Volume 35 — Number 2 — 2022 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Special Issue on Computational Linguistics for Political and Social Sciences

Edited by
Ines Rehbein, Gabriella Lapesa, Goran Glavaš and Simone Paolo Ponzetto

Imprint

Editor
Christian Wartena
Publication supported by the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)

Board of Directors
Committee and Advisory Board of the GSCL

Current issue
Volume 35 2022 Issue 2

Guest Editors
Ines Rehbein, Gabriella Lapesa, Goran Glavaš and Simone Paolo Ponzetto

Address
Christian Wartena
Hochschule Hannover
Expo Plaza 2
D-30539 Hannover
info@jlcl.org

ISSN
2190-6858

Publication
Mostly 2 issues per annum
Publication only electronically on jlcl.org

Contents

Computational Linguistics for Political and Social Sciences

In recent years, an increasing number of studies has been published in the newly emerging text-as-data field. More and more scholars in the areas of political and social science are taking advantage of the ever increasing amount of text available (not only on the internet, but also transcriptions of parliamentary debates, newspaper texts, or party manifestos) to address a heterogeneous set of research questions. While this trend has already brought many promising results, it is not free of risks and challenges. In their seminal paper, Grimmer and Stewart (2013) not only discuss the potential of text-as-data approaches but also highlight the pitfalls that arise when applying NLP methods for the investigation of questions from political and social science.

We therefore argue for a closer collaboration between scholars from the social/political sciences on the one hand, and researchers from the area of computer science, NLP and computational linguistics on the other hand, to overcome those challenges and advance the state of the art for applications in the field of computational social science. As a first step to bridge the gap between the different communities, we organised the 1st Workshop on Computational Linguistics for the Political and Social Sciences (CPSS 2021).[1] The workshop took place in September 2021 as a virtual event, co-located with the Conference on Natural Language Processing (KONVENS 2021) in Düsseldorf, Germany. To meet the diverse needs of our research fields, we not only asked for long and short paper submissions but also for non-archival abstracts, in order to allow researchers to discuss work in progress without committing to a publication. The workshop program included five long and four short paper presentations and six non-archival abstracts that have been presented as posters. The presentations covered a wide range of topics, starting from NLP tools and corpus annotation that support research in the social science (Glaser, Patz, & Stede, 2021; Kahmann, Niekler, & Wiedemann, 2021) to the analysis of framing and formulaic speech (Russo, Comandini, Caselli, & Patti, 2021; Yu & Fliethmann, 2021), work on topic modelling and topic detection for political text analysis, using a variety of supervised and unsupervised techniques (Ahltorp, Dürlich, & Skeppstedt, 2021; Brand, Schünemann, König, & Preböck, 2021; Koh, Boey, & Béchara, 2021; Kreutz & Daelemans, 2021) and methodological studies (De Vos & Verberne, 2021). This JLCL special issue presents four selected long paper contributions from CPSS 2021.[2]

The first paper by De Vos and Verberne addresses a methodological question, namely the problem of data sparsity for the application of machine learning in political research. The authors present a replication study where they investigate the impact of pre-processing when only little data is available, showing the sensitivity of the models to variation regarding training and test splits and pre-processing. Their findings question

[1] https://old.gscl.org/en/arbeitskreise/cpss/cpss-2021

[2] The proceedings of the CPSS 2021 workshop are available online: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2021/workshop-proceedings/352683648-1630596221/cpss2021-proceedings.pdf.

previous results from the literature and highlight the importance of data set size and the validation of model robustness.

The contribution of Yu and Fliethmann studies media framing in German newspaper articles on the European Refugee Crisis (2014–2018). The authors test approaches to frame detection that do not rely on large-scale manual annotations. Their first method is based on LDA topic modelling, the second approach combines static word embeddings with a set of handcrafted keywords based on an expert-curated framing schema. Comparing the two techniques, Yu and Fliethmann show that the embedding-based approach yields better and more interpretable results. This illustrates the benefits to be gained from interdiscipinary work that combines domain knowledge from political science with NLP techniques for exploratory text analyses.

Another approach related to framing is presented in Russo et al. who analyse the use of proto-slogans in political communication before the 2019 European election, based on more than 700,000 comments extracted from the Facebook pages of two Italian leaders of populist parties (Matteo Salvini and Luigi Di Maio). The paper describes how the data has been clustered, followed by a manual annotation step, in order to detect proto-slogans used by the party leaders' supporters. The long-term objective of this work is the identification of stylometric patterns in informal populist social media posts.

The final paper by Glaser and collegues argues for using Named Entity Recognition and Named Entity Linking, two well-established NLP tasks, as an alternative source of information for political text analysis that is more transparent, robust and interpretable than topic modelling. The paper presents an add-on to the United Nations Security Council (UNSC) Debates corpus (Schoenfeld, Eckhard, Patz, van Meegdenburg, & Pires, 2019) and compares two approaches for obtaining this information. The pros and cons of each method are discussed, based on an intrinsic evaluation and an exploratory study that asks which entities are mentioned by different political actors in debates on the agenda of Women, Peace and Security.

Due to space limitations, we have only been able to report some of the results from the papers in this volume, and the short summaries given above surely do not do the work justice. Therefore, we invite the reader to form their own opinion and hope that they will find them insightful and intellectually rewarding.

We would like to thank the authors for their fine contributions and the reviewers for their constructive feedback which helped to improve the quality of the manuscripts: Adrien Barbaresi, Julian Bernauer, Chris Biemann, Christian Gawron, Goran Glavas, Annette Hautli-Janisz, Slava Jankin, Jonathan Kobbe, Sebastian Pado, and Esther van den Berg. Finally, we want to thank the editors of the Journal for Language Technology and Computational Linguistics for their support in putting together this special issue. We hope that the reader will enjoy the result!

The guest editors,
Ines Rehbein, Gabriella Lapesa, Goran Glavaš and Simone Paolo Ponzetto.

References

Ahltorp, M., Dürlich, L., & Skeppstedt, M. (2021). Textual contexts for "Democracy": Using topic- and word-models for exploring Swedish government official reports. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 45–52).

Brand, A., Schünemann, W. J., König, T., & Preböck, T. (2021). Detecting policy fields in German parliamentary materials with Heterogeneous Information Networks and node embeddings. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 53–58).

Glaser, L., Patz, R., & Stede, M. (2021). UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 79-88).

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi: 10.1093/pan/mps028

Kahmann, C., Niekler, A., & Wiedemann, G. (2021). Application of the interactive Leipzig Corpus Miner as a generic research platform for the use in the social sciences. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 39-44).

Koh, A., Boey, D. K. S., & Béchara, H. (2021). Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 67–78).

Kreutz, T., & Daelemans, W. (2021). A semi-supervised approach to classifying political agenda issues. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 59–63).

Russo, I., Comandini, G., Caselli, T., & Patti, V. (2021). Share and shout: Discovering proto-slogans in online political communities. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 25-33).

Schoenfeld, M., Eckhard, S., Patz, R., van Meegdenburg, H., & Pires, A. (2019). *The UN Security Council Debates.* Retrieved from `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH`

de Vos, H., & Verberne, S. (2021). Small data problems in political research: a critical replication study. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 3–12).

Yu, Q., & Fliethmann, A. (2021). Frame Detection in German Political Discourses: How Far Can We Go Without Large-Scale Manual Corpus Annotation? In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 13–24).

Hugo de Vos and Suzan Verberne

Small Data Problems in Political Research: A Critical Replication Study

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 & 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 & 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 & 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,

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

and members (Anastasopoulos & 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 & 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 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 77 times since 2019.[2]

In this paper, we replicate the results by A&W, and analyze their validity. We perform what Belz, Agarwal, Shimorina, and Reiter (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

[2] According to Google Scholar, April 2022

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 & 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 & 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 aggravated by the short length of tweets: They contain few words compared to other document types such as party manifestos (Merz, Regel, & Lewandowski, 2016; Verberne, D'hondt, van den Bosch, & Marx, 2014) or internet articles (Fraussen, Graham, & Halpin, 2018). Because tweets are so short, the bag-of-words representation will be sparse, and in a small data

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

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, Mahmoud, and Hefny (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, Collignon-Delmar, Benoit, and Matsuo (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, Chang, Lee, & Toutanova, 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, Farra, & Nakov, 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,

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

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. All details on the exact replication can be found in the scripts in the supplementary material.

Class distribution : In the data from the original paper 32% of the tweets was on Moral reputation and 58% was the 'other class'. So this is a fairly balanced data set. However only after the classes Performative Reputation (12%), Procedural Reputation (1.5%), Technical Reputation (12%) and 'None' (42.5%) were put together as the 'other' class.

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 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

[5] In their case this seed is 41616

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:

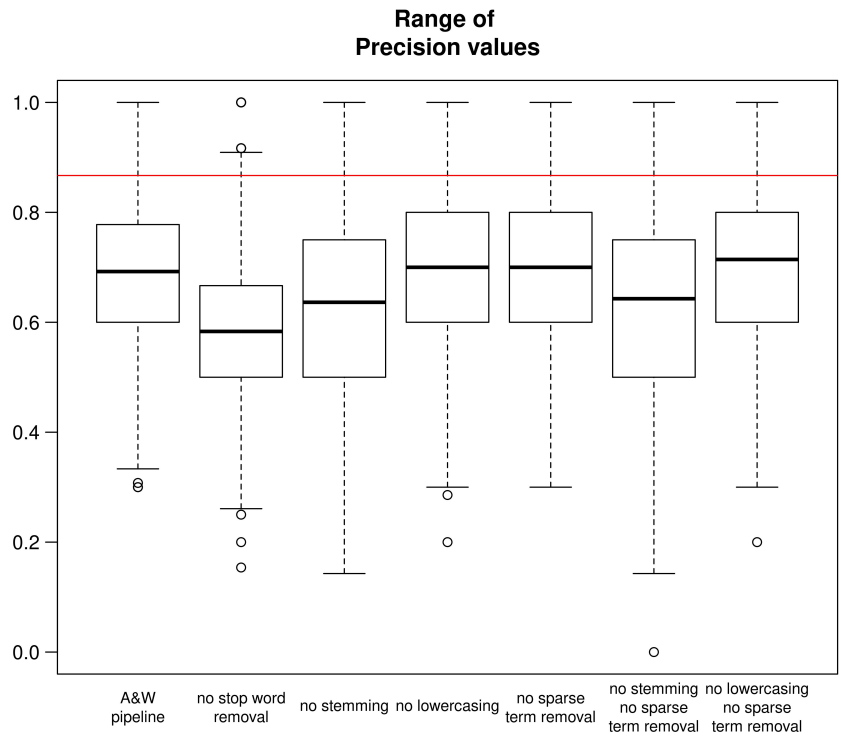

Figure 1: A visualization of the spread of results of the random seed variation experiment. The leftmost box summarizes the results of 1000 different runs with the same settings as A&W, except for the random seeds. The horizontal line depicts the precision that is reported by A&W. The other box plots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

- 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 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 stop-word 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.

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.

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

	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.

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 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.

Also on the topic of evaluation there could be a wider debate on whether F1, precision or recall are the most suitable for this situation. Recently, they have been shown to present overly optimistic results in binary classification tasks, in which case the Matthews correlation coefficient (Chicco & Jurman, 2020) has been argued to perform better.

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 only 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

Amador Diaz Lopez, J. C., Collignon-Delmar, S., Benoit, K., & Matsuo, A. (2017, January). Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data. *Statistics, Politics and Policy*, *8*(1). Retrieved 2020-02-26, from `http://www.degruyter.com/view/j/spp.2017.8.issue-1/spp-2017-0006/spp-2017-0006.xml` doi: 10.1515/spp-2017-0006

Anastasopoulos, J. L., & Whitford, A. B. (2019, June). Machine Learning for Public Administration Research, With Application to Organizational Reputation. *Journal of Public Administration Research and Theory*, *29*(3), 491–510. Retrieved 2020-02-26, from `https://academic.oup.com/jpart/article/29/3/491/5161227` doi: 10.1093/jopart/muy060

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (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* (pp. 381–393).

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Bouchet-Valat, M. (2020). Snowball stemmers based on the c 'libstemmer' utf-8 librar [Computer software manual]. Retrieved from `https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf` (R package version 0.6.0)

Burnap, P., & Williams, M. L. (2015, June). 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. Retrieved 2020-02-26, from http://doi.wiley.com/10.1002/poi3.85 doi: 10.1002/poi3.85

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). San Francisco, California, USA: ACM Press. Retrieved 2020-02-26, from http://dl.acm.org/citation.cfm?doid=2939672.2939785 doi: 10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13.

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., ... Hunter, L. E. (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* (Vol. 2018, p. 156).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elghazaly, T., Mahmoud, A., & Hefny, H. A. (2016). Political Sentiment Analysis Using Twitter Data. In *Proceedings of the International Conference on Internet of things and Cloud Computing - ICC '16* (pp. 1–5). Cambridge, United Kingdom: ACM Press. Retrieved 2020-02-26, from http://dl.acm.org/citation.cfm?doid=2896387.2896396 doi: 10.1145/2896387.2896396

Feinerer, I., & Hornik, K. (2021). tm: Text mining package [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/tm/index.html (R package version 0.7-8)

Fraussen, B., Graham, T., & Halpin, D. R. (2018, October). Assessing the prominence of interest groups in parliament: a supervised machine learning approach. *The Journal of Legislative Studies*, *24*(4), 450–474. Retrieved 2020-02-26, from https://www.tandfonline.com/doi/full/10.1080/13572334.2018.1540117 doi: 10.1080/13572334.2018.1540117

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. Retrieved 2020-02-26, from https://www.cambridge.org/core/product/identifier/S1047198700013401/type/journal_article doi: 10.1093/pan/mps028

Grün, B., Hornik, K., Blei, D., Lafferty, J., Phan, X.-H., Matsumoto, M., ... Cokus, S. (2021). topicmodels [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf (R package version 0.2-12)

Heaps, H. S. (1978). *Information retrieval, computational and theoretical aspects*. Academic Press.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic

researchers. *arXiv preprint arXiv:2108.02497*.

Merz, N., Regel, S., & Lewandowski, J. (2016, April). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 205316801664334. Retrieved 2020-02-26, from http://journals.sagepub.com/doi/10.1177/2053168016643346 doi: 10.1177/2053168016643346

Nikolov, A., & Radivchev, V. (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* (pp. 691–695).

Rosenthal, S., Farra, N., & Nakov, P. (2019, December). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *arXiv:1912.00741 [cs]*. Retrieved 2021-06-09, from http://arxiv.org/abs/1912.00741 (arXiv: 1912.00741)

Verberne, S., D'hondt, E., van den Bosch, A., & Marx, M. (2014, July). Automatic thematic classification of election manifestos. *Information Processing & Management*, *50*(4), 554–567. Retrieved 2020-02-26, from https://linkinghub.elsevier.com/retrieve/pii/S0306457314000168 doi: 10.1016/j.ipm.2014.02.006

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Correspondence

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

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

Qi Yu and Anselm Fliethmann

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

Abstract

Automated detection of *frames* in political discourses has gained increasing attention in natural language processing (NLP). However, earlier studies in this area focus heavily on frame detection in *English* using *supervised* machine learning approaches. Addressing the difficulty of the lack of annotated data for training and 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 an original 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 exploratory 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 consistent influence on citizens' political opinions (Druckman, 2004; Nelson & Oxley, 1999; Slothuus, 2008). In the field of NLP, recent years have witnessed growing attention on the automated detection of frames in political discourse (e.g., Baumer, Elovic, Qin, Polletta, & Gay, 2015, Card, Gross, Boydstun, & Smith, 2016, Field et al., 2018, Khanehzar, Turpin, & Mikolajczak, 2019, Cabot, Dankers, Abadi, Fischer, & Shutova, 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, Boydstun, Gross, Resnik, & Smith, 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, Ng, & Jordan, 2003), and 2) a combination of word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and carefully selected 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 the 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, as these results are more explainable. We release the resulting framing keywords as a publicly available lexical resource under: `https://github.com/qi-yu/refugees-and-migration-framing-vocabulary`

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, Card, Gross, Resnik, and Smith (2014), a large amount of previous studies on frame detection have focused 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). Further studies using similarly supervised or weakly supervised settings, but based on other manually annotated datasets than the MFC, include Baumer et al. (2015); Johnson, Jin, and Goldwasser (2017); Liu, Guo, Mays, Betke, and Wijaya (2019); and Mendelsohn, Budak, and Jurgens (2021).

Frame detection in languages other than English remains greatly neglected so far. 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 performance 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

In our work here, we investigate the effectiveness of NLP approaches in frame detection that do not require large-scale corpus annotation from scratch. For this purpose, we use 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 the slightly right-leaning FAZ and the center-left-leaning SZ (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 did not only incorporate a tabloid, but also brought together the three most highly-circulated printed newspapers in Germany (Deutschland.de, 2020).

From each newspaper, articles containing at least one match with the following quasi-synonyms of 'refugee' (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. 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 publishers. Unfortunately, due to their copyright regulations, we cannot make the corpus publicly available.

4 Experiment 1: Detecting Frames Using Topic Modelling

As the task of detecting frames strongly resembles the detection of sub-aspects within the event under discussion, it is reasonable to give topic modelling a trial as a first bottom-up, data-driven method for exploring differences in frames between the newspapers. We therefore trained one LDA-based model per newspaper to explore frame differences between the publications.

4.1 Training

We used the Python library *Gensim* (Řehůřek & 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, Zhang, Zhang, Bolton, & Manning, 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, Both, and Hinneburg (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 resulting 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

[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.

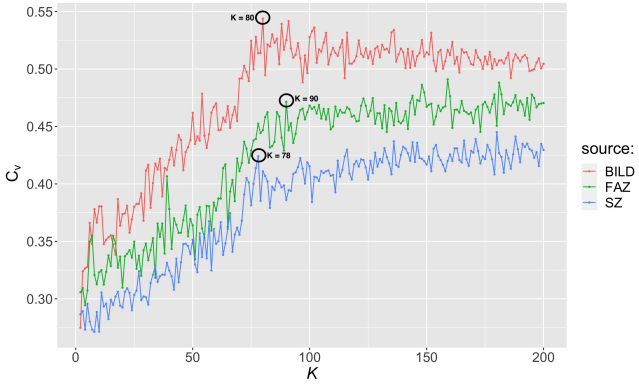

Figure 1: C_v coherence score of topic number $K \in [2, 200]$ in BILD, FAZ and SZ.

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 considerably complicates human interpreting of the overall topic differences between newspapers. The results can therefore barely inform further analyses of 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 fairly high, since all articles focus thematically on issues related to refugees and migration. This contrasts to other more vocabulary-heterogeneous datasets on which LDA-based topic modelling has been shown to achieve much clearer topic division, e.g., the 20 NewsGroups corpus used in Harrando, Lisena, and Troncy (2021), the Wikipedia corpora used in Markoski, Markoska, Ljubešić, Zdravevski, and Kocarev (2021), or the IMDB movie review dataset used in Kherwa and Bansal (2020). 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 on 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 real stop words, bear highly relevant information for the context of refugees and migration. We leave further theoretical and empirical investigation on the reason of the poor performance of topic modelling for future studies, as this is beyond the scope of the current paper.

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 terms 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.

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-driven method. First, we 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. Next, 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 (Akyürek et al., 2020; Field et al., 2018; Johnson et al., 2017).

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, generally distinguishes 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 an *extremely broad* range of 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, de Boef, and Boydstun (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 parts 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 predominantly provides detailed and meaningful differentiations 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 exploratory reading of a sample of articles from our corpus, 5 domain experts (graduate students of political science) listed words and phrases that they found highly relevant to each frame category in our schema. These seed vocabulary lists were then expanded by synonyms of each item, found using GermaNet (Hamp & Feldweg, 1997; Henrich & 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

categories by Habermas (1991)	**frame**	**description: frames...**	**exemplary keywords**
utilitarian	economy*	... related to jobs, education, financial issues, etc., incl. *human resources frames*, *material resources frames*	Armutsflüchtling (poverty refugees), Arbeitskräftemangel (labor shortage), Ausbildung (training)
	legal	... related to legal questions, incl. *jurisprudence frames*, *law frames*	Rechtsanspruch (legal entitlement), Bleibeperspektive (perspective to stay), Asylrecht (asylum right)
	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)
moral-universal	morality*	... concerning ethics and moral concepts, incl. *humanitarianism frames*, *fairness and equality frames*	Menschenwürde (human dignity), Willkommenskultur (welcoming culture), solidarisch (showing solidarity)
identity-related	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)

Table 3: *Refugees and Migration Framing Schema* and corresponding example keywords to each category extracted with methods described in Section 5.2. *Category names following Boydstun et al. (2014).

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 & Hanks, 1990) to C:

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

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.

5.3 Mention Rate of Frames

As a first exploratory 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 is a non-parametric alternative of *analysis of variance* (ANOVA), and we chose it because the mention rate values in single articles of each newspaper 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 (which can also be observed from the article numbers and token numbers in Table 1) 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.

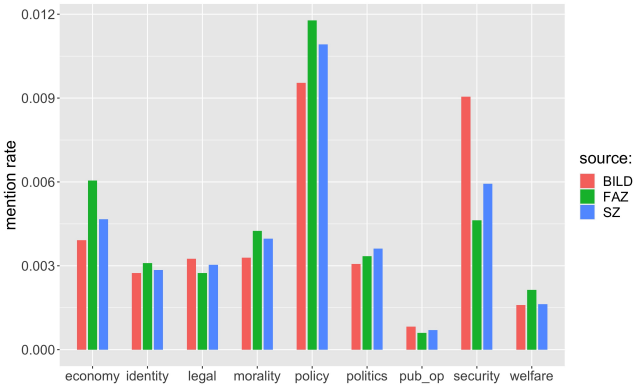

Figure 2: Mention rates of different frames in articles from 2014-2018 in BILD, FAZ and SZ.

	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 usage differences can be observed by measuring the mention rate, this metric is coarse and 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') belong both 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`. This 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 in each newspaper. In all four frame categories interesting 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').

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 the `refugee_centroid` within the *Security*, *Economy*, *Policy* and *Morality Frames* in BILD, FAZ and SZ.

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'). This is 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 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'). For the FAZ, top ten items 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 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

In this article we addressed the difficulty that for many low-resource languages there are no large-scale annotated datasets available for training and/or evaluating models of automated frame detection. We did so by experimenting 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 this method for detecting topic preferences in a corpus where the vocabulary is highly homogeneous. Our second experiment with word2vec embeddings and the carefully selected *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. Given the broadness of our corpus from which we took parts of our vocabulary lists, as well as the inclusion of additional vocabulary from an additional corpus, we are confident in having achieved unbiased word lists of acceptable quality. Nevertheless, achieving a reliable and objective evaluation of the quality of vocabulary lists is a generally inevitable difficulty for dictionary-based approaches. In future work we will therefor attempt to further strengthen the quality of our vocabulary lists by exploring the potential of more sophisticated keyword mining techniques, such as the method proposed by Jin, Bhatia, and Wanvarie (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 the workshop CPSS 2021.

References

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., & Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8614–8624).

Bade, K. J. (2015). Zur Karriere abschätziger Begriffe in der deutschen Asylpolitik. In *Aus Politik und Zeitgeschichte* (pp. 3–8).

Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (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 (pp. 1472–1482).

Baumgartner, F. R., de Boef, S., & Boydstun, A. E. (2008). *The decline of the death penalty and the discovery of innocence.* New York, Cambridge: Cambridge University Press.

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

Boydstun, A. E., Card, D., Gross, J. H., Resnik, P., & Smith, N. A. (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 30 May 2022)

Cabot, P.-L. H., Dankers, V., Abadi, D., Fischer, A., & Shutova, E. (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* (pp. 4479–4488).

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (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* (pp. 438–444).

Card, D., Gross, J. H., Boydstun, A., & Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1410–1420).

Church, K., & Hanks, P. (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 30 May 2022)

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

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

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3570–3580).

Habermas, J. (1991). *Erläuterungen zur Diskursethik.* Frankfurt am Main: Suhrkamp.

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications* (pp. 9–15).

Harrando, I., Lisena, P., & Troncy, R. (2021). Apples to apples: A systematic evaluation of topic models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 483–493).

Helbling, M. (2014). Framing immigration in Western Europe. *Journal of Ethnic and*

Migration Studies, *40*(1), 21–41.

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

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

Ji, Y., & Smith, N. A. (2017). Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 996–1005).

Jin, Y., Bhatia, A., & Wanvarie, D. (2021). Seed word selection for weakly-supervised text classification with unsupervised error estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 112–118).

Johnson, K., Jin, D., & Goldwasser, D. (2017). Modeling of political discourse framing on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 556–559).

Khanehzar, S., Turpin, A., & Mikolajczak, G. (2019). Modeling political framing across policy issues and contexts. In *Proceedings of The 17th Annual Workshop of the Australasian Language Technology Association* (pp. 61–66).

Kherwa, P., & Bansal, P. (2020). Topic modeling: a comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24).

Lapesa, G., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J., & Padó, S. (2020). DEbateNet-mig15: Tracing the 2015 immigration debate in Germany over time. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 919–927).

Liu, S., Guo, L., Mays, K., Betke, M., & Wijaya, D. T. (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)* (pp. 504–514).

Markoski, F., Markoska, E., Ljubešić, N., Zdravevski, E., & Kocarev, L. (2021). Cultural topic modelling over novel Wikipedia corpora for South-Slavic languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 910–917).

Mendelsohn, J., Budak, C., & Jurgens, D. (2021). Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2219–2263).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Naderi, N., & Hirst, G. (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* (pp. 1–9).

Nelson, T. E., & Oxley, Z. M. (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 30 May 2022)

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (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* (pp. 101–108).

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408).

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

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

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

Correspondence

Qi Yu
Department of Linguistics & Cluster of Excellence "The Politics of Inequality"
University of Konstanz, Germany
`qi.yu@uni-konstanz.de`

Anselm Fliethmann
Department of Politics and Public Administration & Cluster of Excellence "The Politics of Inequality"
University of Konstanz, Germany
`anselm.fliethmann@uni-konstanz.de`

Irene Russo, Gloria Comandini, Tommaso Caselli and Viviana Patti

Share and Shout: Proto-Slogans in Online Political Communities

Abstract

This paper proposes a methodology for investigating populism on social media by analyzing the emergence of proto-slogans, defined as 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 created by the parties' supporters.
Our manual annotation consists of four layers, namely:*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 that nonetheless express 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 (Stieglitz & Dang-Xuan, 2012; Stier, Bleier, Lietz, & Strohmaier, 2018).

The vast amount of comments on political topics daily produced 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, most comments on social media are characterized by the synthetic expression of a point of view. Analyzing this type of content is challenging: due to their brevity, a topic-based analysis of users' comments performs poorly. Nonetheless, short comments have a pragmatic function in online debates, and often through the use of nominal utterances (NUs) (Comandini & Patti, 2019; Comandini, Speranza, & Magnini, 2018), help to build a common view among supporters of the same party/politician.

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 characterize one of the political communities investigated in this paper, i.e. Lega Nord (LN, Northern League). 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 realized 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. We find that proto-slogans are a communicative device exploited by populist supporters, a stylistic feature usable in the future to detect emerging populist attitudes.

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. Their parties were gaining consensus, with the LN winning the European elections. However, both parties have lost consensus at the time of writing, and their leaders have changed their communication. These changes are mainly due to the new roles these leaders are now covering (Salvini being part of the ruling majority with no position in the government and Di Maio being Foreign Minister). Besides this difference, the data analyzed still represent a valuable tool for gaining insights into the communication strategies of populist leaders and parties.

To investigate the linguistic behavior of these communities, 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 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 for 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 summarized as follows:

- an operational definition of proto-slogan as a key aspect in bottom-up populist communication on the web (Section 3.1);
- a methodology that combines unsupervised approaches (i.e., clustering) and manual annotation to identify political proto-slogans (Sections 3.3 and 3.4).

Article Outline The remainder of this article is structured as follows: in Section 2, we describe related work on populism on social media, what characterizes the language of online communities, and how slogans have been linguistically analyzed. Section 3 focuses on the methodology, presenting a definition of proto-slogan. It also describes how data

from social media have been extracted and pre-processed. The results emerging from the manual annotation of automatically clustered instances are discussed. In Section 4, we provide some preliminary comparisons between Facebook and Twitter data, before concluding in Section 5.

2 Related Work

Populist discourses have been analyzed from different perspectives, collecting them in corpora that ease qualitative and quantitative analyses. However, it is still unclear if populist discourses can be linguistically identified - independently from the historical moment, the latitude, and the political orientation. Moreover, social media has had a disruptive effect on the propagation of populist discourses, affecting some of its features in a way still under scrutiny. This section reports on relevant literature about populism on social media and how language is shaped by communication in online communities in order to frame the reception of populist discourses in online contexts. In particular, slogans are presented as stylistic devices related to the emergence of a shared attitude among users.

2.1 Populism on Social Media

Social media are fundamental for understanding populist ideologies, which are mainly identified by their communication style (Kriesi, 2014; Aslanidis, 2016; Stanyer, Salgado, & Strömbäck, 2016). In particular, Facebook seems to be the preferred social network of populist parties (Ernst, Engesser, Buchel, Blassnig, & Esser, 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 & 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, analyzing how supporters react to it.

Populist leaders often adopt an emotional and straightforward communication style to be more persuasive and trigger a more emphatic response on social media (Oliver & Rahn, 2016). Indeed, it has been shown 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 supporters, appearing more approachable (Jacobs & 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 (Ernst, Esser, Blassnig, & Engesser, 2018; Engesser, Fawzy, & Larsson, 2017).

Together with the sense of belonging to the in-group due to the general resentment toward the out-group (Hameleers, Reinemann, Schmuck, & Fawzi, 2019), this perceived intimacy with the leader creates a strong sense of being part of a homogeneous community, supportive of 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 & Mackie, 2008).

Previous computational linguistics studies of populism are scarce. Recently, (Huguet Cabot, Abadi, Fischer, & Shutova, 2021) present a crowdsourcing annotated dataset for populist attitudes that collects comments about news on Reddit that mention a set of social groups (i.e., immigrants and Muslims), manually classifying attitudes toward them as supportive, critical, or discriminatory. In detecting the overall stance of comments, their analysis does not target exclusively populistic content and how populist attitudes are expressed. Instead, our work starts with the assumption that comments on the chosen politicians' Facebook public pages are mainly supportive and sympathetic to the populist rhetoric. Thus, they constitute the ideal starting point for a stylistic investigation of the reception of populist discourse online.

2.2 The language 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 see the new word and if a member of their network uses it or not (Eisenstein, O'Connor, Smith, & Xing, 2014). The central members of the network introduce lexical innovations that are successfully adopted by other members if there is a subset of adopters with strong ties (Tredici & Fernández, 2018). Creating a shared vocabulary is both a prerequisite and a consequence of being part of a cohesive community, even online.

With a data-oriented analysis, Khalid and Srinivasan (2020) show that different communities have peculiar styles, and the stylistic choices of users are a good predictor of group membership, more than the topics discussed.

The adaptations at the stylistic level contribute to being well-received by a community. In (Tran & Ostendorf, 2016) the reception of content posted by users with positive feedback is investigated through a hybrid n-grams and topic models to characterize the style and the topic of language in Reddit online communities. Stylistic features have discriminatory power for distinguishing between communities: the style is a better indicator of community identity than the topic. The authors found a positive correlation between the community reception of a contribution and the style similarity to that community. On the contrary, this does not hold for topic similarity.

In this paper, we argue that nominal utterances are a stylistic device that characterizes the reception of populist discourse online, being the pragmatic choice that supports the creation of a shared vocabulary among supporters.

2.3 Slogans as Linguistic Devices

Slogans are usually short, expressive, and assertive utterances, easy to memorize and spread (Amălăncei, Cîrtiţă-Buzoianu, & Daba-Buzioianu, 2015). They 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 that 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 & Toivonen, 2020).

The slogans that have received attention in previous works are those used in advertising. Political slogans are less studied, although they generally follow the same advertising rules and have the same goal: influencing people's behaviors (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 linguistic features.

Top-down slogans 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. In this paper, we present a methodology to detect bottom-up slogans that, if widely adopted, can shed light on the emerging attitudes of political supporters online.

3 Methodology

This section presents a definition of proto-slogan and reports on how data have been extracted from Facebook and pre-processed before automatically clustering them. In addition, the details of the manual annotation performed on clustered data are documented.

3.1 Definition of proto-slogan

Regarding the general political rhetoric, it is possible to differentiate two different kinds of slogans: top-down and bottom-up. Top-down political slogans are produced by the politician's communication team and generally convey a complex message in a carefully crafted short, sharp form, such as *porti chiusi* (*closed harbors*); on the other hand, bottom-up political slogans are produced by the political electorate and generally convey a less complex message, as in *avanti tutta* (*full steam ahead*).

However, analyzing bottom-up political slogans, it would appear that some of the linguistic devices used by the political electorate, such as *forza Salvini* (*go Salvini*), are even simpler than the others, conveying an even less complex message. These very simple linguistic devices do not appear to be real slogans, because they convey only the user's stance (positive or negative) regarding a target, which is always explicitly mentioned.

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 political slogans produced spontaneously by the politician's supporters, which still convey a relatively complex message.

For example, these slogan-like constructions usually appear as concise messages supporting (example 1) or denigrating (example 2) a politician or a group of people.

1. *Grande Di Maio!* [*Great Di Maio!*]

2. *Giornalisti venduti!* [*Corrupted journalists!*]

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. More specifically, proto-slogans are classified as a subset of bottom-up slogans, as illustrated in Figure 1.

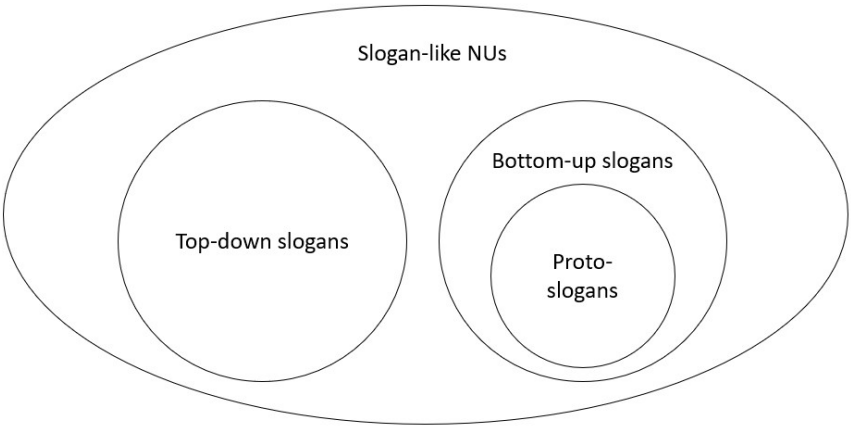

Figure 1: The subsets of slogan-like NUs

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 them in social media.

Even if peculiar syntactic structures do not characterize slogans, slogans and especially proto-slogans are often realized syntactically as nominal utterances (Comandini et al., 2018), also known as fragments without an overt antecedent (Merchant, 2005). NUs are linguistic constructions without a verb in a finite form in their syntactic nucleus and are very common in informal spoken English (Merchant, 2005) (example 3) and Italian (Cresti, 1998) (example 4).

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

4. (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, 2011a, 2011b). It has been found that NUs are often used as a device to convey hate speech, as shown in the POP-HS-IT corpus, by frequently taking the form of a verbless, hateful slogan (Comandini & Patti, 2019), such as *TOLLERANZA ZERO* (*ZERO TOLLERANCE*) and *rimpatriare subito tutti gli immigrati irregolari* (*immediately repatriate all irregular immigrants*). 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, actual slogans (Benveniste, 1990). Thus, not all NUs are slogans, but slogans are often NUs.

Therefore, many slogan-like constructions are NUs. Slogan-like NUs can be separated into two groups, as illustrated in Figure 1: top-down slogan-like NUs, produced by the politician's communication team, and bottom-up slogan-like NUs, produced spontaneously by the politician's supporters. Bottom-up slogan-like NUs also have an additional subset: proto-slogans, which convey a positive or negative stance towards a target.

3.2 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 the 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,[1] 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

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

FB page	Avg. post length	Avg. comm. lenght
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. All numbers refer to tokens.

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

Table 2: Eligible NUs after preprocessing

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.

As Table 1 shows, the two groups have similarities and differences. 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 nominal utterances (NUs) as syntactic declarative constructions built around a nonverbal head, framing them as the minimal unit of meaning in online communication. Therefore, 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, Klein, & Loper, 2009);
- its content has been PoS-tagged with TreeTagger (Schmid, 1995);
- finally, 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.

The dimensions of the dataset before and after the preprocessing steps are reported in Table 2. The preprocessed data, more than 240k comments, have been the input for the clustering algorithm based on semantic similarity.

3.3 Aggregating Data Through Clustering

The amount of data after preprocessing is such that a manual exploration is not feasible (see Table 2 for details). We thus decided to aggregate the pre-processed 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*. In this case, using known estimate methods such as the Elbow curve does not represent a solution.

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. First, each message representing an eligible NU has been converted as a 300-dimensions vector using FastText (Bojanowski, Grave, Joulin, & Mikolov, 2016). We then computed the pairwise cosine similarity scores between vectorized messages. The result is a N by N matrix of similarity scores.

Similarity scores below 0.6 were excluded and replaced with zeros to reduce noise in the data.[2] 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 quick manual exploration of the data, ensuring a fine-grained level of analysis. We plotted their centroids and observed their distributions for each group of clusters. 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 daily, creating eight sets of comments, while aggregating those for Di Maio in three blocks. These differences are due to the number of messages available for the two politicians.

3.4 Manual annotation

Since centroids have been obtained by means of semantic similarity scores, 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 list of centroids (1,650 in total) obtained from 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. The agreement between annotators is calculated after discussion on divergent choices.

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

[2] With a similarity equals to zero, messages are considered to be very different.

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

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, as they do not convey informations about Tense, Aspect and Mood.

The list below provides several examples of NUs retrieved in our dataset:

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

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

7. <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 2.3, 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 involves some form of subjective interpretation. Below we report examples of slogans in our dataset:

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

9. <NU> *Orgogliosi della propria identità* </NU> [*Proud of our identity*]

10. <NU> *Forza Salvini* </NU> [*Go Salvini*]

The third layer has been applied only to those items previously annotated as slogans by both annotators, 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:

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

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

As illustrated by example 12, 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 illustrates, 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

FB page	NUs	Top down slogans	Bottom up slogans
Salvini	926	22	204
Di Maio	369	5	57

Table 3: Eligible NUs after preprocessing

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

- <NU> *via i ladroni* </NU> [*away the robbers*]
- <NU> *m5s tutta la vita* </NU> [*m5s for the rest of my life*]

In Table 4 we report how many NUs have been labeled as proto-slogans. Bottom-up NUs are proto-slogans when they express a positive or negative stance towards a discourse target (always explicitly mentioned), whose identity is common knowledge for the electoral base.

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

Table 4: Proto-slogans after annotation

Comparing the bottom-up slogans and proto-slogans produced by the users to those produced by the politicians, it is clear that Salvini uses both these kinds of slogans very frequently, while Di Maio generally uses only top-down slogans. Salvini often uses 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-slogans, 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 it 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*).

Furthermore, 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 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.

4 Facebook and Twitter Data Comparison

Slogan-like NUs are specific linguistic items for a political community, when 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 need a more complex investigation based on more data.

We propose a qualitative classification of slogan-like NUs complementary to proto-slogans' characterization. In order to investigate this aspect, after extracting and annotating nominal utterances from Facebook public pages, the list of NUs 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] The aim of this analysis is the identification of three types of slogan-like NUs:

- Generic slogan-like NUs: nominal utterances whose content does not directly concern populism or are 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.

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

[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)

Examples	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

coherent with the perceived intimacy of Facebook communication, which makes leaders seem more approachable.

On Twitter, top-down slogans are more productive (see examples in Table5) 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 to answer Matteo Salvini's tweets, while in 2019 appeared more frequently in free-standing tweets. *Porti chiusi* is an example of an attested slogan-like NU that is distinctive for a political community but can also be used to address this community, criticizing 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*) or *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.

5 Conclusions and Future Work

Political communication on social media can be investigated 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 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 supporters 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 the 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 publicly available data.

References

Alnajjar, K., & Toivonen, H. (2020). Computational Generation of Slogans. *Natural Language Engineering*.

Amălăncei, B.-M., Cîrtiţă-Buzoianu, C., & Daba-Buzioianu, C. (2015). Looking for the best slogan: an analysis of the slogans of the 2016 Romanian parliament campaign. *Studies and Scientific Researches. Economics Edition*, *26*.

Aslanidis, P. (2016). Is populism an ideology? A refutation and a new perspective. *Political Studies*, *64*.

Benveniste, É. (1990). *Problemi di linguistica generale*. Milano, Italia: Mondadori.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

Bobba, G. (2019). Social media populism: features and 'likeability' of lega Nord communication on facebook. *European Political Science*, *18*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Comandini, G., & Patti, V. (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.*

Comandini, G., Speranza, M., & Magnini, B. (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.* (Vol. 2253). CEUR-WS.org.

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

Eisenstein, J., O'Connor, B. T., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE*, *9*.

Engesser, S., Fawzy, N., & Larsson, A. O. (2017). Populist online communication: introduction to the special issue. *Information, Communication and Society*, *20*.

Ernst, N., Engesser, S., Buchel, F., Blassnig, S., & Esser, F. (2017). Extreme parties and populism: an analysis of facebook and twitter across six countries. *Information Communication and Society*, *20*.

Ernst, N., Esser, F., Blassnig, S., & Engesser, S. (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*.

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

Ferrari, A. (2011b). Stile nominale. *Enciclopedia dell'Italiano*. Retrieved from `http://www.treccani.it/enciclopedia/stile-nominale_(Enciclopedia-dell'Italiano)/` (`http://www.treccani.it/enciclopedia/stile-nominale_(Enciclopedia-dell'Italiano)/`)

Ferrier, A. (2014). *The advertising effect. How to change behaviour.* South Melbourne: Oxford University Press University Press.

Hameleers, M., Reinemann, C., Schmuck, D., & Fawzi, N. (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, & C. H. de Vreese (Eds.), *Communicating Populism. Comparing Actor Perceptions, Media Coverage, and Effects on Citiziens in Europe.* New York and London: Routledge.

Huguet Cabot, P.-L., Abadi, D., Fischer, A., & Shutova, E. (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. Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2021.eacl-main.165

Jacobs, K., & Spierings, N. (2016). *Social media, parties, and political inequalities.* Basingstoke: Palgrave Macmillan.

Khalid, O., & Srinivasan, P. (2020). Style matters! investigating linguistic style in online communities. In *nternational aaai conference on web and social media.*

Kriesi, H. (2014). The populist challenge. *West European Politics*, *37*(2). Retrieved from https://doi.org/10.1080/01402382.2014.887879 doi: 10.1080/01402382.2014.887879

McCulloch, G. (2019). *Because internet: Understanding the new rules of language.* Riverhead Books.

Merchant, J. (2005). Fragments and ellipsis. *Linguistics and Philosophy*, *27*.

Mudde, C. (2004). The Populist Zeitgeist. *Government and Opposition*, *39*(4).

Oliver, J. E., & Rahn, W. M. (2016). Rise of the trumpenvolk: Populism in the 2016 election. *The ANNALS of the American Academy of Political and Social Science*, *667*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference.* New York, NY, USA: ACM.

Rooduijn, M., & Akkerman, T. (2017). Flank attacks: Populism and left-right radicalism in western europe. *Party Politics*, *23*.

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

Smith, E. R., & Mackie, D. M. (2008). Intergroup emotions. In L. F. B. M. Lewis J. M. Haviland-Jones (Ed.), *Handbook of emotions.* New York: Guilford.

Stanyer, J., Salgado, S., & Strömbäck, J. (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, & C. H. de Vreese (Eds.), *Populist Political Communication in Europe.* New York: Routledge.

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, *3*.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political Communication*, *35*.

Tran, T., & Ostendorf, M. (2016). Characterizing the language of online communities and its relation to community reception. In *Emnlp.*

Tredici, M. D., & Fernández, R. (2018). The road to success: Assessing the fate of linguistic innovations in online communities. *ArXiv*, *abs/1806.05838*.

Westen, D. (2007). *The political brain. the role of emotion in deciding the fate of the nation.* New York: PublicAffairs.

Zappavigna, M. (2015). Searchable talk: the linguistic functions of hashtags. *Social Semiotics*, *25*.

Correspondence

Irene Russo
Institute for Computational Linguistics "A. Zampolli" (ILC-CNR)
University of Pisa
irene.russo@ilc.cnr.it

Gloria Comandini
Italian Institute of German Studies
University of Trento
gloria.comandini@unitn.it

Tommaso Caselli
Center for Language and Cognition
University of Groningen
t.caselli@rug.nl

Viviana Patti
Department of Computer Science
University of Turin
patti@di.unito.it

5

Luis Glaser, Ronny Patz, Manfred Stede

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

Abstract

We present the Named Entity (NE) add-on to the previously published United Nations Security Council (UNSC) Debates corpus (Schoenfeld, Eckhard, Patz, Meegdenburg, & Pires, 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, Dasandi, & Mikhaylov, 2017), discursive landscapes through topic modeling (Eckhard, Patz, Schönfeld, & van Meegdenburg, 2021) or inter-state agreement in international negotiations through linguistic style matching (Bayram & 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, Montani, Van Landeghem, & Boyd, 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 corpus is 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 & 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 & Mikhaylov, 2020).

However, applications of NER published in political science journals are still rare. Most existing contributions focus on geographical locations (Nardulli, Althaus, & Hayes, 2015), demonstrating how geolocated event data using NER can be used to identify places of conflict or protest (Lee, Liu, & Ward, 2019). Geolocation is also applied by

[1] Version 2.0 of the UNSC-NE is accessible at `https://doi.org/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.

Fernandes, Won, and Martins (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 & Binderkrantz, 2021). An additional novel contribution comes from the NLP community: Kerkvliet, Kamps, and Marx (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 & 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 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, Chiarcos, McCrae, & Gracia, 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.

[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:`

TAGME (Ferragina & Scaiella, 2010) resp. WAT (Piccinno & 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, Ganea, and Hofmann (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 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, Hofer, Węcel, & Lewoniewski, 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

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

	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

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]

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] The latest version can be found at https://github.com/glaserL/unsc-ne, snapshots are distributed together with the corpus.

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

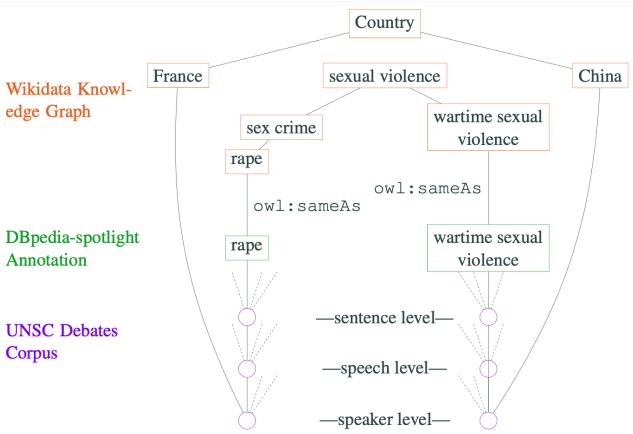

Figure 1: Example structure for two DBpedia concepts with their relationships to both the base UNSC corpus and the added Wikidata knowledge graph

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.

The two following sections will explain the different data types contained in the UNSC-NE in detail. Figure 1 provides a more visual intuition for this corpus structure.

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
- 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 3 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

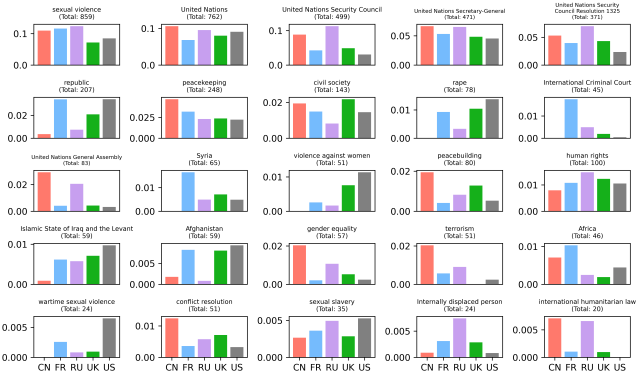

Figure 2: Most frequently used NEs by P5 countries in WPS debates

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 (Eckhard et al., 2021) 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 & 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

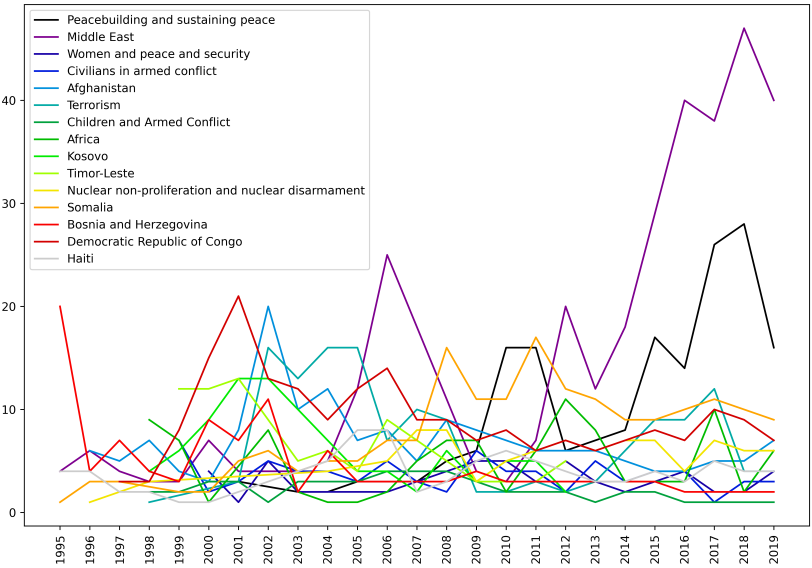

Figure 3: Top 15 agenda items of UNSC meetings

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 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 Eckhard et al. (2021) 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".

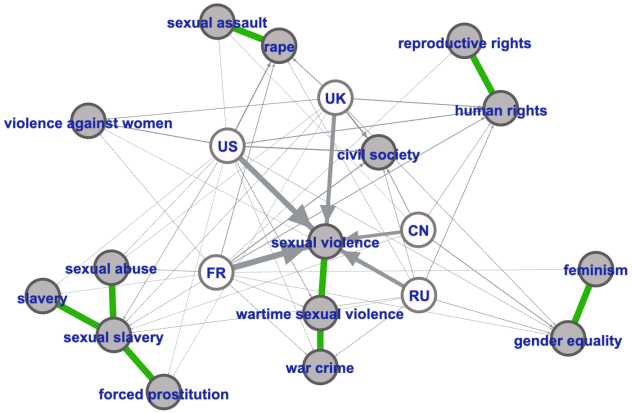

Figure 4: Network visualization of P5 countries' mentions of the most frequent policy-related NEs and directly related conceptual NEs in WPS debates. Directed edge strength represents frequency of mentions. Undirected edges (thick and green) are links in the knowledge graph.

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 4 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 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 & 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.

Acknowledgements

We thank the anonymous reviewer for their helpful comments and recommendations.

The third author acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG), project (448421482) "Verläufe von Konflikten: Die Dynamik der Argumentation im UN Sicherheitsrat“.

References

Aizenberg, E., & Binderkrantz, A. S. (2021). Computational approaches to mapping interest group representation: A test and discussion of different methods. *Interest Groups & Advocacy*, *10*(2), 181–192. Retrieved 2021-06-28, from `https://link.springer.com/10.1057/s41309-021-00121-4` doi: 10.1057/s41309-021-00121-4

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (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)* (pp. 54–59).

Amicarelli, E., & Di Salvatore, J. (2021). Introducing the PeaceKeeping Operations Corpus (PKOC). *Journal of Peace Research*. Retrieved 2021-06-28, from `http://journals.sagepub.com/doi/10.1177/0022343320978693` doi: 10.1177/0022343320978693

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research &*

Politics, *4*(2). Retrieved 2021-06-28, from http://journals.sagepub.com/doi/10.1177/2053168017712821 doi: 10.1177/2053168017712821

Bayram, A. B., & Ta, V. P. (2019). Diplomatic chameleons: Language style matching and agreement in international diplomatic negotiations. *Negotiation and Conflict Management Research*, *12*(1), 23–40. doi: 10.1111/ncmr.12142

Benoit, K., & Herzog, A. (2017). Text analysis: estimating policy preferences from written and spoken words. *Analytics, policy and governance*, 137–159.

Chatsiou, K., & Mikhaylov, S. J. (2020). Deep Learning for Political Science. In *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 1053–1078). SAGE Publications Ltd. Retrieved 2021-06-28, from https://methods.sagepub.com/book/research-methods-in-political-science-and-international-relations/i8596.xml doi: 10.4135/9781526486387.n58

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing. Retrieved 2020-05-23, from http://link.springer.com/10.1007/978-3-030-30225-2 doi: 10.1007/978-3-030-30225-2

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

Eckhard, S., Patz, R., Schönfeld, M., & van Meegdenburg, H. (2021). International bureaucrats in the un security council debates: A speaker-topic network analysis. *Journal of European Public Policy*, 1-20. Retrieved from https://doi.org/10.1080/13501763.2021.1998194 doi: 10.1080/13501763.2021.1998194

Fernandes, J. M., Won, M., & Martins, B. (2020). Speechmaking and the Selectorate: Persuasion in Nonpreferential Electoral Systems. *Comparative Political Studies*, *53*(5), 667–699. Retrieved 2021-06-28, from http://journals.sagepub.com/doi/10.1177/0010414019858964 doi: 10.1177/0010414019858964

Ferragina, P., & Scaiella, U. (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* (p. 1625). ACM Press. Retrieved 2021-08-16, from http://portal.acm.org/citation.cfm?doid=1871437.1871689 doi: 10.1145/1871437.1871689

Gray, J., & Baturo, A. (2021). Delegating diplomacy: Rhetoric across agents in the United Nations General Assembly. *International Review of Administrative Sciences*. Retrieved 2021-06-28, from http://journals.sagepub.com/doi/10.1177/0020852321997560 doi: 10.1177/0020852321997560

Hellmann, S., Hofer, M., Węcel, K., & Lewoniewski, W. (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* (p. 1-15).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1212303 doi: 10.5281/zenodo.1212303

Jurafsky, D., & Martin, J. H. (2018). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd draft ed.). Retrieved 2018-07-11, from https://web.stanford.edu/~jurafsky/slp3/

Kerkvliet, L., Kamps, J., & Marx, M. (2020, May). Who mentions whom? recognizing political actors in proceedings. In *Proceedings of the second parlaclarin workshop* (pp. 35–39). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.parlaclarin-1.7

Kolitsas, N., Ganea, O.-E., & Hofmann, T. (2018). End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 519–529). Association for Computational Linguistics. Retrieved 2021-08-16, from http://aclweb.org/anthology/K18-1050 doi: 10.18653/v1/K18-1050

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

Lee, S. J., Liu, H., & Ward, M. D. (2019). Lost in Space: Geolocation in Event Data. *Political Science Research and Methods*, *7*(4), 871–888. Retrieved 2021-06-28, from https://www.cambridge.org/core/product/identifier/S2049847018000237/type/journal_article doi: 10.1017/psrm.2018.23

Macfarlane, E. K. (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.

Nardulli, P. F., Althaus, S. L., & Hayes, M. (2015). A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, *45*(1), 148–183. Retrieved 2021-06-28, from http://journals.sagepub.com/doi/10.1177/0081175015581378 doi: 10.1177/0081175015581378

Piccinno, F., & Ferragina, P. (2014). From TagME to WAT: A new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation* (pp. 55–62). ACM Press. Retrieved 2021-08-16, from http://dl.acm.org/citation.cfm?doid=2633211.2634350 doi: 10.1145/2633211.2634350

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

Schoenfeld, M., Eckhard, S., Patz, R., Meegdenburg, H. V., & Pires, A. (2019). *The UN Security Council Debates.* https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH. Harvard Dataverse. Retrieved 2021-06-29, from https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH doi: 10.7910/DVN/KGVSYH

True, J., & Wiener, A. (2019). Everyone wants (a) peace: the dynamics of rhetoric and practice on 'women, peace and security'. *International Affairs*, *95*(3), 553–574.

Correspondence

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

Ronny Patz
Hertie School
patz@hertie-school.org

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