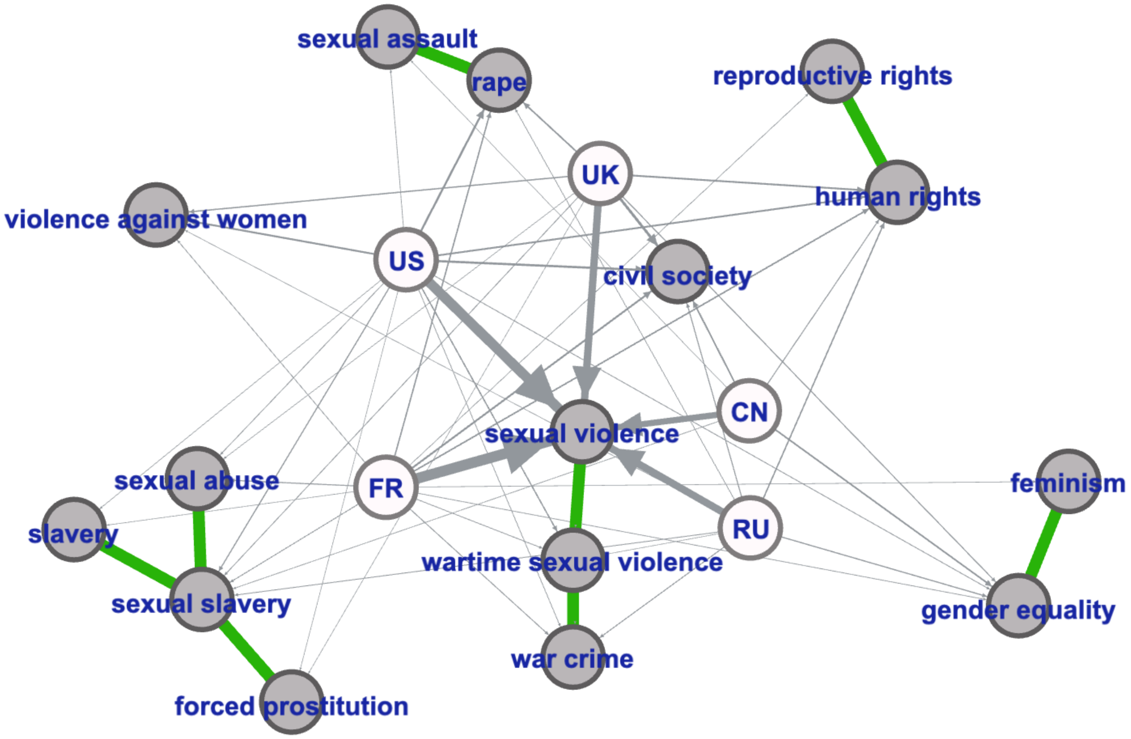


Figure 3: Network visualization of P5 countries' mentions of the most frequent policy-related NEs and directly related conceptual NEs in WPS debates. Directed edge strength represents frequency of mentions. Undirected edges (thick and green) are links in the knowledge graph.

concepts that are directly linked in the knowledge graph. As to be expected, the most often used conceptual entity—"sexual violence"—is most central in the network. However, adding less frequent NEs that are directly linked to frequent NEs adds further insights about speakers' choices: While China never mentions "rape", it makes use of the conceptually related "sexual assault". And while multiple speakers mention the more general "human rights" and "gender equality", France more explicitly mentions the more concrete "reproductive rights" and the more political term "feminism".

In sum, the NE-tagged corpus allows for observations that are in line with existing qualitative research on WPS debates and that link to previous insights based on quantitative research on the UNSC Debate corpus. A simple descriptive analysis of NE use already indicates differences in geographic focus between P5 members as well as similarities and differences in legal or institutional focus, while making use of the knowledge graph helps to find further differences between speakers' policy focus or framing of the debates. This suggests that further exploration of the corpus may reveal various domains of agreement and disagreement between the global powers. This may be most interesting in instances that are not along the most commonly known dividing lines, i.e. between France, the UK, and the US on one side and China or Russia holding different views on key issues (as represented by NEs), or on issues where this has not yet been noticed.

6 Limitations

Despite its potentials for political science research on language use in the UNSC, there are a few limitations.

Although the differentiation between entity recognition and labeling that NEL offers allows users to customize and filter the annotations, it is still not fully tailored towards usage in political sciences. There are erroneous classifications that we noticed during inspection: For instance, "president" is often falsely linked to the President of the United States while in the UNSC this is rather the President of the UNSC. This is a bias emerging from the the training data, highlighting that the choice of knowledge graph matters. Also, a direct mapping from text to Wikidata instead of going through the intermediary in DBpedia-spotlight may improve annotation quality in future research. Next, the quality metrics of the DBpedia-spotlight NEL pipeline compared to spaCy's NER pipeline show that the basic annotations of DBpedia are of lesser quality, due to the increase in granularity and linking to a knowledge graph. This has to be weighted against the additional depth the knowledge graph provides. Additionally the tagging could be compared to other NER pipelines like flair (Akbik et al., 2019). Lastly, there are alternative options for the format of the corpus: A more straightforward representation could be to represent the UN Security Council debates NE addon in RDF directly, instead of merely mentioning the URIs within the jsonlines format. The present format was chosen in favor of usability, especially for social scientists already familiar with json from working with json-based APIs (Benoit and Herzog, 2017), who should be able to inspect and analyse the corpus add-on easily and with the tools they prefer. Providing it in RDF requires users to be familiar with not only RDF but also SPARQL to interact with the corpus.

7 Conclusion

This paper presented the UN Security Council debates NEL add-on. Based on the previous work of Schoenfeld et al. (2019) we annotated NEs to the corpus using DBpedia-spotlight. We have demonstrated the potential for political scientists to turn to using NEL or NER based methods in their work. Compared to topic modeling, for example, NEL and NER provide more stable (i.e. reliable) results and they are more transparent. Through links to existing knowledge graphs or pre-trained classifiers they provide categorizations that can be directly used for social science analysis, e.g. showing agreement and disagreement between speakers in a speech corpus. While existing NER taggers may be good enough for many use cases, NEL methods can add the richness required for such analysis. However, despite these advantages, the analytical quality of the tags and links depends on the quality of the taggers—here: DBpedia-spotlight—used. Further validation across the entire UNSC-NE corpus add-on can show which tags, links, and categorization are most valid for research on diplomatic debates and thus to make choices of how to filter the corpus for different research questions.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and recommendations. We also thank Nathan Shuftan for the helpful remarks on this paper.

References

Ellis Aizenberg and Anne Skorkjær Binderkrantz. 2021. Computational approaches to mapping interest group representation: A test and discussion of different methods. *Interest Groups & Advocacy*, 10(2):181–192.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Elio Amicarelli and Jessica Di Salvatore. 2021. Introducing the PeaceKeeping Operations Corpus (PKOC). *Journal of Peace Research*.

Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2).

A Burcu Bayram and Vivian P Ta. 2019. Diplomatic chameleons: Language style matching and agreement in international diplomatic negotiations. *Negotiation and Conflict Management Research*, 12(1):23–40.

Kenneth Benoit and Alexander Herzog. 2017. Text analysis: estimating policy preferences from written and spoken words. *Analytics, policy and governance*, pages 137–159.

Kakia Chatsiou and Slava Jankin Mikhaylov. 2020. *Deep Learning for Political Science*, pages 1053–1078. SAGE Publications Ltd.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Inis L Claude. 1966. Collective legitimization as a political function of the united nations. *International organization*, 20(3):367–379.

Jorge M. Fernandes, Miguel Won, and Bruno Martins. 2020. Speechmaking and the Selectorate: Persuasion in Nonpreferential Electoral Systems. *Comparative Political Studies*, 53(5):667–699.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, page 1625. ACM Press.

Julia Gray and Alex Baturo. 2021. Delegating diplomacy: Rhetoric across agents in the United Nations General Assembly. *International Review of Administrative Sciences*.

Sebastian Hellmann, Marvin Hofer, Krzysztof Węcel, and Włodzimierz Lewoniewski. 2020. Towards a Systematic Approach to Sync Factual Data across Wikipedia, Wikidata and External Data Sources. In *Proceedings of the Conference on Digital Curation Technologies*, pages 1–15.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daniel Jurafsky and James H. Martin. 2018. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd draft edition.

Lennart Kerkvliet, Jaap Kamps, and Maarten Marx. 2020. Who mentions whom? recognizing political actors in proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 35–39, Marseille, France. European Language Resources Association.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529. Association for Computational Linguistics.

Anne-Kathrin Kreft. 2017. The gender mainstreaming gap: Security council resolution 1325 and un peacekeeping mandates. *International peacekeeping*, 24(1):132–158.

Sophie J. Lee, Howard Liu, and Michael D. Ward. 2019. Lost in Space: Geolocation in Event Data. *Political Science Research and Methods*, 7(4):871–888.

Emma K Macfarlane. 2021. Resolutions without resolve: Turning away from un security council resolutions to address conflict-related sexual violence. *Michigan Journal of Gender & Law*, 27(2):435–472.

Peter F. Nardulli, Scott L. Althaus, and Matthew Hayes. 2015. A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, 45(1):148–183.

Francesco Piccinno and Paolo Ferragina. 2014. From TagME to WAT: A new entity annotator. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, pages 55–62. ACM Press.

Olivier Schmitt. 2020. How to challenge an international order: Russian diplomatic practices in multilateral security organisations. *European Journal of International Relations*, 26(3):922–946.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. 2018. Discursive Landscapes and Unsupervised Topic Modeling in IR: A Validation of Text-As-Data Approaches through a New Corpus of UN Security Council Speeches on Afghanistan.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde Van Meegdenburg, and Antonio Pires. 2019. The UN Security Council Debates. `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH`.

Jacqui True and Antje Wiener. 2019. Everyone wants (a) peace: the dynamics of rhetoric and practice on ‘women, peace and security’. *International Affairs*, 95(3):553–574.

Legal Element Classification on German Parliamentary Debates

Christopher Klamm*, Martin Hock*
University of Kassel, TU Dresden
klamm@uni-kassel.de, martin.hock@tu-dresden.de

1. Abstract Submission

The use of force by states is unlawful. The Charter of the United Nations (UNC) (the most important treaty of international law) prohibits every use of or threat of the use of force (Art. 2 (4) UNC). There are two undisputed exceptions to this prohibition: self-defence (art. 51 UNC) and authorization by the Security Council of the United Nations (art. 39 + 42 UNC). Two further concepts - humanitarian intervention (HI) and responsibility to protect (R2P) - are legally disputed. Arguments supporting the lawfulness of the latter concepts are often based on customary international law (CIL) in addition to the treaty law of the Charter (Gray, 2018, p. 40-64). CIL consists of state practice that is accompanied by a sense of legal obligation, the so-called opinio iuris (Lepard, 2010, p. 6-7). All legal concepts are composed of *legal elements*[1], *these are the requirements that have to be fulfilled in order to achieve legal consequences and effects* (Wienbracke, 2013, p. 25-39). This highlights the importance of the legal elements. In order to prove the existence of opinio iuris the arguments brought forward to substantiate a legal concept are an important factor: if a state backs its practice by referring to a legal concept in general or to the legal elements of that concepts the existence of opinio iuris and thus CIL can be assumed (Lepard, 2010, p. 6-7). The legal elements can be found inter alia in parliamentary debates. For example, Ludger Volmer during the `KOSOVO` debate[2]: *"Es kann keinen Zweifel darin geben, daß es überfällig war, den boshaftesten Despoten in Europa [Element 1], der Krieg gegen sein eigenes Staatsvolk führt, es entwurzelt, in die Wälder treibt und ermorden läßt, [Element 2] in seine Schranken zu verweisen, um eine humanitäre Katastrophe noch größeren Ausmaßes zu verhindern [Element 3]."* Traditionally CIL does not explicitly include parliamentary debates as a source of opinio iuris (International Law Commission, 2018). We close this gap and treat parliamentary debates as a source of CIL. We aim to provide a new framework for annotating legal elements in parliamentary debates and an annotation of four debates with this new framework.

Legal expertise beyond word search is needed since legal elements are often ambiguous (i.e. a single sentence can be applied to more than one legal element). Furthermore, the legal concept referred to by a speaker is often not made explicit in parliamentary debates. For example, Minister of Defence Volker Rühe in the `KOSOVO` debate[3]: *"Es geht aber um die Abwehr einer humanitären Katastrophe."* Implicit in this claim is the legal concept of HI. It is not, however, explicitly mentioned. Nevertheless, the legal element of humanitarian catastrophe is stated. In order to deal with these ambiguities and the lack of explicit references to legal concepts the present system goes beyond word search and shows the need for a more comprehensive approach. From an international law point of view the paper asks wether legal elements can be found in parliamentary debates and thus substantiate the claim that opinio iuris regarding HI and R2P exists. Furthermore, it is asked whether the applied methods of Natural Language Processing (NLP) are sufficiently precise in order to automate the subsumption of parliamentary debates under legal elements. Advantages in NLP show the possibilities of applying new contextualized language models (Devlin et al., 2018) to identify such justifying elements. The models can deal with the automatic identification of supporting and opposing sentences within natural language (Reimers et al., 2020; Schaefer and Stede, 2020; Toledo-Ronen et al., 2020; Chakrabarty et al., 2020; Wang et al., 2020). These two fields of research are to be combined in order to enable the analysis of legal elements regarding the validity of legal concepts. This helps international law scholarship to ascertain the opinio iuris of states and substantiate the claim to validity of a given legal concept faster and on a broader empirical basis. At the same time, this represents a big challenge for the NLP-area since there are few prior works (Haigh, 2018; Yamada et al., 2019; Poudyal et al., 2020; Zhong et al., 2020) on legal texts. Parliamentary debates can be considered a cross-domain use case inasmuch as they treat questions of international law in an genuinely political setting. As of yet, there are no sufficiently fine-grained analyses regarding legal elements in the context of HI and R2P discussed in debates. Overall, the contributions of our work will address several points: (1) a new insight-driven task on the legal element classification in a cross-domain environment, (2) a theoretical-based framework to annotate parliamentary debates, (3) expert-based annotations and (4) evaluation as well as (5) an analysis of transformer-based contextualized embeddings for legal element classification.

*equal contribution, random order

[1] German: Tatbestandsmerkmale

[2] https://dserver.bundestag.de/btp/13/13248.pdf

[3] https://dserver.bundestag.de/btp/13/13248.pdf

2. Bibliographical References

Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., and Hwang, A. (2020). AmperSand: Argument mining for persuasive online discussions. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2933–2943.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). {BERT:} Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.0.

Gray, C. (2018). *International Law and the Use of Force*. Oxford University Press, 2.

Haigh, R. (2018). *Legal English*. Routledge, Abingdon, Oxon; New York, NY, 6.

International Law Commission. (2018). Draft conclusions on identification of customary international law, with commentaries.

Lepard, B. D. (2010). *Customary International Law*. Cambridge University Press, Cambridge.

Poudyal, P., Savelka, J., Ieven, A., Moens, M. F., Goncalves, T., and Quaresma, P. (2020). ECHR: Legal Corpus for Argument Mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75. Association for Computational Linguistics.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2020). Classification and clustering of arguments with contextualized word embeddings. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 567–578.

Schaefer, R. and Stede, M. (2020). Annotation and Detection of Arguments in Tweets. In *Proceedings of the 7th Workshop on Argument Mining*, number 2014, pages 53–58.

Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A., and Slonim, N. (2020). Multilingual Argument Mining: Datasets and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, H., Huang, Z., Dou, Y., and Hong, Y. (2020). Argumentation Mining on Essays at Multi Scales. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5480–5493, Stroudsburg, PA, USA. International Committee on Computational Linguistics.

Wienbracke, M. (2013). *Juristische Methodenlehre*. C.F. Müller, Heidelberg ; München ; Landsberg ; Frechen ; Hamburg.

Yamada, H., Teufel, S., and Tokunaga, T. (2019). Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law*, 27(2):141–170, 6.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Stroudsburg, PA, USA. Association for Computational Linguistics.

The role of interjections in Austrian parliamentary debates

Klaus Hofmann[1], Tanja Wissik[2]
Universität Wien[1], Österreichische Akademie der Wissenschaften[2]
klaus.hofmann@univie.ac.at, Tanja.Wissik@oeaw.ac.at

1. Introduction

Parliamentary discourse is among the most prototypical types of political speech. While much research on parliamentary discourse is focused on the regular speeches given by members of parliament (MPs), interjections and heckling have received less attention, despite the fact that they play an integral part in the communicative dynamics between representatives and parties in parliamentary sessions. In the past, research on interjections has mostly relied on small data sets and/or qualitative approaches (Ilie, 2003; Stopfner, 2013; Truan, 2017). However, more recent studies on German data have analyzed the use of interjections (*Zwischenrufe*) in the *Bundestag* and regional parliaments with quantitative methods (Brunner, 2021; Vögele and Thomas, 2019).

Inspired by these research activities, we investigate the distribution and usage of interjections in the Austrian Parliament. In particular, we ask whether interjections are employed asymmetrically by various groups of MPs and to what extent they differ with respect to style and function.

2. Data & Methods Used

To this end, we avail ourselves of the TEI-annotated Corpus of Austrian Parliamentary Records (ParlAT). In its current version, the corpus covers more than 20 years of recent parliamentary discourse, comprising c. 75 000 000 tokens. ParlAT is lemmatized, PoS-tagged and densely annotated with metatextual information, including dates, speaker IDs and utterance type classifications (Wissik and Pirker, 2018; Wissik, forthcoming). This allows us to systematically extract interjections and analyze their distribution.

In addition, we combine the linguistic data with lexical ratings of abstractness (concreteness, imageability) and emotion (arousal, valence) (Köper and Schulte im Walde, 2016) to characterize the style and communicative functions of interjections along those dimensions. In terms of analytical methodology, we rely on regression modeling using R (Baayen, 2008).

3. Preliminary Results

Preliminary results confirm that interjections are very unevenly distributed among MPs. Multivariate logistic regression reveals that (a) female members are much less likely to utter interjections than their male colleagues; (b) right-wing parties are more likely to use interjections than liberal and left-leaning parties (Figure 1); (c) members of the opposition are more prone to verbal interjections than members of governing parties; and (d) the relative incidence of interjections varies considerably between legislative periods, even when confounds such as gender and party membership are controlled for.

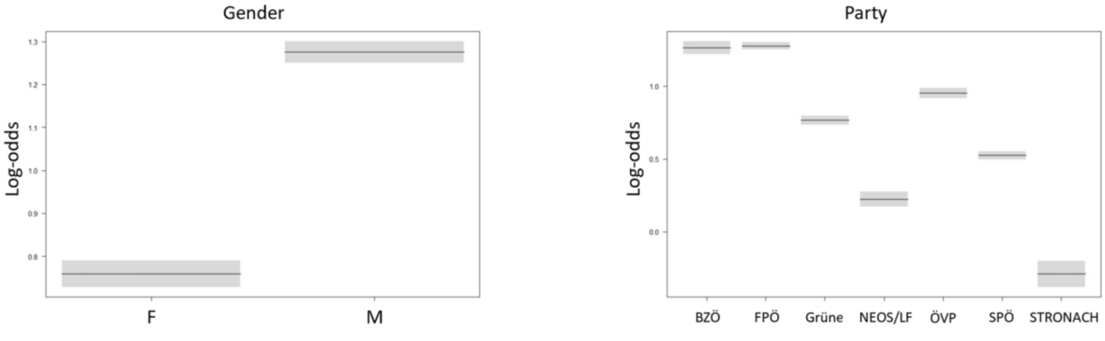


Figure 1: Distribution of interjections by gender (left) and by party (right)
(compared to regular speeches, log odds)

In terms of style and function, the analysis suggests that (a) liberal parties' interjections use language that is more abstract, more imageable (Figure 2), more positive, and less arousing; (b) the interjections from opposition parties are less positive and more arousing than those from coalition parties; (c) women's interjections are more abstract and more positive than those from their male colleagues.

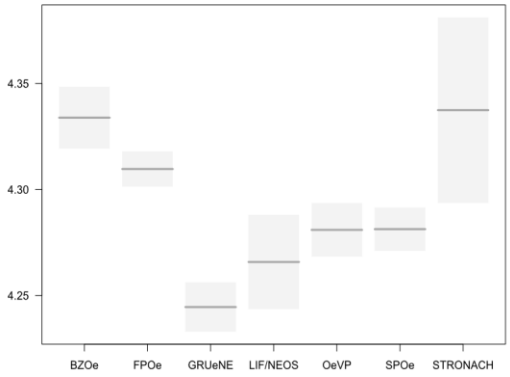


Figure 2: Imageability of interjections by party

4. Bibliographical References

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press, DOI: https://doi.org/10.1017/CBO9780511801686.

Brunner, Katharina; Ebitsch, Sabrina; Gierke, Sebastian; Schories, Martina (2018). Das gespaltene Parlament. In: *SZ online*, 24 April 2018, https://projekte.sueddeutsche.de/artikel/politik/die-afd-im-bundestag-e362724/ (last accessed: 30 June 2021).

Ilie, Cornelia (2003). Interruption patterns in British parliamentary debates and drama dialogue. In: Betten, Anne; Dannerer, Monika (eds.). *Dialogue analysis IX: dialogue in literature and the media: selected papers from the 9th IADA Conference, Salzburg 2003. Part 1: Literature*, 415–430.

Köper, Maximilian; Schulte im Walde, Sabine (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC16)*, 2595–2598.

Stopfner, Maria (2013). *Streitkultur im Parlament. Linguistische Analyse der Zwischenrufe im österreichischen Nationalrat.* Göttingen: Narr.

Truan, Naomi (2017). Zwischenrufe zwischen parlamentarischer Routine und Kreativität. Bundestagsdebatten über Europa aus dem Blickwinkel von unautorisierten Unterbrechungen. *Cahiers d'Études Germaniques* 73, 125–138, DOI: https://doi.org/10.4000/ceg.2370.

Vögele, Catharina; Thomas, Claudia (2019). Die isolierte Fraktion. Zwischenreaktionen, Zwischenrufe und die AfD im Baden-Württembergischen Landtag. *Zeitschrift für Parlamentsfragen* 2019(2), 306326, DOI: https://doi.org/10.5771/0340-1758-2019-2-306.

Wissik, Tanja; Pirker, Hannes (2018). ParlAT beta Corpus of Austrian Parliamentary Records. In: *Proceedings of the LREC 2018 Workshop ParlaCLARIN: LREC2018 workshop on creating and using parliamentary corpora*, 20–23.

Wissik, Tanja (forthcoming). Encoding interruptions in parliamentary data: From applause to interjections and laughter. In: *Journal of the Text Encoding Initiative* 14.

Lexical Convergence and Divergence in Austrian Parliamentary Debates: A Network-Based Approach

Anna Marakasova[1], Klaus Hofmann[2], Andreas Baumann[2], Julia Neidhardt[1], Tanja Wissik[3]
TU Wien[1], Universität Wien[2], Österreichische Akademie der Wissenschaften[3]
anna.marakasova@tuwien.ac.at, klaus.hofmann@univie.ac.at, andreas.baumann@univie.ac.at,
neidhardt@ec.tuwien.ac.at, Tanja.Wissik@oeaw.ac.at

1. Background and Research Aim

Parliamentary debates are a key source for studying political discourse. Ostensibly, debates have the function to discuss the merits of legislative proposals and governmental policies. From a sociology-of-politics perspective, however, debates are at least equally important for developing the image of a party or of individual politicians in contrast to their political opponents (Huang, Perry and Spirling, 2020; Atzpodien, 2020).

We investigate some of the dynamics at work in the debates of the Austrian Parliament, especially focusing on topical and discursive unity and divergence within and across parties.

2. Data & Methods

Our data comes from the Corpus of Austrian Parliamentary Records (ParlAT). The corpus covers parliamentary discourse in the National Chamber between 1996 and 2017. ParlAT is lemmatized, PoS-tagged and includes metatextual information such as speaker ID or type of utterance (Wissik and Pirker, 2018).

The study is situated at the interface of natural language processing and quantitative linguistics. For each speaker in parliament providing sufficient text, we construct similarity-based network representations of the speaker's typical lexical repertoires in every year (Figure 1). This allows us to model the topical and discursive patterns in their speech. We rely on adaptations of the skip-gram algorithm for representing semantic similarities between words (Mikolov et al., 2013).

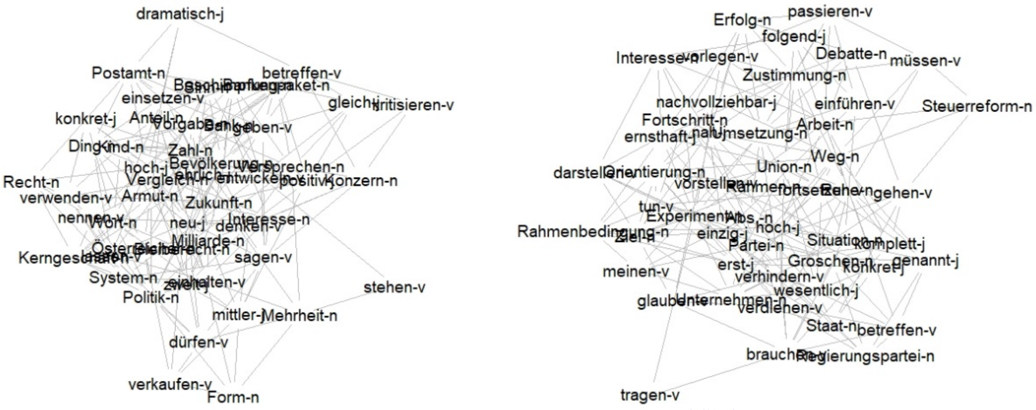


Figure 1: Network representations of Heinz-Christian Strache (left, FPÖ, 2009) and Eva Glawischnig (right, Grüne, 2015). Number of vertices and edges reduced for visualization.

The politicians' discourse patterns are compared by calculating the Canberra distances between their networks (Lance 1967; Levshina, 2015). To identify diachronic trends, network distances are related to party membership and other predictor variables by means of general additive models. Multidimensional scaling is performed on network distances for visual exploration of group coherence within parties and government coalitions (Wood, 2006; Baayen, 2008).

3. Results

The similarity/distance between the discourse patterns of Austrian members of parliament fluctuates over time. There is greater similarity within parties than between parties, although there is also a trend towards greater similarity across all networks. Female members of the parliament (MPs) show a tendency towards more similar discourse patterns in certain years. The networks of MPs from opposition parties become markedly more similar over time, while the networks of members of government parties show the opposite trend.

4. Bibliographical References

Atzpodien, D. S. (2020). Party competition in migration debates: The influence of the AfD on party positions in German state parliaments, *German Politics*, DOI: 10.1080/09644008.2020.1860211.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press, DOI: https://doi.org/10.1017/CBO9780511801686.

Huang, L.; Perry, P. O.; Spirling, A. (2020). A general model of author "style" with application to the UK House of Commons, 1935–2018. *Political Analysis* 28, 412434, DOI: https://doi.org/10.1017/pan.2019.49.

Lance, G. N.; Williams, W. T. (1967). Mixed-data classificatory programs. I. Agglomerative Systems. *Australian Computer Journal*, 15–20.

Levshina, N. (2015). *How to do linguistics with R. Data exploration and statistical analysis.* Amsterdam: John Benjamins.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26, 3111-3119, arXiv:1310.4546v1.

Wang, Shirui; Zhou, Wenan; Jiang, Chao (2020). A survey of word embeddings based on deep learning. *Computing* 102, 717–740, DOI: 10.1007/s00607-019-00768-7.

Wissik, Tanja; Pirker, Hannes (2018). ParlAT beta Corpus of Austrian Parliamentary Records. In *Proceedings of the LREC 2018 Workshop ParlaCLARIN: LREC2018 workshop on creating and using parliamentary corpora*, 20–23.

Wood, S. N. (2006). *Generalized additive models. An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC, DOI: https://doi.org/10.1201/9781315370279.

Crime and the Electoral Success of Populist Radical Right Parties

Uwe Remer, Raphael Heiberger, Marius Kaffai
University of Stuttgart,Institute for Social Sciences, CSS Lab
Seidenstraße 36, 70174 Stuttgart, Germany
{uwe.remer, raphael.heiberger, marius.kaffai}@sowi.uni-stuttgart.de

1. Abstract

Rising public approval and electoral gains for radical right parties and populist movements contest liberal democracies all over Europe (Mudde 2007; Norris & Inglehart 2019: 9; Guth & Nelsen 2021). By positioning themselves as law and order parties together with a pronounced framing of the migration crisis in terms of security threats, radical right parties aim to obtain issue ownership on immigration and link it with crime (Mudde 2007: 146; Dinas & van Spanje 2011: 661; Arzheimer 2018: 151). The strategic use of agenda setting and priming to combine these issues is main part of the electoral strategy of populist radical right parties (Arzheimer 2018: 157). On the empirical side, the findings are mixed and scarce. Neither the effect immigration, nor the effect of crime on the electoral success of radical right parties is uncontested (Coffé et al. 2007; Mudde 2007: 224; Smith 2010; Dinas & van Spanje 2011; Arzheimer 2018: 156–157; Dennison 2020: 398; Deiss-Helbig & Remer 2021). Potential sources for the inconclusiveness of these findings are differences in scale and level of aggregation, heterogeneous operationalization of the theoretical constructs (Kaufmann & Goodwin 2018; Deiss-Helbig & Remer 2021: 3), and the complex interaction between the variables at play (Dinas & van Spanje 2011). Our contribution connects to this research puzzle. We ask, whether and how crime has an influence on the success of populist radical right parties [and how this effect is moderated by the local presence of immigrants]. Based on previous research, we assume that immigration evokes a perceived threat within parts of the electorate (Deiss-Helbig & Remer 2021) which leads to increased vote proportions of populist radical right parties (Green et al. 2016). We extend the state research as we study the proposed effects on the local level at several elections at three levels of the political system over six years: national, state, and local elections in the state of Baden-Württemberg, Germany. Beside official criminal statistics, we utilize original data to measure crime with a resolution down to the municipal level. The analysis is based on a corpus of nearly 500.000 police press reports published since 2015 by police departments in the state of Baden-Wuerttemberg, Germany. To be able to match the crime reports with the official electoral results on municipal level, first, the documents are geolocated. In a second step, human coders annotate a sample of the corpus for text classification. The labeled data are then used for supervised machine learning to classify the documents regarding the reported crime. The crimes that are identified by the classified documents are aggregated on the level of municipal administrative units. With this measure of crime prevalence on the local level, we are able to test the local influence of reported crime on the local vote shares of populist radical right parties and its interaction with immigration. As controls, we account for potential confounders like official crime statistic, urbanization and, wealth. Prospectively, we hope to be able to differentiate between different types of crime and to Our preliminary results reveal heterogeneous relationship between reported crime and votes for radical right parties.

2. Bibliographical References

Arzheimer, K. (2018). Explaining Electoral Support for the Radical Right. In J. Rydgren (ed.), The Oxford handbook of the radical right, pp. 143–165. New York, NY: Oxford University Press.

Coffé, H., Heyndels, B. & Vermeir, J. (2007). Fertile grounds for extreme right-wing parties: Explaining the Vlaams Blok's electoral success. Electoral Studies 26(1): 142–155. doi: 10.1016/j.electstud.2006.01.005.

Deiss-Helbig, E. & Remer, U. (2021). Does the Local Presence of Asylum Seekers Affect Attitudes toward Asylum Seekers? Results from a Natural Experiment. European Sociological Review. doi: 10.1093/esr/jcab036.

Dennison, J. (2020). How Issue Salience Explains the Rise of the Populist Right in Western Europe. International Journal of Public Opinion Research 32(3): 397–420. doi: 10.1093/ijpor/edz022.

Dinas, E. & van Spanje, J. (2011). Crime Story: The role of crime and immigration in the anti-immigration vote. Electoral Studies 30(4): 658–671. doi: 10.1016/j.electstud.2011.06.010.

Green, E.G.T. et al. (2016). From Stigmatized Immigrants to Radical Right Voting: A Multilevel Study on the Role of Threat and Contact. Political Psychology 37(4): 465–480. doi: 10.1111/pops.12290.

Guth, J.L. & Nelsen, B.F. (2021). Party choice in Europe: Social cleavages and the rise of populist parties. Party Politics 27(3): 453–464. doi: 10.1177/1354068819853965.

Kaufmann, E. & Goodwin, M.J. (2018). The diversity Wave:A meta-analysis of the native-born white response to ethnic diversity. Social science research 76: 120–131. doi: 10.1016/j.ssresearch.2018.07.008.

Mudde, C. (2007). Populist radical right parties in Europe. Cambridge, New York: Cambridge University Press.

Norris, P. & Inglehart, R. (2019). Cultural Backlash. Cambridge University Press.

Smith, J.M. (2010). Does Crime Pay? Issue Ownership, Political Opportunity, and the Populist Right in Western Europe. Comparative Political Studies 43(11): 1471–1498. doi: 10.1177/0010414010372593.

Representing Political Topics with Sentence Transformers - Transfer Learning with Topic Centroids

Moritz Laurer
Centre for European Policy Studies (CEPS)
moritz.laurer@posteo.de

Political scientists have collected large-scale textual datasets over the past decades, classifying texts in political categories. The most prominent dataset is the Manifesto Corpus, which classifies party manifestos in 7 domains and 56 sub-categories (Burst et al., 2020). In the past years, several researchers have leveraged this dataset to train machine learning classifiers which are then applied to texts from a different domain – an approach called transfer learning (Ruder, 2019). Some research groups have used Manifesto data to train classifiers to identify topics in COVID-19 press releases (Chatsiou, 2020), to classify the related Comparative Agendas Dataset (Terechshenko et al., 2020), parliamentary debate motions (Abercrombie et al. 2019) and political speeches (Osnabrügge et al., forthcoming). These analyses mostly use softmax based classifiers like convolutional neural networks (CNN) or transformers. The accuracy of these analyses, however, remains relatively low due to issues of noisy source data (Mikhaylov et al., 2012) and transfer learning challenges.

This paper proposes an alternative approach to classifying texts into the Manifesto topical categories, which might be more suitable for working in a transfer learning setting and with noisy data: centroid classification with sentence transformers. Sentence transformer models transform sentences to dense vectors which are designed for finding other, semantically similar sentences (Reimers et al., 2019; sbert.net, n.d.). Sentences of the same class can be transformed to vectors and their centroid can be calculated. This centroid can be understood as the average representation of the respective class – for example the topic 'political corruption' in party manifestos. These 'topic centroids' can then be compared to any other sentence with cosine similarity, indicating their similarity to the topic. *These assumptions lead to two research questions: Are 'topic centroids' created with sentence transformers accurate representations of topics in a given domain? What are the advantages and disadvantages of 'topic centroids' in a transfer learning setting when applied to data from a different domain?*

The approach is tested on 130.000 English quasi-sentences from the Manifesto Corpus categorised in 56 sub-categories of political topics (Burst et al. 2020). First, in-domain classification is tested. Previous papers have trained text classifiers like CNNs or transformer models on the manifesto dataset, but have mostly treated the task as a classification task with a softmax layer (Terechshenko et al., 2020; Chatsiou, 2020; Abercrombie et al., 2019). Abercrombie et al. report a 0.42 F1 macro score, Bilbao-Jayo and Almeida (2018) also report 0.42 F1 and on 56 manifesto classes and Osnabrügge et al. (forthcoming) report 0.417 F1 macro and 0.388 balanced accuracy on 44 merged classes (using logistic regression). Terechshenko et al. and Chatsiou only classify the 7 domains and report up to 0.84 and 0.87 accuracy respectively. When we apply 'topic centroids' of 56 manifesto classes to a 25% test set for this paper, we obtain 0.40 balanced accuracy, 0.35 F1 macro and 0.45 F1 micro. While the comparison strongly depends on the number of classes and the metric, this indicates that 'topic centroids' are roughly on par with the classifiers in the literature for in-domain classification.

The main argument of this paper is that 'topic centroids' have important advantages over these classifiers in real-world transfer learning settings. A key challenge in transfer learning is the different label space in source and target data Ys != Yt (Ruder, 2019). First, when applied to another domain, softmax classifiers are forced to only classify text into the classes they have been trained on and perform badly for out-of-distribution detection (Zhang, 2020). A state-of-the-art softmax classifier is 99% sure that the sentence "the unicorn ate my shoes" should be classified in "welfare and quality of life". With centroid classification, on the other hand, any sentence can be compared to topic centroids and the comparison will return a low similarity score if the sentence is unrelated. Second, the centroids created with sentence transformers are modular. Centroids for 50+ classes can be calculated, but only two of them can be used to analyse a target text if desired. Third, sentence transformers can be used for multi-label classification, even if the training data is only annotated with single labels. If a sentence is close to the centroid of two topics, the sentence can be attributed to both topics based on a manually defined threshold. This is particularly useful for social science datasets like the manifesto corpus which suffers from noisy and overlapping labels, given the higher complexity of the labelling scheme (Mikhaylov et al., 2012).

We therefore use the topic centroids for multi-label classification: For each sentence, we select the 3 nearest topic centroids and then discard those centroids that are below a similarity threshold. The threshold (0.61) is the average distance of training sentences to their gold label centroid minus one standard deviation. This multi-label classifier obtains 0.61 balanced accuracy, 0.56 F1 macro and 0.66 F1 micro. Thanks to the threshold, it can also discard out-of-distribution data in a transfer learning setting.

The analysis is in active development at the time of writing. In the next step, the predictions from the multi-label classifier will be tested through manual annotations of both in-domain data, as well as data from a different domain (news articles).

Bibliographical References

Abercrombie, Gavin, Federico Nanni, Riza Theresa Batista-Navarro, and Simone Paolo Ponzetto. 2019. 'Policy Preference Detection in Parliamentary Debate Motions'. CoNLL. https://doi.org/10.18653/v1/K19-1024.

Burst, Tobias, Krause Werner, Pola Lehmann, Lewandowski Jirka, Theres Mattheiß, Nicolas Merz, Sven Regel, and Lisa Zehnter. 2020. 'Manifesto Corpus'. WZB Berlin Social Science Center. https://manifesto-project.wzb.eu/information/documents/corpus.

Chatsiou, Kakia. 2020. 'Text Classification of Manifestos and COVID-19 Press Briefings Using BERT and Convolutional Neural Networks'. *ArXiv:2010.10267 [Cs]*, November. http://arxiv.org/abs/2010.10267.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *ArXiv:1810.04805 [Cs]*, May. http://arxiv.org/abs/1810.04805.

Mikhaylov, Slava, Michael Laver, and Kenneth R. Benoit. 2012. 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos'. *Political Analysis* 20 (1): 78–91. https://doi.org/10.1093/pan/mpr047.

Osnabrügge, Moritz, Elliot Ash, Massimo Morelli (Forthcoming): 'Cross-Domain Topic Classification for Political Texts'. *Political Analysis*. DOI: 10.1017/pan.xxxx.xx

Reimers, Nils and Iryna Gurevych. 2019. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. *Proceedings of EMNLP 2019*. http://arxiv.org/abs/1908.10084

Ruder, Sebastian. 2019. 'Neutral Transfer Learning for Natural Language Processing'. PhD thesis. Available at: https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf

sbert.net. n.d. 'Pretrained Models — Sentence-Transformers Documentation'. Accessed 26 June 2021. https://sbert.net/docs/pretrained_models.html.

Terechshenko, Zhanna, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2020. 'A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics'. SSRN Scholarly Paper ID 3724644. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3724644.

Zhang, Jian-Guo, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. 2020. 'Discriminative Nearest Neighbor Few-Shot Intent Detection by Transferring Natural Language Inference'. *ArXiv:2010.13009* [Cs], October. http://arxiv.org/abs/2010.13009.

Politics in between crises. A political and textual comparative analysis of budgetary speeches and expenditure.

Alice Cavalieri, Dario Del Fante
University of Turin (Italy), Institute of Computational Linguistics“A. Zampolli” - CNR (Italy)
alice.cavalieri@unito.it, dario.delfante@ilc.cnr.it

1. Abstract

European countries have recently been heavily hit by two dramatic crises (i.e, the great recession and the Covid-19 pandemic) which have dramatically smashed national and even supranational economic policies. From an agenda-setting perspective, the moment of crisis, by concentrating the attention on the trend topic, overpowers the usual political dynamics, transforming political actors from agenda-setters to agenda-takers. In ‘normal’ times, namely in those periods when there is not a sudden necessity to immediately respond to an external shock, instead, budget changes tend to be very small and largely determined by previous years’ spending choices. To efficaciously analyze budgetary changes over time, for a long time scholars have been using the annual percentage change as dependent variable, both considering the total expenditure or the expenditure in single budget categories. However, this method neglects a fundamental aspect, that is the complementary nature of spending allocation across budget functions. Plainly, when the government decides to increase expenditure for a certain budget category, a parallel reduction of expenditure in another budget category follows. As budget spending of multiple categories add up to 100% of total spending, budget trade-offs can be treated as a compositional dependent variable. Using public expenditure data from the Eurostat database which split expenditure into 10 macro categories, we engage in a preliminary analysis of budgetary trade-offs across spending categories focusing on Italy and the UK between 2013 and 2019, the period in-between two crises when – we suppose – governments had more chances to steer the allocation of expenditure according to their ideological and/or strategic considerations. The choice of this time frame is driven also by methodological reasons, as we use for the very first time the ParlaMint dataset which, for now, encompasses the period from 2013 onwards. ParlaMint, is a linguistically annotated corpus of parliamentary debates across European countries, financed by CLARIN ERIC, the research infrastructure for language as social and cultural data, which aims to convert existing contemporary multilingual and diverse cross-national parliamentary data into comparable and interpretable resources. This dataset is meant to provide one of the main – if not the most important – independent variable(s) explaining budgetary trade-offs. More precisely, starting from the available data, we build a sub-corpus of debates about the budget, then we create additional sub-corpus referring to specific budget categories and verify whether the emphasis during parliamentary debates leads to swinging in the allocation of expenditure. Including the analysis of parliamentary debates into the study of budgetary changes is crucial, as political parties exploit the parliamentary arena to strategically emphasize their policy position and because is in the parliament that political conflicts unfold and also both majority and opposition parties can shape government’s decisions. The combined analysis of parliamentary debates and of the final expenditure, carried out by adding quantitative text analysis techniques to the analysis of expenditure trade-offs, despite its preliminary stage, helps to better grasp dynamics which public budgeting is subject to and constitute a very promising venue for future research both for political science and linguistic scholars..

INVITED TALK I
Using NLP to Track Health Implications of Climate Change

Slava Jankin
Hertie School of Governance
Berlin, Germany
jankin@hertie-school.org

Abstract

Climate change is undermining the foundations of good health; threatening the food we eat, the air we breathe, and the hospitals and clinics we depend on. However, the response to climate change could be the greatest global health opportunity of the 21st century. The Lancet Countdown: Tracking Progress on Health and Climate Change brings together 35 leading academic institutions and UN agencies from every continent to monitor this transition from threat to opportunity. We track annual indicators of progress, empowering the health profession and supporting policymakers to accelerate their response. In the talk we discuss the application of natural language processing to develop and track a set of Lancet Countdown indicators, focusing on political speech, corporate commitments, social media, and parliamentary debates. We also discuss challenges in establishing attributional, causal links between statements about health and climate change.

Index of Authors