Proceedings of the 2^{nd} Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022)
Potsdam, Germany

Ines Rehbein ♠, Gabriella Lapesa ♣, Christopher Klamm ♠, Simone Ponzetto ♠
University of Mannheim ♠, University of Stuttgart ♣

Sep 12, 2022

Conference Program

September, 12

9:10 – 9:15 Welcome

9:15 – 10:15 Oral Session 1

10:15 – 10:45 Coffee Break

10:45 – 11:20 Oral Session 2

11:20 – 12:15 Poster Session

12:15 – 13:30 Lunch Break

13:30 – 14:10 Oral Session 3

14:10 – 15:30 Panel

Veronika Batzdorfer, Digital Society Observatory, GESIS
Valentin Gold, Center of Methods in Social Sciences, University of Göttingen
Camille Roth, Centre Marc Bloch, Berlin
Henning Wachsmuth, Computational Social Science Group, University of Paderborn

15:30 – 16:00 Coffee Break

16:00 – 17:00 Invited Talk

David Jurgens, University of Michigan

17:00 – 17:05 Closing & Good-Bye

How toxic is antisemitism? Potentials and limitations of automated toxicity scoring for antisemitic online content

Helena Mihaljević
HTW Berlin
mihalje@htw-berlin.de

Elisabeth Steffen
HTW Berlin
steffen@htw-berlin.de

Abstract

The Perspective API, a popular text toxicity assessment service by Google and Jigsaw, has found wide adoption in several application areas, notably content moderation, monitoring, and social media research. We examine its potentials and limitations for the detection of antisemitic online content that, by definition, falls under the toxicity umbrella term. Using a manually annotated German-language dataset comprising around 3,600 posts from Telegram and Twitter, we explore as how toxic antisemitic texts are rated and how the toxicity scores differ regarding different subforms of antisemitism and the stance expressed in the texts. We show that, on a basic level, Perspective API recognizes antisemitic content as toxic, but shows critical weaknesses with respect to non-explicit forms of antisemitism and texts taking a critical stance towards it. Furthermore, using simple text manipulations, we demonstrate that the use of widespread antisemitic codes can substantially reduce API scores, making it rather easy to bypass content moderation based on the service's results.

1 Introduction

The current COVID-19 pandemic has been accompanied by an increase in insults, hostility, and hate speech, often directed at Jews who (once again) have been singled out as one of the main culprits in times of crises. A recent large scale study of multiple online platforms reveals that "almost 35% of all posts mentioning Jews or Jewishness expressed negativity toward Jews", with toxic speech against Jews amounting for two to five percent of the posts in some forums (Cohen et al., 2021). With regard to the current 'infodemic', antisemitism is of special relevance as it shares relevant features and is often deeply intertwined with conspiracy theories: Both are based on simplifying forms of personification in combination with a Manichean worldview and the ontological construction of group identities (Haury, 2002). The hostility towards Jews and other targeted groups expressed in these narratives has a negative impact not only on digital spheres but also reaches out to the real world, amplifying verbal and physical acts of violence. It is thus of great importance for a variety of stakeholders such as content moderators, researchers or NGOs monitoring societal developments to have access to tools for automated detection of antisemitic online content.

Despite the current rise of antisemitic conspiracy theories, and the hateful, toxic characteristics of antisemitism, the phenomenon is still under-explored in large-scale research of online content in general, and hate-speech in particular (Steffen et al., 2022). To the best of our knowledge, there are currently no services for automated detection of antisemitic content. However, progress has been made in form of datasets, code packages, and production-ready web services regarding the recognition of other linguistic phenomena intersecting with antisemitism, such as hate speech and toxic language. Perspective API, a free service created by Jigsaw and Google's Counter Abuse Technology team, is one such widely used technology. It allows the detection of abusive content by computing scores for different attributes such as toxicity, insult or threat. Perspective API could thus provide a low-threshold approach to detect certain forms of antisemitic speech and include it in monitoring and moderation efforts. This paper aims to explore the possibilities of the service with regard to this objective.

Accordingly, we address the following research questions:

- RQ1: As how toxic are antisemitic texts rated?
- RQ2: Are encoded manifestations of antisemitism rated as toxic?

- RQ3: Are critiques of antisemitic statements rated as toxic?
- RQ4: How do modifications of antisemitic statements affect their toxicity score?

We use a set of 3.642 German-language Telegram and Twitter posts published during the COVID-19 pandemic, annotated in terms of content and stance with respect to antisemitism. We evaluate different attributes of the Perspective API that, by definition, should produce higher scores when confronted with antisemitic texts. We further analyze the scores depending on the subform of antisemitism and the stance towards it. Finally, we perform adversary attacks to assess in how far modifications of antisemitic statements influence their scores.

Content Warning This article contains examples of hateful content including offensive, insulting and threatening comments targeting Jewish people but also other individuals and groups frequently targeted by antisemitic hate speech. It might therefore cause anxiety among members of various population groups.

2 Related Work

Similar to other sociolinguistic phenomena such as offensive or abusive language, toxicity is not uniquely defined across existing research and rather used as an umbrella term. Horta Ribeiro et al. (2021) refer to it as "socially undesirable content" that includes "sexist, racist, homophobic, or transphobic posts, targeted harassment, and conspiracy theories that target racial or political groups". In Cohen et al. (2021) it is understood as "blatantly aggressive and demeaning messages about a group or person, such as dehumanization, incitement of hatred or discrimination, or justification of violence". The authors note that "toxic language includes but is not limited to hate speech", but in fact utilize machine learning models developed for hate speech detection.

Perspective API defines "rude, disrespectful or unreasonable" content that is "likely to make people leave a discussion" as toxic (Google, 2022a; Thain et al., 2017) and provides scores representing the likelihood that a reader will perceive a text as e.g. toxic. The service is used in a variety of applications such as The New York Times website and the social news platform Reddit (Google, 2022a), while also being applied by social and online media scholars. It has been used as a prefiltering method for analyses of moderation measures on Reddit (Horta Ribeiro et al., 2021), investigations of political online communities (Rajadesingan et al., 2020) such as the QAnon movement (Hoseini et al., 2021), or to identify antisemitic and islamophobic texts on 4chan that are subsequently used for the detection of hateful images via contrastive learning (González-Pizarro and Zannettou, 2022).

Perspective API also measures severe toxicity of a text as, roughly speaking, an even stronger form of toxicity. A severe toxicity score of 0.8 is chosen as the lower limit in a number of studies to preselect particularly toxic texts (Horta Ribeiro et al., 2021; Hoseini et al., 2021; Rajadesingan et al., 2020; Zannettou et al., 2020). The toxicity scores provided by Perspective API have been validated on random manually labeled text samples in e.g. Horta Ribeiro et al. (2021) and Gehman et al. (2020). Horta Ribeiro et al. (2021) compared its toxicity scores with results from HateSonar, a tool developed for the detection of hate speech and offensive language (Davidson et al., 2017), deducing that Perspective API yields better results (however, the evaluation is based on a rather small sample of data).

Despite its broadness and ambiguity, in the definition of Perspective API and beyond, the term toxicity encompasses antisemitic speech with its widely accepted operational definition as "a certain perception of Jews, which may be expressed as hatred toward Jews" (International Holocaust Remembrance Alliance, 2016). However, the design of the service already indicates the possibility of certain shortcomings with respect to detecting toxic antisemitic texts. As the developers of the Perspective API themselves point out, the very definition of toxic language has a subjective character (Borkan et al., 2019). Some labeled datasets used for training the respective models were published as part of Kaggle competitions to improve models and reduce unintended model bias (Thain et al., 2017; Wulczyn et al., 2017; Borkan et al., 2019). Labeling by a larger number of crowdworkers is used as a vehicle to make the dataset and the models trained on it more robust. However, a look at the annotated data exposes various examples confirming that this is not sufficient to label antisemitic content as toxic. For instance, the following text was annotated by 54 crowdworkers, with an av-

erage toxicity score of 0.33 (Thain et al., 2017): "The US has finally cut bait on the occultist blood suckers. Obama and Trump just drop kicked bibi down to size. This has been a long time coming and that is why the zionists wanted Hitlery to win and start ww3." Research has shown that the annotation of antisemitic content poses considerable difficulties, even for scholars with respective backgrounds (Ozalp et al., 2020; Steffen et al., 2022), thus it is plausible that annotators assess antisemitic texts in diverging ways. The task is further complicated by the fact that antisemitism is often expressed implicitly, using codes which annotators need to be familiar with (Jikeli et al., 2019) or additional context (Jikeli et al., 2022) in order to recognize them as antisemitic language. Furthermore, toxicity can be significantly lowered by undertaking minor changes such as single character-level insertions or perturbations in words associated with toxicity (e.g. 'stupid' → 'st.upid'), while the scores remain relatively high, if the statement is negated (Hosseini et al., 2017).

The service has also been shown to be biased with respect to differences in dialect, computing a significantly higher score for texts in African American English (Sap et al., 2019). Recent work demonstrated that systems tend to produce false positive bias by overestimating the level of toxicity if minorities are mentioned (Dixon et al., 2018; Hutchinson et al., 2020). Röttger et al. (2021) developed functional tests for hate speech detection models and evaluated the Perspective API and three other models. Their results indicate that all models have critical weaknesses, namely an oversensitivity to certain keywords, a common misclassification of non-hateful content (such as counter-speech), and statements including reclaimed slurs. Furthermore, the models were biased across the different target groups included in the test data (women, trans people, gay people, Black people, disabled people, Muslims, and immigrants; Jews were not included).

3 Antisemitic language

As a basic definition, we apply the working definition by the International Holocaust Remembrance Alliance of antisemitism as "a certain perception of Jews, which may be expressed as hatred toward Jews" which can be "directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities" (International Holocaust Remembrance Alliance, 2016). Among several other narrative strategies this may include calls for the killing or harming of Jews as well as false, dehumanizing, demonizing, or stereotyping accusations against Jews or the power of Jews as a collective.

We extend the working definition to also include certain subforms of antisemitism we consider as specifically relevant in the context of the COVID-19 pandemic, namely *encoded antisemitism* and *post-Holocaust antisemitism*.

Encoded forms of antisemitism are statements which do not mention Jews or the State of Israel, but instead turn generally against presumed or actual economic or political elites while deploying antisemitic codes or stereotypes, e.g. narratives holding Bill Gates or 'Big Pharma' accountable for inventing and/or benefiting from the COVID-19 pandemic and the resulting global crisis instead of explicitly mentioning and accusing Jews as initiators. Common stereotypes which are by no means exhaustive are: Actually or allegedly Jewish persons or dynasties such as Rothschild, Rockefeller, George Soros, Mark Zuckerberg, or Bill Gates; animal metaphors: e.g. octopus, snake, pig, rat; disease and cancer metaphors such as virus, germ, parasite, cancer; codes referring to the 'lying press' trope ('Lügenpresse', 'Pinocchio-Presse', 'Systemmedien'); and codes referring to a financial (Jewish) elite in control of global events ('financial elite', 'high finance', 'East coast', 'Wall street'). Note that the occurrence of a single code is typically not sufficient to label a text as antisemitic and that antisemitic codes can be articulated consciously as well as unconsciously.

Manifestations of post-Holocaust antisemitism explicitly name Jews as part of argumentation strategies which instrumentalize the victims of the Holocaust for a political agenda and at the same time shift the perpetrator-victim coordinates by undertaking relativizing Holocaust comparisons. In the context of the COVID-19 pandemic, we encounter forms of post-Holocaust antisemitism in comparisons or equations of the state measures against the pandemic with the Nazi persecution of Jews. Common examples are the use of the term 'Giftspritze' for COVID-19 vaccinations as a more or less implicit reference to the illegal and often lethal experiments performed on human beings by the Nazis, the use of the yellow star with the imprint 'Ungeimpft' (unvaccinated) by which anti-

vaccination protesters compare themselves to Jews under the Nazi regime, and references to known victims of and/or resistance fighters against the Nazi regime such as Anne Frank or Sophie Scholl.

4 Data and methods

Our data was annotated using a comprehensive annotation scheme developed as part of a research project on online antisemitism and conspiracy narratives in the context of the COVID-19 pandemic. The scheme consists of two main categories, antisemitism and conspiracy theory, and sub-labels to specify the content and stance of a message. The scheme and the annotated dataset are described in detail in Steffen et al. (2022).[1]

The annotation was performed by a team of nine researches with scientific backgrounds in political science, sociology, or data science. We annotated a corpus consisting of a few thousand messages from Telegram and Twitter. While most of the messages were labeled by a single individual, the annotation process was continuously reflected in regular discussions and a joint workshop. We furthermore evaluated inter-annotator reliability on an additional sample of 445[2] records, yielding Cohen's kappa of $\kappa = 0.84$ and thus strong agreement for the category antisemitism.

For the experiments on the Perspective API presented here, we use a dataset consisting of 3.642 texts, with ∼3.200 Telegram messages and ∼400 Twitter tweets. Around 19% of all posts were classified as articulating and/or addressing antisemitism, with ∼ 40.6% labeled as encoded antisemitism, ∼29% as Post-Holocaust antisemitism, and ∼29% as explicit forms of antisemitism. The stance expressed was predominantly affirmative (∼ 68%), while ∼ 24.4% of the texts expressed a critical stance (∼ 24.4%), and the fewest were classified as neutral or uncertain (∼7.5%).

The two sources differ not only in terms of size but also regarding the sampling approaches: While we selected tweets from a dataset about the German 'Querdenken' movement based on antisemitism-related keywords, the Telegram messages were sampled from pre-selected channels disseminating conspiracy theory content as well as critique of the anti-COVID-19 measures in Germany. It is thus not surprising that the Telegram dataset contains a greater variety of topics and thus a smaller proportion of antisemitic content (∼ 14%) than the analyzed tweets (∼56%)[3].

At the same time, almost all antisemitism-related texts from Telegram are affirmative towards antisemitism, while Twitter users in our dataset talk about antisemitism, but not necessarily support antisemitic worldviews (cf. Figure 1).

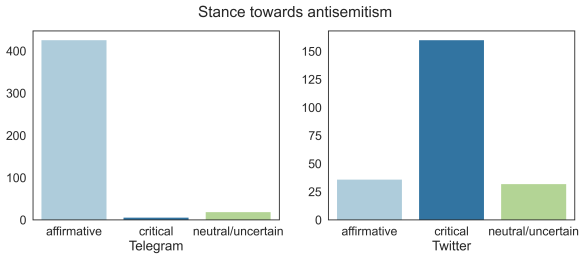

Figure 1: Stance towards antisemitic content: while only ∼16% of tweets are classified as affirmative and over ∼70% as critical, almost 95% of all Telegram messages with antisemitic content were classified as affirmative.

Encoded antisemitism makes up for almost half of the antisemitic content in our Telegram dataset (∼49%), followed by explicitly articulated antisemitism (∼ 37%), and post-Holocaust antisemitism (∼13%). On Twitter, post-Holocaust is clearly the dominant subform with 60%, followed by encoded (∼ 23%) and explicit antisemitism (17%). We suppose that the prevalence of post-Holocaust antisemitism in our Twitter data is due to the fact that users critically addressed the German 'Querdenken' movement and its comparisons of anti-COVID-19 regulations with the Nazi regime.

We have retrieved the scores for the attributes *insult*, *identity attack*, *threat*, *toxicity* and *severe toxicity* from Perspective API, since their definition (Google, 2022a) shares relevant features with antisemitic language as defined above (cf. Table 1 in Appendix). The returned score is a value between 0 and 1 that "indicates how likely it is that a reader would perceive the comment provided in the request as containing the given attribute" (Google, 2022b).

[1]The cited manuscript has been submitted for publication, thus the dataset and the annotation scheme are not yet publicly available. Until the publication, all documents and data can be made available to researchers upon request.

[2]Of the 500 texts originally selected at random, 55 were excluded because, for example, they were not in German, were too short, or were incomprehensible.

[3]Against this background, we believe that our results should not lead to the conclusion that antisemitism is generally more prevalent on Twitter than on Telegram.

5 Results

Our results indicate that our dataset has a strong toxic bias: A median severe toxicity score of ~ 0.18 clearly exceeds not only the baseline Telegram dataset compiled in Hoseini et al. (2021) with a median severe toxicity score of 0.03 but also their QAnon Telegram dataset (median: 0.07). The CDF further reveals that only $\sim 20\%$ of the texts are assigned a severe toxicity score lower than 0.1, while this holds for around 60% of all texts in the mentioned baseline set. While the Telegram subset is almost identical to all texts regarding its CDF and its median of 0.18, the Twitter subset is more toxic with a median of 0.29. Overall, the median scores range between ~ 0.18 for severe toxicity, and ~ 0.35 for insult and identity attack.

5.1 Antisemitic content

Texts classified as antisemitic have higher scores than those not classified as such with respect to all attributes. As shown in Figure 2, the median scores are around twice as high for texts containing antisemitism, with the greatest difference for severe toxicity (0.16 versus 0.35) and identity attack (0.33 versus 0.7). These findings support our hypothesis that texts with antisemitic content share relevant features with messages classified as threatening, toxic, etc. by the Perspective API. Furthermore, these types of texts often contain other insults and threats, which further contributes to the increase of their scores.

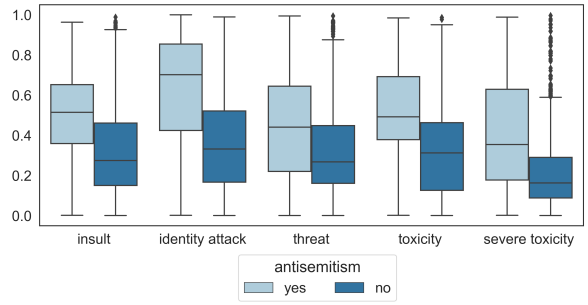

Figure 2: Distribution of scores in relation to the presence of antisemitic content.

5.2 Subforms of antisemitism

Explicit forms of antisemitism rank highest in all categories with medians between 0.56 and 0.83 (identity attack), while those for encoded forms of antisemitism range between 0.29 and 0.53. This supports our hypothesis that Perspective API generally does recognize antisemitic content as toxic, but finds it more difficult to recognize rather implicit forms of antisemitism. Texts communicating (about) narratives related to post-Holocaust antisemitism are ranked very similar to those classified as encoded antisemitism except for identity attack and severe toxicity, where they yield significantly higher values (0.53 vs. 0.63 and 0.29 vs. 0.35). The higher score, in particular for identity attack, might originate from more frequent mentions of the Holocaust and Nazis, which might also explain why there are no texts with a score of 0 in this category. The fact that most of these texts criticize the antisemitism of the 'Querdenken' movement indicates that these two endpoints do not perform well on capturing the stance of the messages.

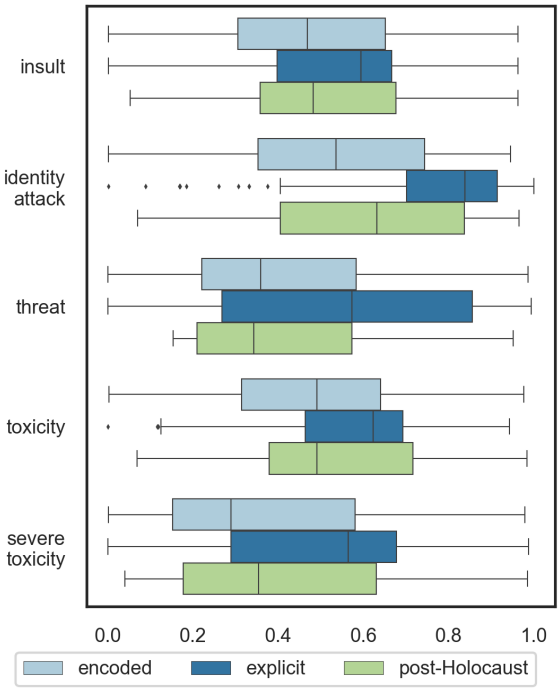

Figure 3: Distribution of scores in relation to the subform of antisemitism, with significantly higher values for explicitly antisemitic texts,

5.3 Stance towards antisemitism

As indicated in the previous exploration, Perspective API might not be able to differentiate the stance of texts even at a basic level. We thus performed experiments with single sentences and phrases extracted from our dataset that address a similar topic but with different stance. One such example is presented in Table 1. Note that the critical example does not contain any other kinds of toxic language or hate speech, thus side effects can be ruled out.

The scores for the critical example clearly outnumber the affirmative text in all categories. Furthermore, all scores are in a remarkably high range with ~ 0.7 for insult and > 0.9 for identity attack. By contrast, even the highest score for the affirmative example is lower than ~ 0.2, and in most of the categories close to 0.

As shown in Figure 4, our statistical exploration confirms the observed tendency for the entire dataset, with texts taking a critical stance towards antisemitism ranked highest in almost all categories. Texts with affirmative or neutral/uncertain stance towards antisemitism are ranked roughly equally.

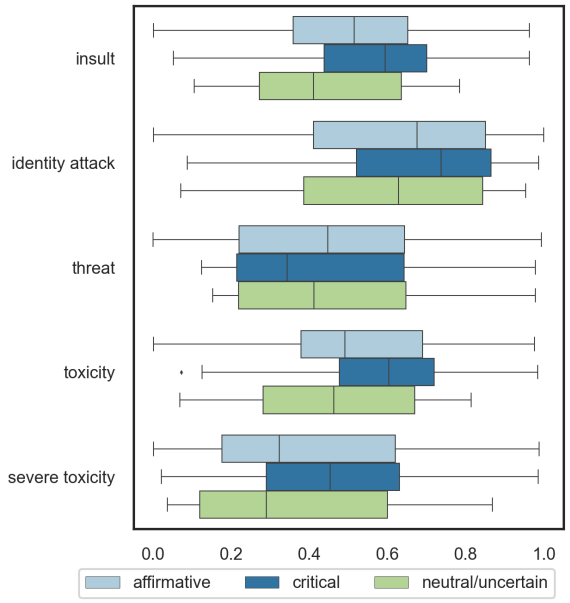

Figure 4: Texts taking a critical stance are ranked highest in all categories except threat, with medians between 0.45 and 0.74 (threat: 0.34).

5.4 Adversarial attacks: the impact of text manipulations

We are interested in how the Perspective API reacts to explicit mentions of Jews or Israel in comparison to antisemitic codes. To analyze this, we conducted two experiments:

First, we added '#Israel' and '#Juden' ('#Jews') to all texts in the corpus labeled as antisemitic ($n = 679$). Since prior work indicates an oversensitivity of hate speech detection models to certain keywords (Röttger et al., 2021), our aim was to evaluate if the presence of words explicitly related to Jewishness would affect the API's assessment.

Our second experiment focussed on antisemitic codes frequently observed in online content (Zannettou et al., 2020; Finkelstein et al., 2020). Such codes allow to express antisemitic worldview without explicitly expressing hatred against Jewish individuals or communities, thus avoiding social ostracism as well as platform bans or criminal prosecution. We believe that they can play a crucial role for spreading antisemitism, since their subtlety hampers their detection and facilitates their dissemination. To evaluate the API's assessment of encoded antisemitism, we replaced words related to Jews or Jewishness by frequently observed antisemitic codes, namely 'zionist', 'globalist' and 'satanist Freemason'[6], in texts labeled as explicitly and affirmatively antisemitic ($n = 89$). Examples of replacements are listed in Table 2 in the Appendix. In addition, we used triple parentheses, a "widely used antisemitic symbol that calls attention to supposed secret Jewish involvement and conspiracy" (Zannettou et al., 2020), known from online communication in English-language context.

We focus on the attributes identity attack, toxicity and severe toxicity since we expect these to be affected most by the performed manipulations.

Figure 5 shows the difference between the scores achieved when adding either '#Israel' or '#Juden' to the end of a text and a text's original score, in relation to the length of a text. Clearly, the effect is negatively correlated with the text length, with longer texts ($> 2,000$ characters) being almost not affected at all by the performed manipulation. Shorter texts, however, can be heavily affected. In almost all cases, adding one of the two expressions yields an increased score, which can grow up to, e.g., 0.7 for identity attack. For all three attributes, the difference is larger when adding '#Juden' than '#Israel'. Moreover, the effect is least pronounced for the attribute toxicity (mean: 0.04 for '#Israel' and 0.08 for '#Juden'), and strongest for identity attack (mean: 0.13 for '#Israel' and 0.2 for '#Juden'). The latter is not so surprising given the fact that both additions are strongly related to Jewish identity. Assuming a lower threshold of 0.5 for a text to be further analyzed by scientists or monitored by content moderators, this would imply an increase from 72% to 92% resp. 96% for identity attack, a significantly smaller increase from 49% to 57% resp. 68% for toxicity, and a growth from 40% to 48% resp. 59% for severe toxicity.

[6]German words: 'Zionist', 'Globalist', 'satanistischer Freimaurer'

text	insult	identity attack	threat	toxicity	severe toxicity
The Holocaust was unique in its contempt for humanity and its consequences for the world community.[4]	0.67	0.93	0.85	0.79	0.89
The Holocaust did not happen.[5]	0.05	0.07	0.15	0.13	0.04

Table 1: Expression of the historical significance of the Holocaust is assessed as particularly toxic, while its negation is rated with very low scores.

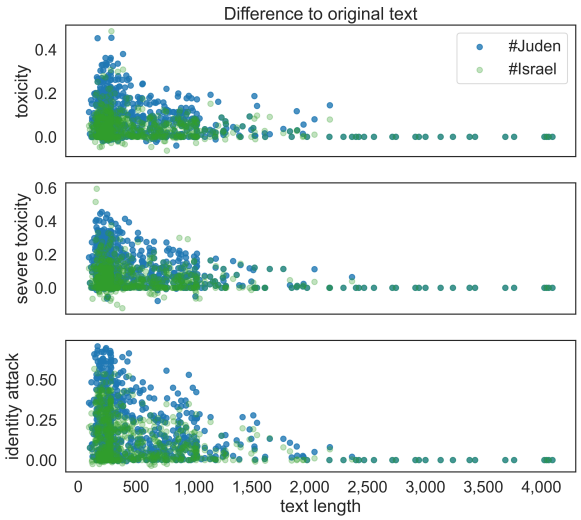

Figure 5: Difference between scores of manipulated and original message, where manipulations consist of adding one of the two string '#Israel' and '#Juden' to the end of the text ($n = 679$).

Our second experiment indicates that using antisemitic codes instead of directly mentioning Jews decreases the scores in most of the cases. Figure 6 shows that the code 'Globalist' has the strongest decreasing effect, making it an attractive term for users interested in disseminating antisemitic content without being moderated or banned from a discussion. Using the code 'Zionist' or adding triple parentheses has some decreasing effects as well, though not as strong as 'Globalist'. In some cases, we observed that using codes actually had an increasing effect on the respective scores. This was mainly the case for the code 'satanistischer Freimauer'. We assume that this is due to the negative connotations of the adjective 'satanistisch'. The effects of manipulations depend on the score of the original message, with those already assessed with a score near 1 being least effected. Interestingly, toxicity is least affected by the manipulations.

Considering again 0.5 as a threshold for all three attributes, the codes 'globalist' and 'zionist' signif-

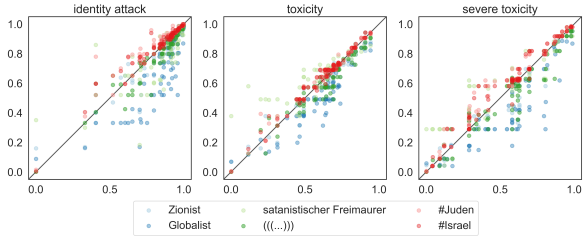

Figure 6: Score of original message vs. manipulated text ($n = 89$): replacing words related to Jewishness by codes decreases the scores in most cases.

	identity attack	toxicity	severe toxicity
original text	85	69	60
Zionist	84	59	50
Globalist	73	49	38
satanistischer Freimaurer	86	76	67
(((...)))	85	66	55
#Israel	87	72	64
#Juden	87	73	67

Table 2: Number of explicitly antisemitic texts with affirmative stance with scores above 0.5 ($n = 89$) after applying different manipulations.

icantly reduce the number of detected messages, while the usage of triple parentheses has a rather small reducing effect, and satanist Freemason as well as appending #Israel or #Juden yield a clear increase (cf. Table 2).

6 Summary

We analyzed Perspective API's assessment of the toxicity of antisemitic online content using a German-language Telegram and Twitter dataset. We conducted two experiments to examine the sensitivity of the API towards the direct mention of Jews and Israel, compared to cases in which these terms were replaced by different codes.

Regarding RQ1, texts with antisemitic content were generally scored as more toxic than texts

without such content, with median scores approximately two times higher. This indicates that texts with antisemitic content share relevant features with texts rated as toxic by the API. We observed the greatest differences between the positive and negative class regarding identity attack and severe toxicity.

Explicit forms of antisemitism are rated as more toxic, more threatening, etc. than forms of encoded antisemitism. In combination with our findings for RQ1, this indicates that Perspective API generally does recognize antisemitic content as toxic, but finds it more difficult to recognize rather implicit forms of antisemitism.

Texts taking a critical stance towards antisemitism are rated with higher scores compared to both texts with a neutral/uncertain and an affirmative stance, the latter two being similarly rated. As demonstrated with two qualitative examples, a statement clearly expressing Holocaust denial is not rated as toxic, while a statement clearly critical of the Holocaust receives very high scores. With respect to RQ3, our findings demonstrate that Perspective API is not able to differentiate the stance of texts even at a basic level.

To assess the effect of modifying antisemitic statements, we performed two experiments. Adding direct mentions of Jews or Israel to the end of a text resulted in an increase of scores, particularly for shorter posts, indicating that Perspective API is sensitive to the use of identity-group words without necessarily taking into account their textual context. In contrast, replacing direct mentions of Jews resulted in a decrease of scores in most of the cases. We observed the strongest decreasing effect for the code 'Globalist', followed by 'Zionist' and the use of triple parentheses around words like e.g. 'Jude' (Jew) or 'jüdisch' (Jewish). In some cases, the use of codes actually resulted in an increase of scores. This was mostly the case for the code 'satanistischer Freimaurer' (satanic Freemason), probably due to the negative connotation of the adjective 'satanisch'.

7 Discussion

Perspective API is already widely used in moderation and monitoring of comments but also as a tool in the study of online communication, including antisemitic speech. Our findings indicate that on a basic level, Perspective API recognizes antisemitic content as toxic. When taking a closer look, however, our investigation reveals several limitations and critical weaknesses of the service, both for research and content moderation tasks. While it reacts to explicit forms of antisemitism, it will most likely miss rather subtle and implicit forms. According to our results, a lot of texts classified as antisemitic would be neglected by research projects with a toxicity threshold of 0.8 for data collection. Even with a lower threshold of 0.5, more than half of the texts expressing encoded or post-Holocaust antisemitism, and around a third of explicitly antisemitic texts in our corpus, would not be considered. This indicates that Perspective API is able to detect only the most blatant manifestations of antisemitism. This is a severe limitation of the API considering the implicitness of antisemitism, its often encoded character, but also regarding forms such as post-Holocaust antisemitism which primarily function via self-victimization instead of direct attacks against Jewish individuals or communities.

Furthermore, the API clearly struggles with correct stance interpretation. This is a critical weakness, for using the service to build research corpora and even more for the task of content moderation. Our results indicate that Perspective API is more sensitive to (potentially harsh) critiques of antisemitism rather than to affirmative antisemitic statements. This calls for further critical research and evaluation, also with regard to the impact of this bias e.g. for content moderation in social media, since such a bias penalizes counter-speech and critical discourse about antisemitism.

Last but not least, our adversarial attacks against the API have demonstrated that even simple text manipulations can noticeably influence the scores. On the one hand, the service showed a bias towards the presence of identity-related keywords such as 'Juden' or 'Israel', assigning higher scores to texts where these words were added (cf. (Jigsaw, 2021; Röttger et al., 2021). This bias can negatively affect content moderation processes since it skews the focus towards identity-related phrases independent of their context.

On the other hand, it takes only simple manipulations in order to noticeably decrease the assigned scores. This makes it rather easy to bypass content moderation based on the API's results by using simple and known antisemitic codes, e.g. replacing terms like Jew with 'Globalist'. This facilitates the inconspicuous expression and dissemination of antisemitism and undermine monitoring efforts

as well as moderation policies - a problem which should not be underestimated regarding the strategical behaviour of users on online platforms to circumvent regulation and moderation policies assisted by machine learning technologies. Weimann and Am (2020) have analyzed the 'new language' of Right-wing extremists, a language which has partly emerged as a direct counter-reaction to the research initiative behind Perspective API. "To prevent violating the abuse policies of social media platforms and also to avoid detection by automatic systems like Google's Conversation AI, Far-right extremists have begun to use code words (a movement termed Operation Google) and thus a new type of hateful online language appears to be emerging: The systematic use of innocuous words to stand in for offensive racial slurs." (Weimann and Am, 2020)

This clearly shows that human language is not only dynamic, it is also used strategically, especially in contested spaces of social and political communication. Both aspects make it difficult for automated tools to meet their claims of making those spaces less toxic, and more inclusive. So when using these tools, be it for research, monitoring, or moderation, we need to be fully aware of the limitations these tools bring along.

When using the API for research purposes, whether for sample selection or for analyzing the toxicity of online narratives, we believe that researchers should not rely solely on the automated assessment provided by the API. Rather, we recommend a thorough manual review of the obtained results before processing them further. The same applies to its use for content moderation. However, evaluating Perspective's results itself can be a challenging task. We believe that this requires continuous training for e.g. content moderators to enable them to recognize antisemitism in its various shapes. Existing annotation efforts regarding antisemitic online content consistently show that classifying antisemitism is a complex and challenging task that depends on many factors such as the context of a message or the background knowledge of the annotators (Ozalp et al., 2020; Jikeli et al., 2022). The constant evolvement of new antisemitic codes adds to this complexity. A profound and constantly updated knowledge of these codes is thus of crucial importance. Last but not least, content moderation should be aware of the API's potential hypersensitivity towards certain keywords, and to expressions of counter-speech. Relying on API scoring might result in unintended punishment of counter-speech, providing another reason for continuous manual sample checks of results. The hypersensitivity towards keywords might result in unintended regulation of educational content, as incidents on a number of social platforms have shown (Sales, 2021). Furthermore, easily accessible feedback procedures for users should be provided to allow for early error correction and monitoring. Last but not least, transparency of the algorithms and tools used for automated detection needs to be increased, to raise awareness of the potentials and limitations of such tools, and to foster research into their strengths and weaknesses so that they can be improved.

References

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500. ACM.

Katie Cohen, Lisa Kaati, and Björn Pelzer. 2021. Antisemitism in Social Media. Conspiracies, Stereotypes, and Holocaust Denial. Technical Report FOI-R–5198–SE, Sweden.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Joel Finkelstein, Pamela Paresky, Alex Goldenberg, Savvas Zannettou, Lee Jussim, Denver Riggleman, John Farmer, Paul Goldenberg, Jack Donohue, and Malav H. Modi. 2020. Antisemitic Disinformation: A Study of the Online Dissemination of Anti-Jewish Conspiracy Theories. Technical report, The Network Contagion Research Institute.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics.

Felipe González-Pizarro and Savvas Zannettou. 2022. Understanding and Detecting Hateful Content using Contrastive Learning. *Preprint*. ArXiv: 2201.08387.

Google. 2022a. About the API - Attributes and Languages.

Google. 2022b. About the API - Score.

Thomas Haury. 2002. *Antisemitismus von links: kommunistische Ideologie, Nationalismus und Antizionismus in der frühen DDR*, 1 edition. Hamburger Ed, Hamburg.

Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_donald and r/Incels. In *Proceedings of the ACM on Human-Computer Interaction (CSCW2)*, pages 1–24.

Mohamad Hoseini, Philipe Melo, Fabricio Benevenuto, Anja Feldmann, and Savvas Zannettou. 2021. On the Globalization of the QAnon Conspiracy Theory Through Telegram. *Preprint*. ArXiv: 2105.13020.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *Preprint*. ArXiv: 1702.08138.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

International Holocaust Remembrance Alliance. 2016. The working definition of antisemitism.

Jigsaw. 2021. Unintended Bias and Identity Terms.

Gunther Jikeli, Damir Cavar, and Daniel Miehling. 2019. Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism. *Preprint*. ArXiv: 1910.01214.

Günther Jikeli, Damir Cavar, Weejeong Jeong, Daniel Miehling, Pauravi Wagh, and Denizhan Pak. 2022. Toward an AI Definition of Antisemitism? In Monika Hübscher and Sabine Von Mering, editors, *Antisemitism on Social Media*, 1 edition, pages 193–212. Routledge.

Sefa Ozalp, Matthew L. Williams, Pete Burnap, Han Liu, and Mohamed Mostafa. 2020. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society*, 6(2).

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58. ArXiv: 2012.15606.

Ben Sales. 2021. Are social media platforms banning holocaust education along with hate speech? *The Times of Israel*. (Accessed on 08/22/2022).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Elisabeth Steffen, Helena Mihaljević, Milena Pustet, Nyco Bischoff, María do Mar Castro Varela, Yener Bayramoğlu, and Bahar Oghalai. 2022. Codes, patterns and shapes of contemporary online antisemitism and conspiracy narratives - an annotation guide and labeled German-language dataset in the context of COVID-19. *Preprint*.

Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity. figshare. Dataset.

Gabriel Professor Weimann and Ari Ben Am. 2020. Digital Dog Whistles: The New Online Language of Extremism. *International Journal of Security Studies (Vol. 2 : Iss. 1 , Article 4)*, page 24.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *12th ACM Conference on Web Science*, pages 125–134, Southampton United Kingdom. ACM.

A Appendix

Score type	API Definition
insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
identity attack	Negative or hateful comments targeting someone because of their identity.
threat	Describes an intention to inflict pain, injury, or violence against an individual or group.
toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
severe toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

Table 3: Attributes from Perspective API that, based on the provided definitions, should react with higher scores towards texts with antisemitic content.

original word	triple parentheses	code: Zionist	code: Globalist	code: satanistischer Freimaurer
Jude	(((Jude)))	Zionist	Globalist	satanistischer Freimaurer
JÜDISCH!	(((JÜDISCH!)))	ZIONISTISCH!	GLOBALISTISCH!	SATANISTISCH!
Antifajuden	(((Antifajuden)))	Antifazionisten	Antifaglobalisten	satanistische Antifafreimaurer
Judenschwein	(((Judenschwein)))	Zionistenschwein	Globalistenschwein	satanistisches Freimaurerschwein

Table 4: Examples illustrating the modifications performed in texts labeled as explicitly antisemitic and containing words referring to Jews and Jewishness.

Why Justifications of Claims Matter for Understanding Party Positions

Nico Blokker[1], **Tanise Ceron**[2], **André Blessing**[2], **Erenay Dayanik**[2], **Sebastian Haunss**[1], **Jonas Kuhn**[2], **Gabriella Lapesa**[2], and **Sebastian Padó**[2]

[1]SOCIUM, University of Bremen, Germany
[2]IMS, University of Stuttgart, Germany
✉ blokker@uni-bremen.de, tanise.ceron@ims.uni-stuttgart.de

Abstract

Positional analyses in political science target the identification of (dis)similarities between parties based on their stance on a given policy or political demand (*claim*) – which are, unsurprisingly, a well explored source of information for text-as-data approaches to this task. Political actors, however, do not only make claims regarding a given policy: they often provide justifications (*frames*) for their claims. Frames are likely to provide novel perspectives on positional analysis: they have, however, been largely neglected so far. Our work fills this gap: In a first experiment, we show how including political frames in the analysis leads to a more accurate picture using categorical (manually labelled) data from two datasets: i) the Manifesto dataset from the 2021 German federal election, and ii) a more structured dataset extracted from the German voting advice application "Wahl-O-Mat". In a second experiment, we investigate whether transformer-based language model representations are able to infer party relations from the textual information, independent of the annotation. Our approach is a) able to identify relevant differences between claims and frames and b) represent the political spectrum when applied to the more structured dataset.

1 Introduction

The analysis of political texts faces the challenge of growing corpora paired with the continuing need for fine-grained analyses (Wiedemann, 2016). This includes, for example, the analysis of political claims (Koopmans and Statham, 1999) and frames (Benford and Snow, 2000). The former can be understood as demands made by politically motivated actors and the latter as the underlying reasoning they put forward, e.g., to persuade or inform voters through programmatic documents. This paper explores the interplay between claims and frames

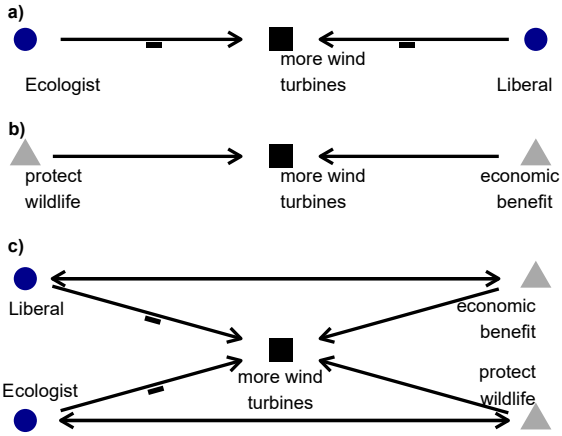

Figure 1: Example constellations of claims (squares), frames (triangles), and parties ('actors', circles). a) Claim–actor representation, b) claim–frame representation, c) joint representation of actors, claims, and frames

in capturing (dis)similarities between parties (*positional analysis*). More specifically, we treat claims as propositions stated by the parties, and frames as the justification(s) given to a certain claim. Figure 1 shows a concrete example: Both ecologists and liberals might reject the call (claim) for more wind turbines (1a); their different world view only becomes apparent when the frames behind the claims – wildlife protection and economic concerns, respectively – are considered (1b). In order to give a meaningful orientation for the voter both levels must be taken into account (1c; for another example see Brenneis and Mauve 2021). Yet, positional analysis has so far focussed on claims, neglecting the impact of frames.

This is not just a theoretical argument: recent empirical results suggest that voters feel more informed when the different propositions are enriched by supporting arguments (Brenneis and Mauve, 2021). This implies that these arguments carry additional information compared to purely positional pledges. Positional studies are conducted

within the framework of the proximity model of voting, according to which voters favor parties whose position is closest to their own standpoint (Downs, 1957; Black, 1958; Enelow and Hinich, 1984). This idea builds on the presuppositional assumptions that a) voters are not only familiar with the propositions (or claims) of elective parties but also b) that these propositions sufficiently represent the partisan viewpoint on a given policy (for a critical discussion see Budge 1994; Budge et al. 2001).

An informed voter may consult the corresponding programs contained in party manifestos to learn about the various viewpoints. However, few voters read these documents to get a picture of the electoral landscape (Budge, 1987; Bara, 2006). Instead, they increasingly resort to voting advice applications (VAAs), such as VoteCompass[1] or the German Wahl-O-Mat (Marschall 2005; Marschall and Schrenk 2021, for an overview see Van Camp et al. 2014). These applications apply the idea of the proximity model to provide voting recommendations by matching the user's policy preferences (propositions) with those of various political parties (Wagner and Ruusuvirta, 2012). Crucially, they go one step further and supply the supporting argumentation for a certain proposition (Marschall and Schrenk 2021; Brenneis and Mauve 2021). Under a framing point of view, a party's argument signifies how the proposition fits into a bigger 'narrative' (Entman 1991; e.g., 'environmental conservation') and is aimed to sway public opinion in their favor (Chong and Druckman, 2007; Slothuus and de Vreese, 2010) and/or mobilize collective action (Benford and Snow, 2000).

Summing up, both the literature on framing and widely used voting applications indicate the crucial role of arguments in party positioning. The natural question to ask is then: What is the (empirical) added value of including them in the positional analysis? We tackle this question by comparing party (dis)similarities based on a) shared usage of claims with b) shared usage of frames.

The comparison is carried out using two datasets from different textual genres. The first dataset contains manually annotated category information for both claims and frames related to the Covid-19 discourse in Germany as found in party manifestos of 2021 (MaCov21). Categories represent an abstraction from the text by assigning a theme/topic to the passage in question (e.g., "economic benefit" in Figure 1, Mayring and Fenzl 2019).

The second dataset is based on the German VAA 'Wahl-O-Mat' (WoM21) and contains annotations regarding the category of claims. It also provides justifications for the stances taken by the parties, but these justifications remain uncategorized (see Table 1 for an example).

Our study presents two experiments. In the first one, we ask whether we can computationally extract party clusters equally well from claim or frame categories (Section 4). In the second one, we ask how this compares to the use of the corresponding textual segments (Section 5). Here, the two corpora correspond to two scenarios differing in the amount of information available: in (i), frames are grouped by the annotated frame category when measuring text similarity; in (ii), frames are only grouped by the claim category, without the need for annotated frame categories. By comparing these two scenarios, we aim at understanding whether an analysis of frames can be carried out automatically without annotated categories, which would facilitate the analysis of texts by political scientists.

Contributions. This paper makes three contributions: First, we empirically assess whether connecting frames to claims in fact offers additional information. Results indicate that the two appear to capture two different dimensions of the data and thus complement each other. Secondly, we lay out how methods from computational argument mining may scale up this task: In a first step, we confirm that there *are* differences derived from the claims and frames from the MaCov21 dataset using a transformer-based language model. Here, we deploy Sentence-BERT (Reimers and Gurevych, 2019) to extract text representations of claims and frames and measure the semantic similarity among them. In a second step, we observe that the more informative the scenario, the more in line clusters are with the ones emerging from the categorical data. However, we also find that the more structured the dataset, the less informative the labels need to be for the 'optimal' clustering. Lastly, we release the dataset which our observations are based on.

2 Background

The present paper utilizes the political claims analysis (PCA) framework (Koopmans and Statham, 1999, 2006), in which actors are linked to the *politically motivated* demands or actions they make (Panel A, Figure 1). PCA is typically applied to

[1] https://votecompass.com/

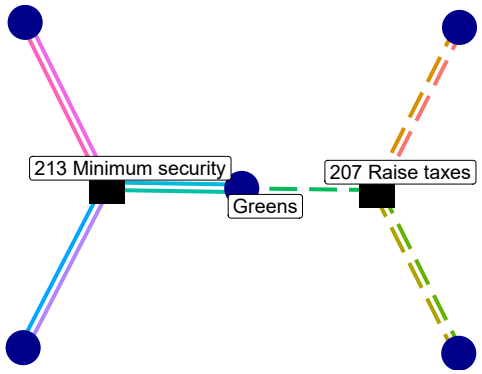

Figure 2: Claim–Actor representation as bipartite network based on categorical MaCov21 data (claims are squares and parties circles). Dashed lines indicate opposition and solid lines support. Example built from the MaCov21 dataset (introduced below).

newspaper articles and increasingly used to extract bipartite discourse networks (Leifeld, 2016; Haunss et al., 2013; Padó et al., 2019). In the electoral context of party manifestos, political claims can be thought of as pledges and analogously analysed (Blokker et al., 2020). The definition of a political claim in combination with *frames* - broadly understood as arguments or justifications for said claims - is the conceptual foundation of the current paper (Benford and Snow, 2000). We define these concepts following their usage in political science, which can vary from the definitions of arguments, frames, and claims employed in NLP.

These concepts can also be found in the context of VAAs, however, with a different notation: Claims map to *propositions* and frames to *justifications* (Table 1). In the case of the Wahl-O-Mat, propositions are compiled by researchers and selected voters.[2] Parties are then asked to take a stance (in favour, neutral, or against) regarding the proposition and additionally to provide the reasoning for their decision. In this sense, VAAs and PCA utilize a similar data structure. As a result, the relations between parties and claims are often visualized as a bipartite network (cf. Figure 2), e.g., in guides (Hunger, 2021) or journalistic publications (Kriesel, 2017).

By and large, the positions presented in VAAs appear to align with party positions derived from expert opinions or party manifestos. Although, the comparison seems to suffer more, the less propositions or theses dedicated to a policy field are covered in the VAA (Wagner and Ruusuvirta, 2012). However, VAAs and PCA differ with regard to domains covered and depth of analysis. While the analysis of policy debates following PCA or DNA typically focuses in-depth on one policy field (see Leifeld 2020), VAAs cover a broad range of issues (Van Camp et al., 2014).

3 Datasets

We use two datasets with a similar data structure (see examples in Tables 1 and 5). The first dataset (MaCov21) was manually annotated by the authors and consists of 204 texts spans with political claims and frames related to the ongoing Covid-19 pandemic, based on the party manifestos of the 2021 federal election in Germany of 6 parties.[3] Of the 204 cases, 170 are distinct claim–frame–actor triplets and used for the subsequent analysis. The employed annotation guidelines were developed inductively based on the manifestos under scrutiny: 55 different claim categories (e.g., 'Mandatory vaccinations') and 43 frame categories (e.g., 'Deficits in the healthcare system') occur in the data set, which are distributed into several higher-level categories based on the respective policy fields they cover ('Health', 'Economy', 'Education', etc).

The Wahl-O-Mat dataset (WoM21) consists of 38 theses containing the same number of claims, which are answered and justified by 38 different parties.[4] The subset used in the present paper only views the 6 parties also contained in the MaCov21 dataset (n = 228 triplets).[5]

4 Experiment 1: Categorical analysis

We investigate the question concerning the added value of frames alongside claims for positional analyses both on a categorical and a textual level. This enables us to check to what extent claims and frames yield in fact different information when it comes to the (latent) positions of parties and whether these are consistent with theoretical and empirical expectations regarding the political spectrum in Germany (e.g., left-right scale). At the

[2] https://www.bpb.de/themen/wahl-o-mat/45292/die-entstehung-eines-wahl-o-mat/

[3] The dataset and replication files can be found here: https://github.com/nicoblokker/cpss. Annotation started before the final versions of the manifestos were available, hence small differences may exist between the annotated and final versions.

[4] The dataset and additional information can be found here: https://www.bpb.de/themen/wahl-o-mat/bundestagswahl-2021/

[5] Parties included are: The Christian Democratic Union (CDU), the Social Democratic Party (SPD), the Green Party (Greens), the Free Democratic Party (FDP), the Left Party (Left), and the Alternative for Germany (AfD).

Table 1: Example annotations from the datasets (quotes translated from German and truncated). First row shows annotations from MaCov21 the second row from WoM21. MaCov21 entry labels are assigned a code-category (213 or f208).

	Actor	Proposition			Justification	
Data set	Party	Claim text	Claim category	Polarity	Frame text	Frame category
MaCov21	Greens	That is why there is a need for a minimum short-time allowance that is independent of the industry.	213 Minimum security	+	In times of Corona it is particularly evident that short-time benefits are too low for workers with small wages.	f208 Social benefits too low
WoM21	Greens	A general speed limit is to apply on all highways.[1]	Speed limit on freeways	-	In order to reduce the number of serious accidents and enable relaxed driving without blatant differences in speed, the introduction of a general safety tempo as in all other European countries is called for [...].	*no category*

[1] Identical wording across all parties. Hence, it cannot be used meaningfully to determine textual similarity.

same time, it requires methods tailored to the analysis for each level and a suitable metric with which one can assess and measure (dis)similarity.

4.1 Methods

MaCov21 - Claims and Frames The similarity between parties' positions and justifications can easily be compared on a network level: Piecing together the three distinct elements forming a triplet (claims, frames, and parties), one may build and visualize the result as a tripartite network (Figure 3). In addition to the typical way to combine parties and propositions (Figure 2), we added justifications as a third node type. This has profound consequences for the network topology: While the bipartite network contains 7 nodes (5 actor nodes and 2 claim nodes), the tripartite network displays an additional number of 9 frame nodes. What stands out in this example is that while parties address the same claims, they rarely corroborate them with the same frames. This is a first indication that claims and frames might convey different information.

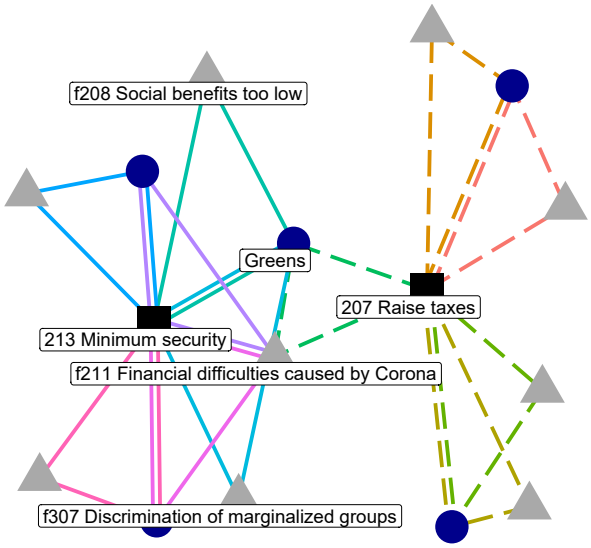

Figure 3: Claim–Actor–Frame representation as tripartite network based on categorical MaCov21 data (Claims are squares, parties circles, frames triangles). Edges express support or opposition (solid vs. dashed lines).

However, instead of directly analyzing the joint representation of parties, claims, and frames, we aim to isolate the impact of frames on the network. Therefore, to evaluate the concrete benefit of including justifications in the analysis of party positions, we tease the network apart: We measure the similarity between a) the party-claim network and b) the party-frame network as derived from the MaCov21 dataset based on the category labels. More specifically, we take the cosine similarity of the incidence matrices underlying the two graphs. The positions parties take are represented within the matrix as 1 (support), -1 (opposition; only relevant for claims), and 0 (no position). This breakdown of the joint, tripartite network into its bipartite components can be seen as a preparatory step. Firstly, it allows to establish whether there are in fact differences and secondly, it serves to gauge the nature

of these differences to guide future analyses.[6]

4.2 Evaluation

To test whether the incidence matrices are significantly different from each other, we compute Mantel correlation to test their significance (Mantel, 1967). This test allows to estimate correlations (such as Pearson's r) in the face of distance or relational data, in which observations are dependent of each other. It addresses the problem of dependency by calculating correlations on all permutations of the flattened distance matrix. The one-tail hypothesis tests whether the correlations from the permuted matrices are higher than the first observed correlation. In a second step, we then apply agglomerative hierarchical clustering procedure to the distance matrices to explore what different party-clusters can be extracted from their usage of claims and frames in terms of categories.

4.3 Results and discussion

MaCov21 The resulting similarity scores between parties are presented as heatmaps in Figure 4. The idea is to potentially uncover different aspects (or dimensions) of observed party behavior by comparing how parties manoeuvre claims and frames in their programs. The correlation between the two matrices is high and significant (Mantel statistic r = 0.68, p $<$ 0.05). However, there appears also to exist enough variance between how parties use claims and frames (r^2 = 0.46). Therefore, the clustering tree based on these scores suggests two different clusters for both representations. The clustering based solely on shared claim usage suggests a division that places the parties into their corresponding party families following a left-right pattern (Cluster 1 = AfD, CDU, FDP; Cluster 2 = Left, SPD, Greens). On the other hand, when considering the party frames representation, the clusters become more heterogeneous — at least from a ideological perspective. The first cluster contains the four mainstream parties (CDU, SPD, FDP, and Greens), while the second cluster contains the two parties at the outskirts of the political spectrum (AfD and Left). Incidentally, the partitioning based on the second cluster analysis seems better suited to identify potential coalitions. For instance, the previous governing coalition (2017–2021) consisted of CDU and SPD, while the newly formed coalition comprises SPD, Greens, and FDP.

While the correlation analysis and the clustering suggests that claims and frames indeed contain different information, the difference should not be overestimated. The correlation between matrices (in Figure 4) is still high. Additionally, these results need to be put into perspective given the small and imbalanced sample size as well as the preliminary nature of the dataset. Also, the AfD-Left cluster is somewhat unexpected given the political spectrum in Germany (see Section 6) and requires further analysis.

WoM21 We now turn to the second dataset where the analysis varies slightly for three reasons. First, WoM21 does not contain as much annotated information as MaCov21, for example, there are no frame categories and the claim statements are a standard sentence across parties. While this permits us to apply both correlation and cluster analyses to the claim level, it is not applicable to the frame level, where one has to resort to the textual representation (see Experiment 2). Secondly, we do not compare the results of the matrix distance across datasets because the MaCov data consists of categories related to the discourse around the COVID-19 pandemic whereas WoM21 comprises the parties' justifications regarding various policy fields. And finally, – although they are conceptually similar – the two datasets have different data distributions because of their nature (one being party manifestos and the other an online application). Rather than aiming for a direct comparison, we content ourselves with a comparison of how well the two approaches map to a two dimensional (political) space (Figure 5).

5 Experiment 2: Textual analysis

Whereas the first experiment makes use of only information from the annotated categories of claims and frames, this step seeks to evaluate to what extent a combination of textual data and text similarity approaches can be implemented to understand party relations in terms of their stated claims and frames. Our aim is to evaluate whether textual data alone is able to capture the proximity among parties, and therefore, aid political scientists to scale up their analyses of political texts to larger amounts of unlabeled data. To this end, we frame party characterization as a representation learning task where

[6] Analogous to the procedure underlying Figure 4 in Section 4.2, we included a comparison between the clustering trees of both bipartite and the tripartite networks in the appendix (Figure 6). The latter resembles a mixture of the former with stronger accents of the party-frame representation.

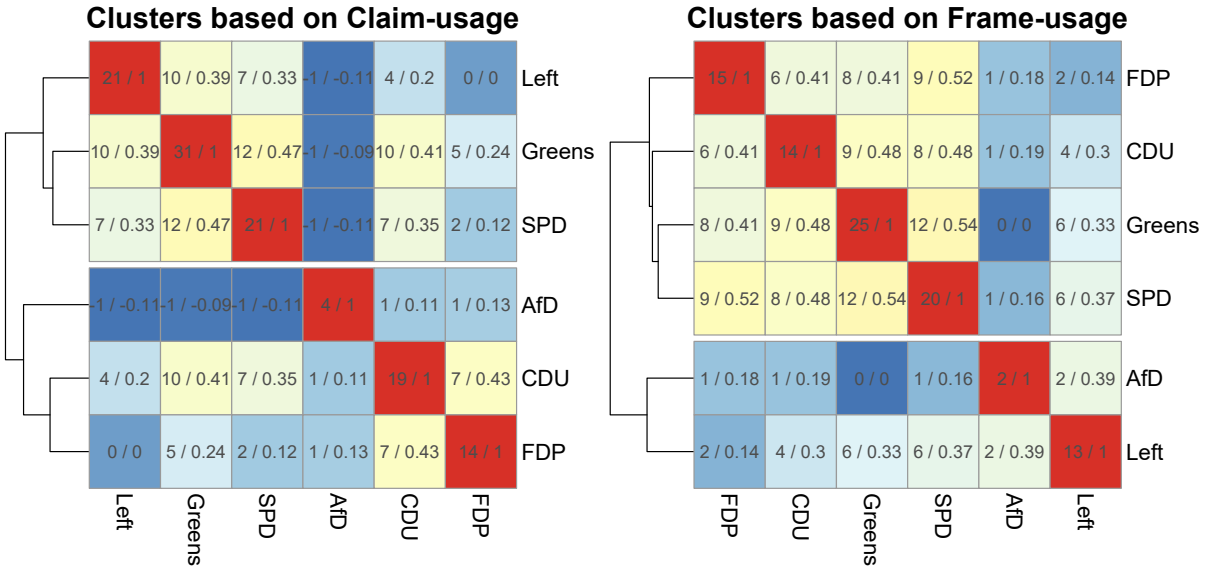

Figure 4: Heatmaps containing similarity scores between party-claim (left) and party-frame (right) representations based on categories from the MaCov21 dataset. Integers show the number of times parties agree with each other adjusted for the number of times they disagree. Floating point numbers correspond to cosine similarities. Clustering trees on the left sides of the matrices are based on cosine distance (1 minus cosine similarity), using Ward's linkage.

	MaCov21	**WoM21**
Num cat. claims	55	38
Num cat. frames	43	-
Tokens claims	9-55-190	-
Tokens frames	8-56-291	9-85-126

Table 2: Number of annotated categories for claims and frames. Minimum, mean, and maximum number of tokens for claims and frames respectively.

the parties are represented by the textual spans of their claims and frames and consecutively are compared by a standard similarity measure such as cosine. Our approach detailed under Section 5.1, however, assumes that the relevant text spans are already identified. The full automated potential of this setting therefore only unfolds when combined with efforts of claim identification (Lippi and Torroni, 2015; Padó et al., 2019) and frame extraction (Card et al., 2015; Nicholls and Culpepper, 2021).

5.1 Methods

Model for text representations We build on Sentence-BERT text representations (SBERT) (Reimers and Gurevych, 2019), a variant of BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) which aims at better encoding the (dis)similarities between sequences of text. To do so, it adds a pair- or triple-based semantic similarity fine-tuning step that uses natural language inference and paraphrasing corpora. As a result, this model is more appropriate for our analysis because of its more semantically meaningful adaptation of transformers-based language models, and its computational efficiency given that it takes into consideration the sentence representations and let aside all the token representations. Pre-trained SBERT models are available in English or in multilingual versions. Since our datasets are in German, we implement the multilingual version of SBERT.[7] [8]

MaCov - Claims For the MaCov21 dataset, since we may have multiple claims belonging to the same claim category and the same party, we first concatenate the texts that belong to the same claim category and the same party, i.e. *claim I of claim category 1* and *claim II of category 1* from party A become a single sequence. The default maximum number of tokens per sequence (128) is kept from the original SBERT since the mean number of tokens does not exceed it, as can be seen in Table 2. Longer sentences are truncated in order to avoid variation from the original fine-tuned sentences which may cause degradation of the representations. The com-

[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[8] It is also the best performing pre-trained multilingual SBERT model in the evaluation tasks overall: https://www.sbert.net/docs/pretrained_models.html.

putation of the similarity is formulated as:

$$\vec{s} = MODEL(s) \quad (1)$$

$$\text{sim}(P_1, P_2, i) = \cos(\vec{cla}_{cat}(P_1, i), \vec{cla}_{cat}(P_2, i)) \quad (2)$$

$$\text{sim}(P_1, P_2) = 1 - \frac{1}{|Cat|} \sum_i \text{sim}(P_1, P_2, i) \quad (3)$$

where the sequence *s* is fed into SBERT to retrieve the representations (Eq. 1). cla_{cat} refers to the claim category and to the way the s representations are grouped in the comparison between party pairs. Then, the cosine similarity of a claim category between two parties P_1 and P_2 is computed (Eq. 2), i.e., if party A and party B both have claims under the category *i*, the cosine similarity is computed (refer to Table 6 for examples). Finally, all accumulated scores between a party pair are averaged for sim(P_1, P_2) as Eq. 3 shows. This results in a distance matrix $\mathbb{R}^{pxp}$ with the mean average sim between parties where *p* is the number of parties (=number of manifestos).

MaCov - Frames A similar procedure is followed to compute the average distance matrix of frames in MaCov21, but as Table 3 illustrates, there are two setups for grouping frames during the similarity computation. The former is less informative because only the claim categories are considered. It follows the same steps as equations 1, 2, 3. The only difference is that Eq. 1 refers to the encoding of frames rather than claims.

On the other hand, the latter is more informative given that it takes the frame annotated categories into account. Therefore, Eq. 2 is replaced by:

$$\text{sim}(P_1, P_2, i) = \cos(\vec{fra}_{cat}(P_1, i), \vec{fra}_{cat}(P_2, i)) \quad (4)$$

where fra_{cat} refers to the frame category.

WoM - Frames We only analyse the textual data from frames since the claim statements are the same across parties. Moreover, there is no annotated data for the categories of frames, so text similarity is grouped in the less informative way, with claim categories following procedures of equations 1, 2 and 3.

Dataset(s)	Text	Group	Informative
MaCov	Claim	Claim cat.	+++
MaCov WoM	Frame	Claim cat.	+
MaCov	Frame	Frame cat.	+++

Table 3: Experimental set-up of the textual data. *Group* represents how the texts were grouped when measuring text similarity. More informative means more information from the annotations.

Categorical	Textual	MaCov	WoM
claim	claim	0.73**	N/A
claim	frame (C)	0.4	0.53**
claim	frame (F)	0.32	N/A
frame	frame (C)	0.41	N/A
frame	frame (F)	0.52**	N/A

Table 4: Mantel correlation between categorical and textual distance matrices. *(C)* means that the text was grouped by the claim categories. *(F)* means it was grouped by the frame categories. ** indicates p-value < 0.05.

5.2 Evaluation

We evaluate to which degree the different settings of text similarity computation can capture the relation between parties through Mantel's procedure (explained in 4.2). We assume that the higher the correlation between them, the better party proximity is captured with textual data. The correlation is calculated between the distance matrix built with textual data and the distance matrix derived from the categorical data which is our frame of reference given that they have been manually annotated. The comparison is carried out within datasets – the categorical distance matrix from MaCov (claims or frames) is compared with the textual distance matrix (claim or frames) from MaCov, and the categorical distance matrix built with claim positioning is compared against the textual distance matrix of frames, both from WoM.

5.3 Results and discussion

MaCov21 Our leading question is whether we can reach comparable results based on text classification as we derive from the categorical data. As Table 4 shows, the overall correlation between categorical and textual data amounts to r = 0.73 ($p < 0.05$) for claims, showing that SBERT representations are relatively good at capturing the relation among parties. For frames grouped by the

frame category, the correlation is r = 0.52 ($p < 0.05$) while for frames grouped by the respective claim category only r = 0.41 (not significant, $p > 0.05$; for examples refer to Table 6). These results show that the more structured information is provided in the grouping of frames when computing text similarity, the better clusters are formed. This points to the fact that frame annotations are still necessary for understanding party proximity. The correlation between the similarity matrix of claims and frames grouped by claims provided by SBERT correlate on a moderate level but not significantly (r = 0.40, $p > 0.05$). The correlation increases once the frames are grouped by frame categories to r = 0.49 ($p > 0.05$), corroborating the idea that claims and frames present different aspects of the parties' view as supported by the category-based analysis.

6 Annotated and predicted similarity

Finally, we compare the results for both experiments on how well they are able to mirror the political spectrum in Germany. More concretely, we use the rigth-left (RILE) score of the Comparative Manifesto Project as empirical reference point (Volkens et al., 2021).[9] To compare the results from both datasets with each other, we used classic multidimensional scaling on the corresponding distance matrices (Figure 5). The cluster analysis reveals similar clusters for both (annotated) frames and claims as observed in MaCov21 with some within-cluster deviations. Once more, the use of propositions can be clustered according to the party family. While the usage of justifications again suggests to distinguish between mainstream parties it now places both AfD and Left in their own cluster. Figure 5 reveals that for the WoM21 data – in which the frame dimension is based on textual similarity and only the claim dimension on categorical information – a three cluster solution is preferred.

Party clusters on the claim dimension (x-axis) are consistent across datasets. Differences exist in their exact position and within cluster arrangements. For example, while the CDU is further right from the FDP according to the WoM21 data (see right panel of Figure 5), they are much closer to the left cluster in the MaCov21 dataset (left panel). Similarly, the Greens are closer to the Left party in the WoM21 data, while they are closer to the SPD in the MaCov21 dataset. In the absence of categorical data we resorted to the use of SBERT to extract similarities between the justifications used by different parties. Again, we observe a first cluster of mainstream parties that is divided from the AfD and the Left. However, instead of placing the latter into the same cluster, a three cluster solution is more in line with our expectation of party alignment in Germany (see Volkens et al. 2021).

[9] According to which the parties' programs in 2021 can be aligned as follows (from left to right): The Left, SPD, Greens, FDP, CDU, AfD.

7 Conclusion

To obtain a comprehensive assessment of a political party, it is not enough to know its position on a given policy, proposition, or claim. We argue that it is also important to know how the position is justified. In this paper, we found indications that the usage of claims and frames by parties carries complementary information that might not only help us better understand how coalitions are formed across ideological divides but also provide a clearer orientation for potential voters. We evaluate the obtained results through comparison to theoretical and empirical expectations, such as the typically used left-right scale (Laver et al., 2003; Slapin and Proksch, 2008; Glavaš et al., 2017).

Furthermore, we investigate a) whether we are able to computationally extract similar party-clusters from the textual representation to the ones derived from the categorical data and b) how much information is needed from the annotations for the clustering. Results are either evaluated against manual annotations derived from party manifestos or a party annotated questionnaire regarding the positions of parties according to certain policies. While the model confirms that frames and claims vary in their semantic similarity, the clustering does not align with the category-based approach. We attribute this to the variance in the domain under study (the Covid-19 debate in Germany) and the preliminary state of the MaCov21 dataset. However, when applied to the more streamlined WoM21 dataset, we find that computational models are able to automatically extract relevant information regarding party (dis)similarities from the WoM dataset that is in inline with theoretical assumption about the political spectrum in Germany.

Given the novelty of the approach, our results are best seen as prototypical attempt to shed light into the dependencies existing in the network representation between the three nodes types (actor, claim, frame) in an partisan setting (Figure 1). This

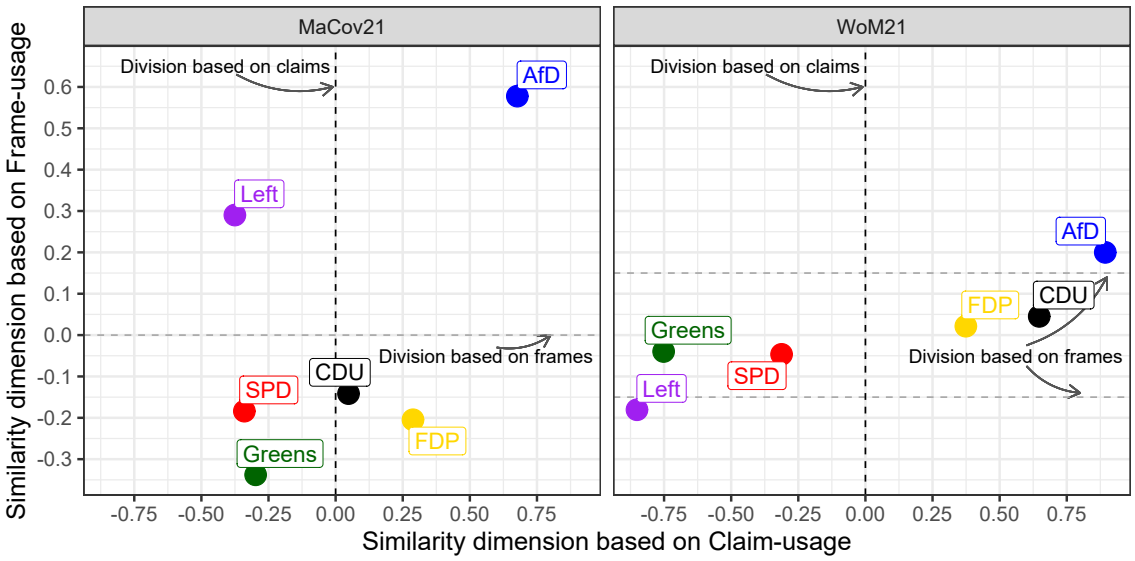

Figure 5: Similarity scores for party-claims and party-frames based on categories from the MaCov21 dataset (left) and on the categorical claim data and textual frame data from WoM21 (right) mapped to two dimensions using multidimensional scaling. Division lines indicate cluster solutions based on either claim usage or frame usage.

is reinforced by the fact that we work with a preliminary and small dataset (MaCov21). As a consequence, the results can only be seen as first indication prompting further inquiries, ideally with larger amounts of data. On a similar note, further inquiry is needed to clarify a) in what exact quality claim and frame dimensions differ and b) to what extent these cues are mirrored by automated procedures.

Nevetherless, we are confident to have found yet another promising point of contact between the political science and NLP communities, as our results demonstrate how party competition research can benefit from the interdisciplinary exchange with argument mining. Comparable results allow political scientists to scale up their analysis to verify if findings hold up across domains. Another still open question is whether the provided insights apply to other media formats as well (e.g., social media (Mendelsohn et al., 2021)).

Acknowledgments

We acknowledge funding by Deutsche Forschungsgemeinschaft (DFG) through MARDY2 (375875969) within SPP RATIO and by Bundesministerium für Bildung und Forschung (BMBF) through E-DELIB (Powering up e-deliberation: towards AI-supported moderation).

Further, we thank our student assistants Leonie Amende and Clara-Sophie Schröder for their dedicated work during annotation and their valuable input in the preparation of the codebook.

References

Judith L. Bara. 2006. The 2005 Manifestos: A Sense of Déjà Vu? *Journal of Elections, Public Opinion and Parties*, 16(3):265–281.

Robert D. Benford and David A. Snow. 2000. Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 26(1):611–639.

Duncan Black. 1958. *The theory of committees and elections*. Cambridge University Press, New York.

Nico Blokker, Erenay Dayanik, Gabriella Lapesa, and Sebastian Padó. 2020. Swimming with the Tide? Positional Claim Detection across Political Text Types. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 24–34, Online. Association for Computational Linguistics.

Markus Brenneis and Martin Mauve. 2021. ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application. In *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies, pages 3–13, Singapore. Springer.

Ian Budge. 1987. The internal analysis of election programmes. In David Robertson, Derek Hearl, and Ian Budge, editors, *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*, pages 15–38. Cambridge University Press, Cambridge.

Ian Budge. 1994. A New Spatial Theory of Party Competition: Uncertainty, Ideology and Policy Equilibria Viewed Comparatively and Temporally. *British Journal of Political Science*, 24(4):443–467.

Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum, editors. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. Oxford University Press, Oxford, New York.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dennis Chong and James N. Druckman. 2007. Framing Theory. *Annual Review of Political Science*, 10(1):103–126.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony Downs. 1957. An Economic Theory of Political Action in a Democracy. *Journal of Political Economy*, 65(2):135–150.

James M. Enelow and Melvin J. Hinich. 1984. *The Spatial Theory of Voting: An Introduction*. CUP Archive.

Robert M. Entman. 1991. Symposium Framing U.S. Coverage of International News: Contrasts in Narratives of the KAL and Iran Air Incidents. *Journal of Communication*, 41(4):6–27.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693, Valencia, Spain. Association for Computational Linguistics.

Sebastian Haunss, Matthias Dietz, and Frank Nullmeier. 2013. Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. *Zeitschrift für Diskursforschung*, 1(3):288–316.

Michael Hunger. 2021. Graph-Analyse Wahl-O-Mat Bundestagswahl 2021. *Medium*.

Ruud Koopmans and Paul Statham. 1999. Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches. *Mobilization: An International Quarterly*, 4(2):203–221.

Ruud Koopmans and Paul Statham. 2006. Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches. *Mobilization: An International Quarterly*.

David Kriesel. 2017. Ähnlichkeiten der Parteien: Erst der Computer löst das Suchbild. *FAZ.NET*.

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

Philip Leifeld. 2016. *Policy Debates as Dynamic Networks: German Pension Politics and Privatization Discourse*. Campus Verlag, Frankfurt/New York.

Philip Leifeld. 2020. Policy debates and discourse network analysis: A research agenda. *Politics and Governance*, 8(2):180–183.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, page 185–191. AAAI Press.

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

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27 2:209–20.

Stefan Marschall. 2005. Idee und Wirkung des Wahl-O-Mat. *Aus Politik und Zeitgeschichte : APuZ*, 55(51/52).

Stefan Marschall and Lea Schrenk. 2021. Der Wahl-O-Mat im „Superwahljahr“ – ein lehrendes und lernendes Tool der politischen Bildung. *GWP – Gesellschaft, Wirtschaft, Politik*, 70(2-2021):164–168.

Philipp Mayring and Thomas Fenzl. 2019. Qualitative inhaltsanalyse. In Nina Baur and Jörg Blasius, editors, *Handbuch Methoden Der Empirischen Sozialforschung*, pages 633–648. Springer Fachmedien Wiesbaden, Wiesbaden.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 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*, pages 2219–2263, Online. Association for Computational Linguistics.

Tom Nicholls and Pepper D. Culpepper. 2021. Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities. *Political Communication*, 38(1-2):159–181.

Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who Sides with Whom? Towards Computational Construction of Discourse Networks for Political Debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Rune Slothuus and Claes H. de Vreese. 2010. Political Parties, Motivated Reasoning, and Issue Framing Effects. *The Journal of Politics*, 72(3):630–645.

K. Van Camp, J. Lefevere, and S. Walgrave. 2014. The content and formulation of statements in voting advice applications: A comparative analysis of 26 VAAs. In Diego Garzia and Stefan Marschall, editors, *Matching Voters with Parties and Candidates : Voting Advice Applications in a Comparative Perspective*, pages 11–31. ECPR Press.

Andrea Volkens, Tobias Burst, Werner Krause, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, Bernhard Weßels, and Lisa Zehnter. 2021. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2021a.

Markus Wagner and Outi Ruusuvirta. 2012. Matching voters to parties: Voting advice applications and models of party choice. *Acta Politica*, 47(4):400–422.

Gregor Wiedemann. 2016. *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. Kritische Studien Zur Demokratie. VS Verlag für Sozialwissenschaften.

A Appendix

Table 5: Examples from the datasets

	Actor	Proposition			Justification	
Dataset	Party	Claim-text	Claim-label	Polarity	Frame-text	Frame-label
MaCov21	Left	Die Lizenzen für die Coronaimpfstoffe müssen freigegeben werden, damit die Impfstoff- produktion beschleunigt werden kann.	709 Einschränkung Patentschutz	+	Die Lizenzen für die Coronaimpfstoffe müssen freigegeben werden, damit die Impfstoffproduktion beschleunigt werden kann.	f102 Wichtigkeit der Impfstoffversorgung
MaCov21	Left	Die Lizenzen für die Coronaimpfstoffe müssen freigegeben werden, damit die Impfstoff- produktion beschleunigt werden kann.	709 Einschränkung Patentschutz	+	Gerade in der Pandemie zeigt sich, dass Pharmaforschung ein öffentliches Gut ist.	f106 Marktversagen
WoM21	Left	Impfstoffe gegen Covid-19 sollen weiterhin durch Patente geschützt sein.	Patentschutz für Impfstoffe	-	Was mit öffentlichen Geldern gefördert wurde, soll auch allgemein und nach sozialen Kriterien zur Verfügung stehen. Impfstoffe, die auf mit öffentlichen Mitteln geförderter Forschung aufbauen, sollen mit sozialverträglicher Patentverwertung (Equitable Licensing) an ärmere Länder und Generikaproduzenten abgegeben werden.	NA
MaCov21	CDU	Deshalb wollen wir die ge- wonnenen Erkenntnisse nutzen, um Aufgaben und Strukturen im Bevölkerungsschutz zu moder- nisieren und weiterzuentwickeln.	101 bessere med. Versorgung	+	Die Bewältigung der Corona-Krise hat die Stärken, aber auch die Schwächen im Zusammenwirken der beteiligten Institutionen verdeutlicht.	f101 Defizite im Gesundheitssystem
WoM21	SPD	Auf allen Autobahnen soll ein generelles Tempolimit gelten.	Tempolimit auf Autobahnen	+	Ein Tempolimit von 130 km/h auf Bundesautobahnen schützt die Umwelt und senkt die Unfallzahlen deutlich. Außerdem erhöht es die Reichweite von Elektrofahrzeugen deutlich - und diese werden in Zukunft eine Großteil der Fahrzeuge ausmachen.	NA

Table 6: Similarity of sentence pairs as provided by SBERT for MaCov21 and WoM21.

Dataset	Sentence 1	Sentence 2	Score
MaCov21:Claim	GRUENE= Das wollen wir ändern: mit einer zeitgemäßen, datenschutzfreundlichen digitalen Ausstattung und mit Strukturen, die die Schulen beim digitalen Lehren und Lernen wirkungsvoll unterstützen – mit kontinuierlichen Fort- und Weiter- bildungsangeboten für das pädagogische Fachpersonal sowie einem zentralen Ort der Beratung und des Austauschs zur Bildung in einer digitalen Welt. Unser Ziel ist ein moderner, engagierter Staat, der mit einer effizienten, zugänglichen Verwaltung transparent, offen und in der Lage ist, Krisen effektiv zu managen, digitale Teilhabe zu sichern und es den Bürger*innen insgesamt leicht macht, ihren Alltag zu bewältigen und ihre Rechte in Anspruch zu nehmen. Wir wollen unsere Verwaltung modernisie- ren, sie kreativer, digitaler und innovativer machen und besser ausstatten.	FDP= Die Digitalisierung von allgemeinbildenden, beruflichen und sonderpädagogischen Schulen muss ganzheitlich von der Ausstattung bis zur Nutzung gedacht werden. Wir wollen deshalb die Kompetenzverteilung zwischen den staatlichen Ebenen neu regeln und die Digitalisierung der Verwaltung vorantreiben. So muss Deutschland bei der Digitalisierung aufholen.	0.81
MaCov21:Frame	CDU= Die Pandemie zeigt, wie wichtig die internationale Zusammenarbeit bei Fragen der Gesundheit und der Gesundheitssicherheit ist. Gleichzeitig hat sich gezeigt, dass die WHO ihr zentrales Mandat in der globalen Gesundheit aufgrund mangelnder Ressourcen aktuell nur unzureichend erfüllen kann.	GRUENE= Sie soll Gesundheitssysteme weltweit stärken können, damit eine bessere Versorgung lokaler Bevölkerungen sichergestellt ist und die Prävention gegen nichtübertragbare wie übertragbare Krankheiten, deren Diagnose und die Reaktion darauf verbessert werden.	0.75
WoM21:Frame	AfD= Die sogenannte ”gendergerechte Sprache” ist eine groteske Verunstaltung der deutschen Sprache. Sie schafft keine Gleichberechtigung. Sprache darf kein Spielball ideologischer Interessen sein. Wir lehnen daher insbesondere die sogenannte ”gendergerechte Sprache” ab und sprechen uns gegen jegliche Verpflichtung aus, sie verwenden zu müssen.	DIE LINKE= Das halten wir für selbstverständlich.	0.20
WoM21:Frame	FDP= Wir Freie Demokraten fordern ein Moratorium für den Weiterbau von ”Nordstream 2”, bis die russische Führung im Fall Nawalny unabhängige und umfassende Ermittlungen gewährleistet und sich die Menschenrechtslage bessert. Die Inbetriebnahme der Pipeline ”Nordstream 2” muss in der EU gemeinsam entschieden werden. Dabei müssen auch die Interessen der Ukraine als Transitland für Energie berücksichtigt werden.	GRÜNE= „Nord Stream 2” zementiert die Abhängigkeit der EU von fossilen Energieimporten und widerspricht den EU-Klimazielen. Die Inbetriebnahme gefährdet die Ukraine. Eine überwältigende Anzahl unserer Partner in der EU hält die Pipeline für falsch. „Nord Stream 2” ist eine wichtige Einnahmequelle der autoritären Regierung Russlands und dient der Bereicherung von Präsident Putin und dessen korrupten Umfelds. Die Pipeline schadet damit auch den Interessen und der Glaubwürdigkeit deutscher Außenpolitik.	0.74

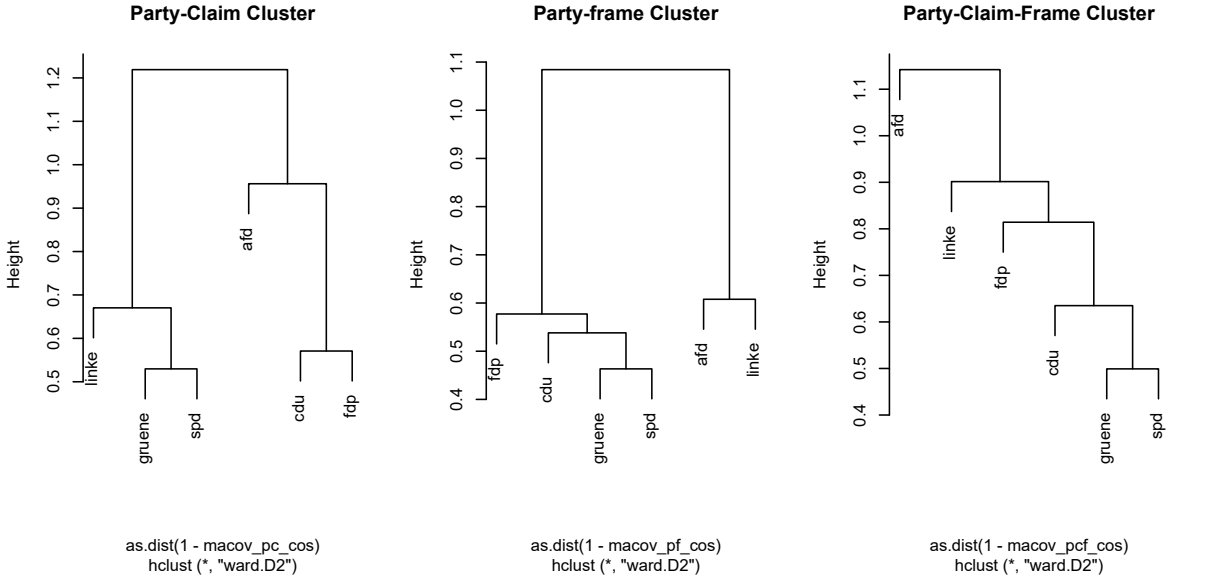

Figure 6: Clustering trees derived from the bipartite Party-Claim (left) and Party-Frame (middle) networks compared to the one derived from the tripartite network (right).

Measuring Plain Language in Public Service Encounters

Wassiliki Siskou[1], **Laurin Friedrich**[1,3], **Steffen Eckhard**[2], **Ingrid Espinoza**[1]
Annette Hautli-Janisz[4]
[1]University of Konstanz, Germany
[2]Zeppelin University Friedrichshafen, Germany
[3]University of Duisburg-Essen, Germany
[4]University of Passau, Germany

Abstract

Face-to-face interactions between public service professionals and citizens constitute an essential point of contact between the public and the state. Of central importance in these settings is the comprehensibility of the conversation in order to reduce the communicative gap between citizens and state authorities. Starting from the criteria available for written communication, we systematically investigate administrative spoken language during public service delivery and propose a plain language score that allow us to measure the comprehensibility of speaker turns. This allows us to track conversation dynamics across public service encounters. Moreover, the results indicate that in the dataset under investigation, there are only minor differences in language use between public service professionals and their clients.

1 Introduction

With public administration professionals being responsible for the practical implementation of existing laws, face-to-face interactions between them and citizens constitute an essential point of contact between the public and the state (Maynard-Moody and Musheno, 2012; Lipsky, 1980). A key point for the success of a fair, non-discriminatory and responsive public service delivery is the comprehensibility of the conversation, regardless of the service domain in which it takes place (Bartels, 2013; Eckhard et al., 2022; Lipsky, 1980; Maynard-Moody and Musheno, 2012). Hardly anyone approaching a state authority, for instance when applying for social or unemployment benefits, disposes of all the necessary legal expertise and procedural knowledge, which is often based on uncommon terminology. This creates a knowledge imbalance between the majority of the population and the administrative staff, making comprehensible forms of communication an absolute necessity.

In addition, countries across the globe are putting in place regulation to reduce this inequality in communication. For instance, in Germany, the Act on the Equality of People with Disabilities ('Gesetz zur Gleichstellung von Menschen mit Behinderungen'), as well as the Barrier-Free Information Technology Ordinance ('Barrierefreie-Informationstechnik-Verordnung') (Bundesamt für Justiz, 2011) regulate since 2002 the accessibility in written texts between people with disabilities and the federal government in Germany aiming at removing communication barriers (Schubert, 2016), both in written and spoken communication.

In this work we look at the concept of *plain language*, a major criterion for bridging the communicative gap between citizens and state authorities. First, we propose a taxonomy for measuring plain language in transcribed face-to-face interaction, going beyond previous work which has mainly considered written communication (Lieske and Siegel, 2014; Löffler, 2015; Bredel and Maaß, 2016; Bock, 2018; Pottmann, 2019; Hansen-Schirra and Maaß, 2020; Harbusch and Steinmetz, 2022). This taxonomy takes into account a variety of factors, from morphology, syntax to discourse structure.

We then use computational means to analyze a novel dataset of German real-world public service encounters, i.e., meetings between citizens and state officials which are targeted at providing support and advise with respect to welfare benefits. Using this dataset, we are able to provide a solid foundation for establishing a metric to measure the use of plain language in such encounters. To the best of our knowledge, this is the first empirically-driven and computational attempt to systematically measure transcribed, spontaneous administrative language and it is also the first attempt to model plain language in this genre.

2 Related Work

The use of plain language in written administrative communication has already been addressed in several studies (see for example Hansen-Schirra and

Maaß (2020); Bock (2018). Baumert (2016) gives an overview of the theoretical background of plain language in German, while Baumert (2018) and Baumert (2019) explore its practical application. Bredel and Maaß (2016) and Maaß (2015) focus more on the provision of rule books and recommendations for text production using plain language. In their research, Oomen-Welke (2015) and Pottmann (2019) have shown that the use of plain language is a useful tool for teaching German as a foreign language, as it allows texts to be adapted to the specific language level of the students. Recently, Harbusch and Steinmetz (2022) developed *ExtendedEasyTalk*, a "natural-language paraphrase generator" helping the user to produce correct and understandable texts. In addition, Seiffe et al. (2022) have recently created a dataset of German sentences that were annotated for perceived sentence complexity and comprehensibility by experts and non-experts. Despite the widespread interest in plain language from various disciplines, spoken interaction and in particular of public-service encounters on the street-level still remains widely under-researched (Döring, 2021).

Building on the findings for written communication, the goal in this paper is to develop a method to measure the use of plain language in face-to-face public service encounters, paving the way for a systematic study of comprehensibility in street-level bureaucracy.

3 Plain Language

In the context of administrative language, we see plain language as one way to lower communication barriers that can prevent a client from following the conversation due to the way that the information and its complexity is presented. By lowering comprehension barriers in administrative language, public service employees can help clients understand the issues that affect them while also ensuring equal citizen treatment and thus building trust and social coherence.

The guidelines developed by the German easy-to-understand language association *Netzwerk Leichte Sprache* (which was founded in 2006) on the practical application of easy language were initially derived from practice and are primarily aimed at facilitating comprehension for a heterogeneous group of people, like for example people with cognitive disabilities, reading difficulties, non-native speakers of German and older people (Bredel and Maaß, 2016). However, in situations involving complex issues and requiring specific knowledge, plain language can also be appropriate for communicating with a broader audience and help them to better understand the content at hand.

Despite the fact that the terms 'easy language' and 'plain language' do not describe the exact same concept,they are often used synonymously in practice. The difference between the two lies in the regulation. While there are fixed rules for Leichte Sprache (Maaß, 2015), there are only recommendations for Einfache Sprache. Although we are aware of the difference, we will use both terms interchangeably throughout the paper.

The rules and recommendations established so far mainly refer to the implementation of comprehensibility in written communication. A systematic and empirically-driven investigation and establishment of criteria for spoken communication in Leichte or Einfache Sprache does not yet exist.

4 Operationalising plain language

Derived from the existing guidelines on plain language that apply to written communication, we establish criteria for evaluating Plain Spoken Language. We distinguish the relevant features into four overarching categories: word-level features, morphological features, syntax-level features, and lastly, utterance-level features, given that we are dealing with spoken language.

4.1 Word-Level Features

Jargon/Specialized Vocabulary Although jargon helps to summarise complex concepts into a single term, it is usually only understood by those with shared background knowledge in a specific area, thus hindering everyone else from understanding. The use of specialized vocabulary should therefore be avoided in expert-lay communication or at least explained adequately in order to create semantic transparency and enhance comprehensibility.

By using Term frequency - inverse document frequency (TF-IDF) we take into account the importance of each non-stopword across transcripts. In contrast to the regular usage of TF-IDF as a way to identify highly characteristic items across documents, we use it to identify those lexical items which are typical across the overall dataset. We then count the number of occurrences of each of these specific words per speaker turn and sum up at the utterance level to then divide the aggregated jar-

gon score by the total number of sentences within the speaker turn.

We are aware that this is a simplification as we do not consider the frequency of those items in non-governmental data. However, collecting data that is similar enough to compare the numbers is difficult, as the jargon used in this type of communication is very specific.

Dialect Another word level criterion that contributes to easier understanding of spoken language is the omission of dialect. Particularly in face-to-face conversations with non-native German speakers, the use of dialect can impede comprehension. We use a self-curated word list and our own set of heuristics to identify instances of dialectal speech, focusing mainly on morphological deviations from Standard German, like for example the use *'zamme'* instead of *'zusammen' (together)*. We then obtain the relative occurrence of dialect items by dividing them by the total sentence number within the specific speaker turn.

4.2 Morphological Features

Nominalizations A nominalization is the transformation from one POS, often verbs and adjectives, into a noun, as illustrated in Example (1).

(1) kündigen (V) → Kündigung (N)
resign → resignation

Following the guidelines for plain language (Netzwerk Leichte Sprache, 2014), nominalizations are rather difficult to understand and should therefore be replaced by the verbal style. With the goal of examining whether the verbal style is much more dominant in spoken interaction than the nominal style, we identify nominalization patterns by annotating the nouns ending *-ung*, *-nis*, *-heit*, *-keit*, *-igkeit*, *-anz* or *-enz* throughout the transcripts. After detection, we sum up the total number and divide by the number of spoken sentences to get the relative frequency of nominalizations within one speaker turn.

Participial constructions Participial constructions, like the one in Example (2), are mainly used in formal language register and should therefore be avoided in plain language. Instead of a finite verb, participial constructions use a participle and form a compound sentence together with the main clause. Based on the morphological analysis we identify those participles and calculate their relative occurrence in the individual speaker turn.

(2) Employee: *Die von Ihnen während der Probezeit erbrachte Leistung war wohl nicht ausreichend.*
'The performance you provided during the probationary period was probably insufficient.'

4.3 Syntactic Features

Sentence Length According to the guidelines for plain language (Netzwerk Leichte Sprache, 2014), the choice of sentence length is a very important characteristic to enhance comprehensibility. Ideally, sentences complying with these requirements should be as short as possible to make any information in the sentence easier to process and understand.

Hence, in addition to the length of each spoken sentence, we also calculate the average sentence length per speaker turn.

Sentence complexity As plain language represents a highly simplified variant of any language, the aim is to avoid complex syntactic structures in the interest of comprehensibility. Sentences ideally contain only one proposition and no subordinate clauses, relative clauses or nested constituents which impede comprehension.

Example (3) is taken from one of the transcripts recorded for this project and shows a sentence, which does not comply with the plain language rules, as it consists of more than one constituent and propositions. Additionally, this sentence also violates the recommendation of using short sentences.

(3) Employee: *Dann schicken Sie mal die Krankmeldung, dann gucken wir das an, ob das wirklich die Tage abdeckt.*
'Then send us the sick note and we'll take a look at it to see if it really covers the days.'

In order to calculate sentence complexity, we divide all utterances into smaller units of text in order to work with a more fine-grained structure of the discourse. Although there is no consensus in the literature on what exactly these units have to comprise, it is generally assumed that each discourse unit describes a single event (Polanyi et al., 2004). Following Marcu (2000), we term these units *elementary discourse units* (EDUs). For German, we approximate the assumption made by Polanyi et

al. (2004) by inserting a boundary at every punctuation mark and every clausal connector (conjunctions, complementizers). The average sentence complexity is calculated by dividing the count of discourse units per sentence by the overall number of sentences per speaker turn. A value of 1 therefore corresponds to the ideal score, whereas higher values are considered to be more difficult to understand.

Passive voice and Genitive Both passive constructions and genitives are used frequently, especially in German written communication, despite the fact that there are alternative and easier to understand ways of phrasing. Compared to the active voice, sentences in passive voice are more difficult to understand as they conceal the acting parties (agents) and require proper "decoding" to be fully understood. They should therefore be replaced by the active voice alternation.

The same applies to genitive constructions, which, according to the guidelines for plain language, should be replaced by more easy-to-understand dative phrases. Example (4) shows how the employee uses a genitive construction that could have been replaced by an easier-to-understand sentence construction.

(4) *Das ist wahrscheinlich von der GEZ. Die Mitteilung über den Ablauf der Befreiuung.*
'This is probably from the GEZ (TV licence fee agency). The notification about the process of exemption.'

We count all the occurrences of passive and genitive constructions per speaker turn and calculate the relative occurrences for each feature separately.

4.4 Utterance-level Feature

Speech rate In order to calculate the words spoken per minute, we first calculate speech time based on the timestamps provided from the transcript per speaker turn. In a second step we divide the number of words by the resulting speech time and add this as a feature to the utterance level. We define a speech rate over 120 words per minute to be more difficult to understand, whereas speech rates below 120 are considered to be easier to follow.

5 A Plain Language Score

5.1 Aggregation

To make plain language measurable in spoken interaction, we propose the 'Plain Language Score', with values ranging between 0 (plain language) and (potentially infinitely) high values for difficult language. Each turn in the transcript is assigned one value of the Plain Language Score, based on the feature structure found in that turn. Across our data set, the Plain Language score varies between 0 and 6 for both clients and employees.

For aggregation, we calculate the relative frequency of each feature based on the total number of sentences in that turn. As all features described above are to be avoided when using plain language, we define a relative score of 0.5 and higher as the threshold for plain language, i.e., if a feature for difficult language is found in more than half of the sentences, the overall Plain Language score is incremented by one. In addition, we add 1 for each turn where the speech rate exceeds 120 words per minute and in cases where the average sentence complexity is greater than or equal to 2. Those thresholds are based on expert heuristics. A Plain Language score of 0 for a turn therefore indicates perfect implementation of the plain language recommendations, whereas a higher score assumes the language to be more difficult according to our operationalisation of plain language.

5.2 Data

The collection of speech data in public service encounters started in 2020 across a number of German local administrations. Here, we recorded dialogues between frontline civil servants and their clients in a local jobcentre in a county administration in Western Germany. After an extensive procedure of making sure that we adhere to data protection regulations, the data was then transcribed and anonymized, including multiple cross-checks. In particular, this involved a manual pre-processing step of replacing all names, city names, telephone numbers, e-mail addresses and other sensitive data with generic tags. In addition, personal information of all speakers is untraceable. The corpus underlying the investigation in this paper consists of 52 word-to-word transcripts of real face-to-face public service encounters, covering around 21 hours of verbatim transcript and containing more than 219,000 words. In total our corpus comprises over 10,000 speaker turns, which are almost equally dis-

tributed between employees and clients. Each transcript contains time stamps per utterance and may include elliptical constructions, morphological imperfections, along with dialectal speech.

5.3 Processing

In a first step, all transcripts are converted into XML format.[1] As these files are verbatim transcripts, any non-verbal and background sound or filler content that is transcribed is cleaned up in a pre-processing step. Any forms of interruption are left as is, they will be used in downstream tasks.

In a next step we use the *Stanza* NLP package (Qi et al., 2020) to conduct sentence splitting, tokenization and lemmatization. We also add POS-tags, morphological features and dependency relations of each token (lexeme). The data is then processed with *LiAnS* (Linguistic Annotation Service) which is based on (Gold et al., 2015), a rule-based NLP pipeline for analyzing linguistic features in spoken dialogue and debates in English and German. For this paper, we extend the pipeline in order to automatically detect the features described in Section 4. This also involves a set of carefully crafted disambiguation rules in order to provide reliable annotation. The features identified in each turn across each transcript provide the basis for calculating the Plain Language Score per turn.

6 Analysing Plain Language in Public Service Encounters

6.1 Distribution

The histogram in Figure 1 shows the frequency distribution of the plain language scores in all transcripts both for clients and employees. For both groups, the majority of speaker turns lie at a score between 0 and 2. Scores between 3 and 4 occur sporadically, while scores at 5 and 6 appear only rarely. The mean score for employees is 1.27 and for clients 0.96, with the median for both groups being 1.0.

Examples (5), (6) and (7) illustrate the difference between score 0, score 1 and score 6 speaker turns, respectively:

(5) Employee: *Mehr habe ich jetzt im Moment nicht. War mir wichtig Ihnen das nochmal zu erklären.*
'That's all I have right now. It was important for me to explain this to you again.'
`SCORE: 0`

(6) Client: *Der kriegt auch eine Rente, aber das nennt sich irgendwie anders.*
'He also gets a pension, but it's somehow called differently.'
`SCORE: 1`

(7) Employee: *Vor allem, weil der ganze Förderzeitraum noch nicht ausgeschöpft wurde, weil diese Förderung nach dem Paragraph 16e wären Sie zwei Jahre lang gefördert.*
'Especially because the whole funding period has not been exhausted, because this funding under the paragraph 16e you would be funded for two years.'
`SCORE: 6`

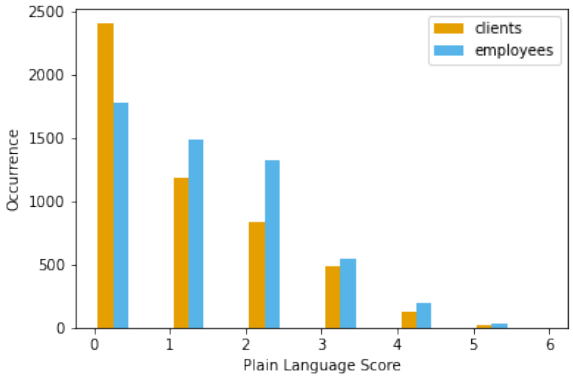

Figure 1: Histogram of Plain Language Scores by employees and clients.

The speech turn in Example (5) receives a score of 0, because no difficult-to-understand constructions and phrases were used. The sentences are short, without subordinate clauses, passivizations and nominalizations. Example (6) is scored with 1 because of the average sentence complexity being 2 (two subclauses in one sentence). The speech turn in Example (7), in contrast, receives the highest score of 6 in the data set, composed as follows: The speech rate calculated from the words spoken and time used is 134 words per minute. The average sentence complexity is over 1. In addition, the sentence contains a passive, a participial construction, a nominalization and use of jargon. Each of these items increases the score by 1.

[1] The full set of anonymized transcripts can be requested via verwaltungssprache@uni-konstanz.de

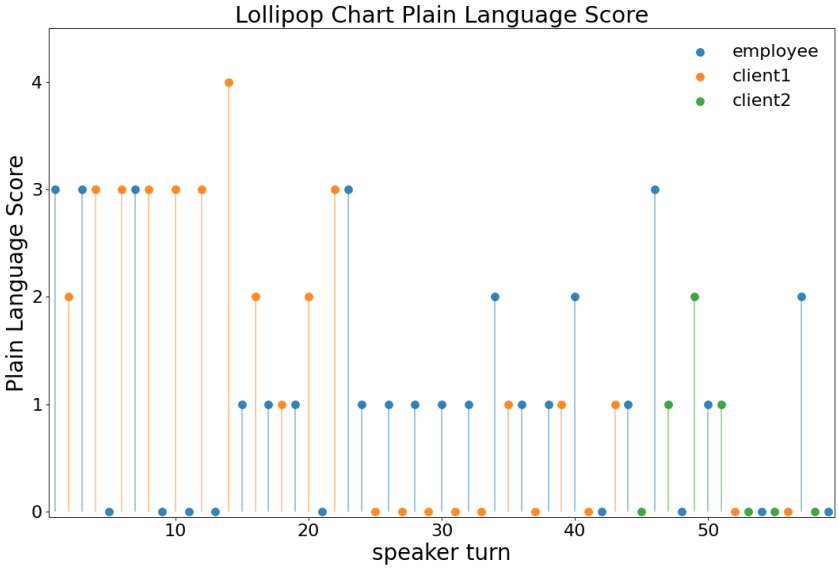

Figure 2: Visualization of conversation dynamics based on the plain language scores of employees and clients within a consultation.

6.2 Conversation dynamics

The lollipop chart in Figure 2 shows the conversation dynamics based on the plain language score per speaker and speaker turn. One employee of the employment agency and two clients participated in the interview depicted here. Based on the detected features and the calculated plain language scores, the visualization shows the scores ranging between 0 and 4. In this particular interaction, both the employee and the clients use equally easy and difficult to understand language, though the level varies across the conversation, with a number of outliers (e.g., at turns 14, 46 and 57).

The outlier in turn 14 is shown in Example (8): it receives a score of 4 due to a high speechrate (280 words/minute), high average sentence complexity, and the use of passive and participial constructions.

(8) Client1: *Der ist halt jetzt nur nicht in dem Haus, weil er dann wieder abgezogen wurde nach Ortsname, und sobald der wieder zurück ist, würde ich den dann unterschreiben.*
'It's just that he's not in the house now, because he was then taken away again by place name, and as soon as he's back again, I would then sign it.'
SCORE: 4

6.3 Feature correlation

In order to see more closely which features of plain language correlate, we plot heatmaps of all features for employees and clients separately in Figure 3. Overall, the correlation between features is very similar for employees and clients: For instance, we find that some features, such as "wordcounter" and "participle_constructions", "passive" or "nominalizations" more strongly correlate than other features, both for clients and employees. The same applies for the feature pair of "passive" and "participle_constructions". "Dialect" and "genitive" also seem to correlate for employees and clients.

The observed correlation between "jargon" and the remaining linguistic features for employees can be explained by the fact that their knowledge of technical terms is better compared to that of clients and leads them to use these terms more often during consultations. At the same time, we observe a higher, albeit defiantly low, correlation between "dialect" and the other features for clients. We attribute this difference to the fact that employees are more careful about speaking Standard German to their clients than the clients themselves are.

(9) Employee: Die Verfahrenskosten *des Bescheides* liegen so zwischen 150 und 180 Euro, dass Sie Bescheid wissen.
'The procedural costs of the notice range from 150 to 180 euros so that you know.'

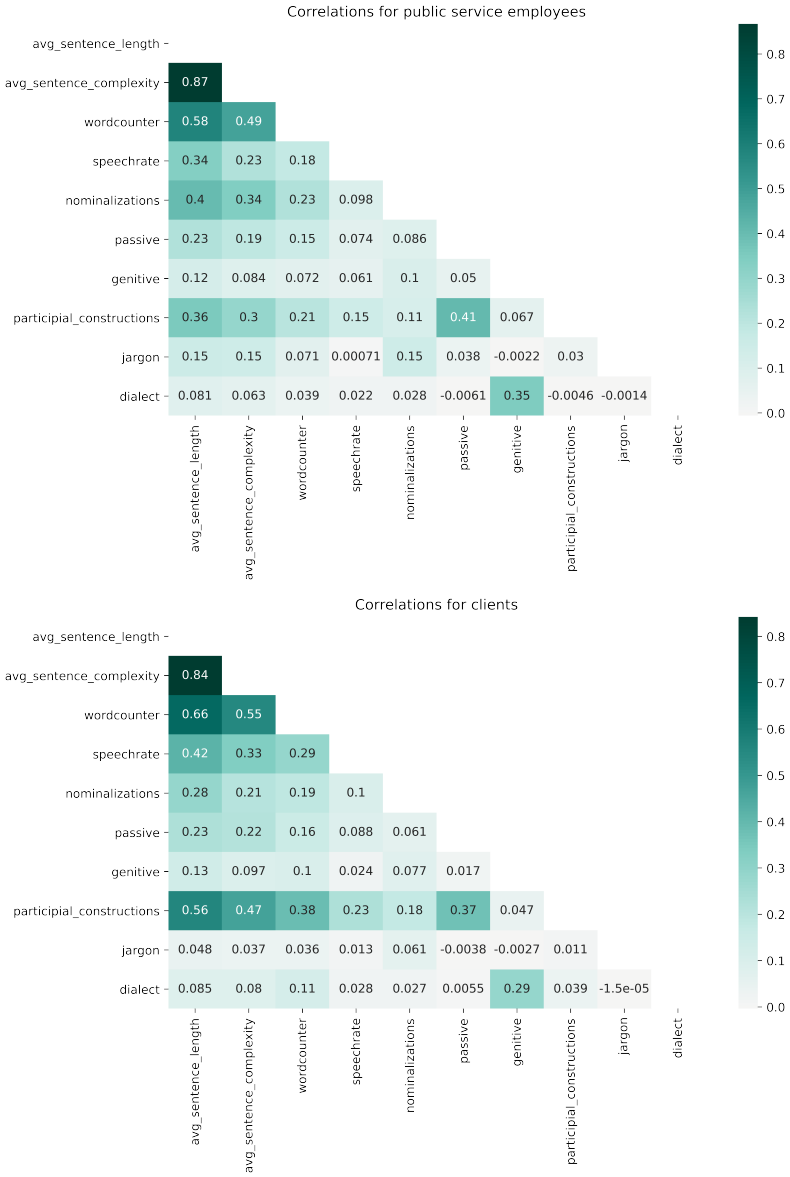

Figure 3: Heatmap of feature correlation in employee (up) and client (down) turns.

Furthermore, we attribute the quite substantial correlation between dialect and genitive to noise in our dataset. An example is given in (9). The definite article *"das"* must be adapted to *"des"* when used with the genitive in German. At the same time, in some dialects, *"des"* is an equivalent of *"das"* without signaling a genitive noun phrase. Teasing apart these two usages will be done in a later step.

In terms of differences between roles, i.e., between employees and clients, we see a higher correlations for "jargon" throughout all relevant features for employees, except for "speechrate". At the same time, "dialect" correlates stronger with other features in client speaker turns in Figure 3.

Overall, the results show only minor differences in language use between employees and clients in public service encounters.

7 Discussion and Outlook

With our analysis of 52 transcripts of face-to-face public service encounters, we conduct the first empirically driven and systematic computational study of plain language in spontaneous administrative language. We offer a taxonomy of different features of plain language, motivated by plain language usage in written communication, extending it with features across morphology, syntax and discourse structure. We then show offer different insights into the dataset regarding plain language: an aggregation of the scores for employees versus clients, the dynamic of plain language within individual

conversations and feature correlations across the taxonomy.

The results presented in Section 6 allow us to draw two conclusions: First, there are only minor differences in correlations for the relevant plain language features between public service employees and their clients, indicating that employees already adapt to their clients language. This is contra our expectations that public employees have on average a higher language complexity than their clients, however by looking into the dynamics of individual encounters we see that the employees adjust their language to match it with the complexity of their clients. Secondly, for longer utterances, we observe an increase in correlation of passive voice, genitive, nominalizations and participle constructions across both roles. In other words, longer utterances with a higher wordcount are more likely to contain complex language than short ones.

A shortcoming of the current approach is that we loose out on phonetic characteristics such as speaking volume or intonation for analysing plain language. However, given the tight data protection regulation involved in recording public service encounters, we are prohibited from using the audio data for any research purposes.

Instead, we focus on a number of other avenues for future work: First, we plan to include more coherence-related measures: inter- and intrasentential discourse relations, topic coherence by way of lexical semantic relatedness between words and entity coherence involving some form of anaphora resolution. Secondly, given that there is survey data available for each encounter, in particular with respect to the socio-demographics of both employee and client, previous experiences with state authorities as well as client satisfaction with the encounter, we will use the features for plain language established in this paper to identify patterns that aid in making public service encounters more satisfactory. We also work on an extension of the data set, both with more transcripts across different regions in Germany, with the particular aim of better accounting for dialectal variance, but also including data on public service encounters in English. This allows us to test whether our measures are in fact cross-linguistically applicable. Lastly, we aim for a reference corpus for plain language which contains external judgements of the complexity of the language, with balanced data of different levels of language complexity. This will allow us to validate the plain language score presented in this paper more objectively.

Overall, we hope to contribute to making public service delivery more accessible and reduce inequality in communication, not only in written but also in spoken communication.

Acknowledgement

The work reported on in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379.

References

Koen P.R. Bartels. 2013. Public encounters: The history and future of face-to-face contact between public professionals and citizens. *Public Administration*, 91:469–483.

Andreas Baumert. 2016. *Leichte Sprache – Einfache Sprache. Literaturrecherche, Interpretation, Entwicklung*. Bibliothek der Hochschule Hannover.

Andreas Baumert. 2018. *Einfache Sprache - Verständliche Texte schreiben*. Spaßam Lesen Verlag.

Andreas Baumert. 2019. *Mit einfacher Sprache Wissenschaft kommunizieren*. Springer Spektrum.

Bettina M. Bock. 2018. *'Leichte Sprache'-Kein Regelwerk: Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt*. Universität Leipzig, LeiSA.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Bibliographisches Institut GmbH.

Bundesamt für Justiz. 2011. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung-BITV 2.0). *www.gesetze-im-internet.de/bitv*, 2(0).

Matthias Döring. 2021. How-to bureaucracy: A concept of citizens' administrative literacy. *Administration & Society*, 53(8):1155–1177.

Steffen Eckhard, Laurin Friedrich, Annette Hautli-Janisz, Vanessa Mueden, and Ingrid Espinoza. 2022. A taxonomy of administrative language in public service encounters. *International Public Management Journal*, pages 1–16.

Valentin Gold, Mennatallah El-Assady, Tina Bögel, Miriam Butt, and Katharina Holzinger. 2015. Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality. *Digital Scholarship in the Humanities*, ISSN 2055-7671. - eISSN 2055-768X.

Silvia Hansen-Schirra and Christiane Maaß. 2020. *Easy Language Research: Text and User Perspectives*. Frank & Timme.

Karin Harbusch and Ina Steinmetz. 2022. A Computer-Assisted Writing Tool for an Extended Variety of Leichte Sprache (Easy-to-Read German). *Frontiers in Communication*, 6.

Christian Lieske and Melanie Siegel. 2014. Verstehen leicht gemacht. *Technische Kommunikation*, 1.

Michael Lipsky. 1980. *Street Level Bureaucracy: Dilemmas of the Individual in Public Services*. Russell Sage Foundation.

Cordula Löffler. 2015. Leichte Sprache als Chance zur gesellschaftlichen Teilhabe funktionaler Analphabeten. *Didaktik Deutsch*, 38(2015):17–23.

Christiane Maaß. 2015. *Leichte Sprache: Das Regelbuch*. Lit Verlag, Münster.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, Mass.

Steven Maynard-Moody and Michael Musheno. 2012. Social equities and inequities in practice: Street-level workers as agents and pragmatists. *Public Administration Review*, 72:S16–S23.

Netzwerk Leichte Sprache. 2014. *Leichte Sprache. Ein Ratgeber*. Netzwerk Leichte Sprache, Frankfurt am Main.

Ingelore Oomen-Welke. 2015. Leichte Sprache, Einfache Sprache und Deutsch als Zweitsprache. *Didaktik Deutsch*, 20(38):24–32.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. Sentential structure and discourse parsing. In *Proc. of the ACL'04 Workshop on Discourse Annotation*, pages 80–87.

Daniel M. Pottmann. 2019. Leichte Sprache and Einfache Sprache - German Plain Language and Teaching DaF German as a Foreign Language. *Studia Linguistica*, 38:81–94.

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

Klaus Schubert. 2016. Barriereabbau durch optimierte Kommunikationsmittel: Versuch einer Systematisierung. In Mälzer Nathalie, editor, *Barrierefreie Kommunikation – Perspektiven aus Theorie und Praxis (German Edition)*, pages 15–33. Frank & Timme.

Laura Seiffe, Fares Kallel, Sebastian Müller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In *Proceedings of the Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Networks of Power: Gender Analysis in Selected European Parliaments

Jure Skubic
Institue of Contemporary History
Ljubljana, Slovenia
jure.skubic@inz.si

Jan Angermeier
University of Stuttgart
Stuttgart, Germany
jan.angermeier@gmail.com

Alexandra Bruncrona
University of Helsinki
Helsinki, Finland
alexandra.bruncrona@gmail.com

Bojan Evkoski
Jozef Stefan Institute
Jozef Stefan Postgraduate School
Ljubljana, Slovenia
bojan.evkoski@ijs.si

Larissa Leiminger
larrisa.leiminger@gmail.com

Abstract

Parliamentary debates maintain an integral part of a country's political power. As social patterns within parliamentary procedures might be reflected in shaping national legislation, identifying these patterns becomes crucial. Therefore, this paper analyzes argumentative and structural power over one term in three national parliaments (UK 2017–19, Spain 2016–19, and Slovenia 2014–18), focusing on gender distribution within these power schemes. In an attempt to recognise gender-related patterns presented in literature, the paper investigates and discusses selected parliamentary debate topics in-depth. Problematic patterns, such as female MPs partaking less actively than male MPs in most debates, could be identified by successfully combining computational methods and social sciences in this multidisciplinary work.

Keywords: ParlaMint, parliamentary debates, social network analysis, digital humanities

1 Introduction

Parliamentary debates are a significant source of highly relevant data, not only for social sciences and humanities but also for computer science. As an institution, parliament is responsible for shaping legislation that impacts people's everyday lives and is a source of power for members of parliament (MPs) and other politicians (Bischof and Ilie, 2018). Therefore, parliamentary debates are at the heart of political decision-making and manifest political power – a complex phenomenon widely theorised in the study of culture and society (Simon, 1953; Parsons, 1963). This makes parliamentary discourse inherently interesting for both qualitative (Van Dijk, 2000; Bayley, 2004; Ilie, 2015) and quantitative research (Abercrombie and Batista-Navarro, 2020; Rheault et al., 2016; Cherepnalkoski and Mozetič, 2016). Parliamentary discourse research is often multidisciplinary since it touches on various academic fields, including history, psychology, linguistics, political science, and computer science. Especially the latter has, in recent years, started to collaborate with humanities and social sciences successfully since various research (Andrushchenko et al., 2022; Blaxill, 2013) have shown that the interconnection of computational methods and those of humanities and social sciences can offer beneficial and relevant results. Today, the publication of the biggest and most richly annotated parliamentary dataset ever, the ParlaMint corpora (Erjavec et al., 2022), is a new call for extensive multidisciplinary research that could make the next step in understanding parliamentary discourse.

This paper was written as the aftereffect of the 2022 Helsinki Digital Humanities Hackathon[1]. During the 10-day intensive work, we exercised multidisciplinary research on the ParlaMint dataset in a project titled Networks of Power. We focused on analyzing power distribution inside parliamentary networks in three European countries; Slovenia, Spain, and the United Kingdom. Our work resulted in various vital findings, presented at the hackathon, which we reflect on in this paper.

The objective was twofold: First, we aimed to

[1] https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/helsinki-digital-humanities-hackathon

shed light on the manifestations of different aspects of power in parliamentary discourse and political debates by analyzing the networks that emerge from parliamentarians mentioning each other. Our second, equally important ambition was to show how intertwining computer science with humanities and social science methods and research approaches can generate powerful and meaningful results, which could have been overlooked if the analysis was done from the perspective of solely one of the disciplines.

The paper is structured as follows: First, we present our research questions, specifying key terminology and definitions. In the following section, we introduce the datasets utilized in our study and our corpora extraction using keyword searches. In the subsequent methodology section, we review the metrics and calculations used in our computational analysis. The paper is completed with an in-depth descriptive section on our results, including a case study on a selected topic (immigration) and a critical discussion of our methods. Finally, in the conclusion, we raise the most important outcomes of our work.

2 Research questions

We determined the power relations in parliamentary discourse to be crucial for decision-making, which is why we decided to focus specifically on different types of political power. We worked with a ground definition of power inspired by Foucauldian power-knowledge (Schwan and Shapiro, 2011) and, in order to reduce the complexity of the theme of power, we approached the networks through two research questions: First, we wanted to analyze the *argumentative power* of the MPs and focused our research on how speeches given by the MPs and the mentions of the MPs can give insights into the power of MPs within political debates. Our main aim here was to show how argumentative power can be seen through the parliamentary discourse itself and how the mentions of the MPs shape the power relations inside the parliament.

The second research question focused on *structural power* within a parliament. We were interested in how the speech practices of female and male MPs relate to topic and power distribution in parliamentary discourse. We approached this question by referring to a research study conducted by Bäck, Debus and Müller (2014) who divide parliamentary discourse into *"hard" topics* (e.g., energy, finances) and *"soft" topics* (e.g., healthcare, education, immigration). Bäck et al. (2014) show that "hard" topics were more dominated by male MPs whereas "soft" topics were more dominated by female MPs in the Swedish parliament 2002–10. We therefore aimed to inspect whether or not this pattern could be observed in the selected parliamentary data.

3 Data

In order to get insight on the distribution of power of parliamentarians in different parliaments across Europe, we exploited the richest corpora on parliamentary data available; the ParlaMint dataset (Erjavec et al., 2022). It contains session transcriptions from 17 (seventeen) European national parliaments, adding up to over half a billion words. The corpora are uniformly encoded, contain rich metadata of 11 thousand speakers, and are linguistically annotated using the Universal Dependencies standard. In addition, and perhaps most importantly for our work, named entities are annotated. The corpora span across multiple parliamentary terms, depending on the country, with transcripts of speeches made between 2009 to 2020.

We selected three parliaments to analyze: the United Kingdom's *House of Commons* (lower house), Spain's *Congreso de los Diputados* (lower house) and Slovenia's *Državni zbor* (lower house). Moreover, in order to have one full balanced set of parliamentarians, we selected a subset of the transcripts which was part of the latest finished parliament term for each country. Table 1 shows the general statistics for each of the three subsets, after parliament guests and chairpersons were removed from the corpora.

With the goal of exploring the structural power defined earlier in this paper, we applied speech selection on the following five topics: energy, finances, healthcare, education, and immigration. The topic modeling was done using a manual procedure using keywords. We identified keyword sets for each topic using careful qualitative and iterative process with the help of the NoSketch Engine system (Rychlỳ, 2007; Kilgarriff et al., 2014) which allows manual browsing through large datasets. Then, a speech belongs to a topic if it contains one or more keywords of the keywords pool for that particular topic. The reason why we decided to apply topic identification using keywords instead of automatic topic modeling such as LDA (Jelodar

Country	From	To	MPs	Female MPs	Speeches	FAR	Mentions	FPR
UK	22/06/2017	17/12/2019	645	32.4%	168k	34.6%	28k	35.1%
Spain	19/07/2016	28/02/2019	388	41.2%	11k	36.8%	6.4k	13.6%
Slovenia	01/08/2014	11/10/2018	101	36.6%	21.9k	27.8%	11.8k	18.6%

Table 1: Basic information on the ParlaMint subsets covered in this work. It includes the country, the first and last date of the speeches; number of MPs and the share of female MPs; number of speeches and the share of speeches given by female MPs (here: Female MP Active Relevance, FAR); number of times MPs are mentioned in speeches and the share of female MPs mentions (here: Female MP Passive Relevance, FPR).

et al., 2019) is because, having such a large dataset, we aimed for a controlled selection of speeches with high topic precision, instead of LDA's low precision with high inclusion. Table 2 shows the general statistics on the topical subcorpora.

4 Methodology

We were interested in measuring the argumentative power MPs had in the three parliaments and observed how this power scheme depended on gender, i.e. structural power, using ParlaMint's traditional gender distinction of female and male. However, measuring power is impossible using but one metric, and one usually adapts multiple measurable aspects that contribute to the general power.

One of the major intellectual challenges while conducting our research was to find a way of analyzing the power of the MPs inside parliamentary debates. We determined that one of the possible approaches of explaining political power is by examining the political influence that MPs hold inside the parliament. This was based on the work of Ilie (2010) who shows that parliamentary interaction exhibits a permanent competition for power and leadership roles which according to van Dijk (2018) is one of the key characteristics of political identities. Ilie (2010) shows that in parliamentary confrontation MPs are taking different roles - from the role of the listener to the role of the speaker. It is this confrontational dialogue that fuels the sense of competitiveness in the parliament and is paralleled by ongoing attempts to destabilize and reestablish the power balance. Thus, we identified the political influence not merely as a quantitative fact of how much the MPs in parliament speak, but also as how much resonance their speeches hold and consequently how much they are referenced by other MPs. We defined the two concepts as relevance and determined that power and relevance are proportionally related - the more power an MP holds, the higher their relevance inside parliament and vice versa. The conceptual use of the term relevance (used henceforth) was thus our own and was used to reflect the political power produced by and given to the MPs.

For that purpose, we defined two types of *relevance* of MPs in the parliament: Active Relevance (AR) and Passive Relevance (PR). By AR, we indicated the portion of speeches an MP or a group of MPs gives. This metric only requires calculating a particular metadata subset's unique speeches in ParlaMint. AR showed us to which extent MPs actively participated in debates. On the other hand, PR refers to the number of times an MP is mentioned by other MPs, as a portion of the total MP mentions in the parliament. PR showed us how much MPs are referenced as being relevant for debates without actively participating.

In order to calculate PR, we used the *named entity* tags in ParlaMint. Once we detected a personal named entity, we measured the Levenshtein distance of the alphabetically sorted named entity with the set of parliamentarians for each parliament, using a token set ratio (Gonzalez, 2021). If a single perfect match was found, we marked it as a mention. If there was no perfect match, that meant a person outside of the parliament had been mentioned. Finally, if multiple perfect matches were detected (e.g., multiple identical surnames), we checked if one of the candidates gave a speech closely before or after the currently processed speech (in our case, ten before and after). If this further heuristic did not solve the ambiguity, the mention was not marked. To determine the precision of mentions, we manually investigated 50 speeches from each parliament and checked whether the named entity correctly corresponded to the detected mention. The manual check showed our process had maintained a high precision in the mention detection (UK 86%, Slovenia 100%, Spain 77%).

To better understand and visualize the power relations between parliamentarians, we created comention directed networks. Nodes represented the MPs, whereas edges represented MPs mentioning

each other, with weights being the number of times a mention happened. We created general networks for each parliament and topic specific networks, which we discuss in the following section.

5 Results and discussion

The presentation of our results begins with a general overview of Active Relevance (AR) and Passive Relevance (PR) for each country, with particular attention to gender distribution. Subsequently, we do the same for the five selected topics and examine how AR and PR are distributed between male and female MPs in the topically separated subcorpora. Ultimately, we analyze the topic of immigration and its networks of mentions qualitatively.

The column diagrams in Figure 1 show the top 20 MPs by their number of speeches made and the top 20 by their number of mentions for Slovenia, Spain, and the United Kingdom. In all diagrams, a power-law could be observed to some extent. For example, a common factor in the distribution of speeches and mentions in the Slovenian parliament was a noticeable gap between the top two MPs and the rest. It should be noted that these top two MPs were not the same for speeches and mentions; Franc Trček and Jožef Horvat gave the most speeches, while Miro Cerar and Janez Janša were mentioned most frequently. In the Spanish parliament, the mentions concentrated in Mariano Rajoy, with a noticeable gap between the three runner-ups and the rest. However, the distribution of speeches in the Spanish parliament lacked such a spike at the left of the diagram. Curiously, in the British parliament, it went the other way around: speeches were distributed with a large gap between Theresa May, Andrea Leadsom and the rest, while mentions were distributed more evenly.

In all top 20 lists, female MPs were the minority. The distributions with the most female MPs were the top 20 mentions for the UK (seven) and the top 20 speeches for Spain (five). The lowest number of female MPs (only one) could be found in the distribution of mentions for Slovenia. Generally, the distribution of mentions was to a great extend determined by the identities of those holding important government positions.

The definition of high or low AR and PR was based on the ground truth of the specific gender distribution in each parliament. The axiom was as follows: if 30% of MPs were female MPs, we assumed 30% of speeches to be given by female MPs, and 30% of name mentions referring to female MPs. A deviation from this expected value was then considered high or low relevance. Hence, having high AR did not mean female MPs gave more speeches than men. Instead, it meant that their share of speeches was more significant than one may expect based on the parliament's gender distribution. In other words, "high AR" means female MPs' AR is higher than expected. The reference to the ground truth of gender distribution allows for a more sensible interpretation of the data because it considers the expected outcome.

In a subcorpus of speeches about a "soft" topic (Bäck et al., 2014), we anticipated a more significant share of female-held speeches and name mentions of female MPs than the share of female MPs, i.e. the expected value. If this was the case, we established that this topic in this specific parliament (during the term we observed) appeared to follow the categorisation found by Bäck et al.

Our expected values for AR and PR, i.e. the share of female MPs in the respective parliaments, were 32.4% for the UK, 41.2% for Spain, and 36.6% (see Table 1). Considering these numbers, one can interpret our findings in Table 2. Finance and energy were considered "hard" topics (Bäck et al., 2014). In our study, this was shown to be true for Slovenia and Spain, where the PR for female MPs was almost 19 p.p (percentage points) and 29 p.p lower than the expected values. AR was also lower by 11.5 p.p and 13.5 p.p, and 8.6 p.p and 5.8 p.p, respectively. For the UK, the values were not as clear (-2.3 p.p and +0.1 p.p for AR, +0.9 p.p and +1.6 p.p for PR). Hence, in the British subcorpora, speeches on finance and energy were less dominated by male MPs but even somewhat dominated by female MPs.

Our selected, predefined "soft" topics: education, healthcare, and immigration (Bäck et al., 2014), partly reflected their expected values. In all three countries, the AR for healthcare was the highest for female MPs (41.8%, 47.8%, 38.3%) compared to the other surveyed topics. For the UK, healthcare PR was also the highest PR for female MPs (42.2%). For Slovenia and Spain, healthcare PR was lower than the expected values, by 10.8 p.p and 21 p.p. Only for the UK, education appeared to be confirmed as a "soft" topic with high AR (+7.3 p.p) and PR (+7.7 p.p) compared to the share of female MPs in the parliament. While the deviation

Topic	The United Kingdom			Spain			Slovenia		
	Tot	FAR	FPR	Tot	FAR	FPR	Tot	FAR	FPR
Finance	15k	30.1%	**33.2%**	2.4k	32.9%	12.6%	7.3k	25.1%	17.9%
*		(-2.3)	(+0.9)		(-8.6)	(-28.9)		(-11.5)	(-18.7)
Energy	1.9k	**32.4%**	**34.0%**	<1k	35.7%	13.1%	<1k	23.1%	17.8%
*		(+0.1)	(+1.6)		(-5.8)	(-28.4)		(-13.5)	(-18.8)
Healthcare	11k	**41.8%**	**42.2%**	1.1k	**47.8%**	20.4%	2.1k	**38.3%**	25.8%
*		(+9.4)	(+9.8)		(+6.3%)	(-21.0)		(+1.7)	(-10.8)
Education	9.3k	**39.7%**	**40.1%**	1.9k	37.2%	13.8%	2.8k	34.2%	22.1%
*		(+7.3)	(+7.7)		(-4.3)	(-27.7)		(-2.4)	(-14.5)
Immigration	2.7k	**41.4%**	**37.6%**	<1k	33.0%	10.9%	1.7k	24.3%	16.1%
*Deviation from exp. (p.p)		(+9.0)	(+5.3)		(-8.5)	(-28.4)		(-12.3)	(-20.5)

Table 2: Information on the topic subsets filtered by the keyword selection. The table is split in three columns (countries) and two rows (types of topics: "hard" and "soft"). For each combination, it shows the total number of speeches (Tot.), Female MP Active Relevance (FAR) and Female MP Passive Relevance (FPR). The bolded numbers present higher percentages than the total share of female MPs in that particular parliament, i.e. the expected value. The deviations from the expected values are marked under each row in percentage points.

from the expected values of Slovenian and Spanish female MPs' AR regarding education was only slightly in the negative (-2.4 p.p and -4.3 p.p), their education PR was distinctly lower (-14.5 p.p and -27.7 p.p). However, these PR results were less extreme than those in the finance and energy subcorpora. Immigration speeches in the UK included highly relevant female MPs (female MPs' results were +9 p.p for AR and +5.3 p.p for PR compared to their expected values). In Slovenia and Spain, female MPs' immigration AR were similarly low to their finance and energy values. The immigration subcorpus featured the lowest PR for female MPs in Spain out of all topical subcorpora (16.1%, or -20.5 p.p compared to expected values).

As the results imply, the dichotomy between "hard" and "soft" topics may not be universally valid. Instead, the identity of "hard" and "soft" topics is different from parliament to parliament. Hence, it would be a positivist fallacy to consider the overall distribution of AR and PR between male and female MPs and conclude that the parliaments in Slovenia and Spain focus more on "hard" topics while the one in the UK has a slight leaning toward "soft" topics.

5.1 Case: Immigration

To demonstrate the diversity and specificity of our results, we evaluated our networks for the topic of immigration in all three countries. Predefined as "soft" (Bäck et al., 2014) but proven ambiguous, the immigration topic showed most prominently how the discussion had varied between parliaments. All networks (see Figure 2) are coloured according to gender, with male MPs appearing in magenta and female MPs in green. The node size represents the indegree and, therefore, how often an MP was mentioned in speeches (PR), while the text size represents the number of speeches (AR) the MPs have given on the topic.

In the Slovenian parliament network (see Fig. 2), the most visually present persons were Miro Cerar, Branko Grims, Vinko Gorenak and Ivan Janša. Miro Cerar, the Prime Minister at the time, had the highest PR and was in the top 10 of highest AR. Ivan Janša, the most prominent oppositional figure, had the second highest PR but, in comparison, a rather low AR. Most notably, both had a significantly lower outdegree than their prominent counterparts, Branko Grims and Vinko Gorenak. From these metrics, we concluded that Cerar and Janša had a high degree of argumentative power and must have been in an overall prominent position within the political landscape. They did not need to actively participate in discussions to be passively relevant to a particular topic. In addition, Cerar and Janša did not have to reference other politicians to be referenced in other politicians' speeches. Their counterparts, Grims and Gorenak, both anti-immigration and far-right, demonstrated remarkably high ARs and comparatively high PRs. They brought themselves into the discussion of immigration and often disputed the decisions made by the Prime Minister and his coalition. Overall, we needed to consider that immigration was a hugely relevant topic for the political landscape in Slovenia after the refugee crisis in 2015–16 (Stoyanova and

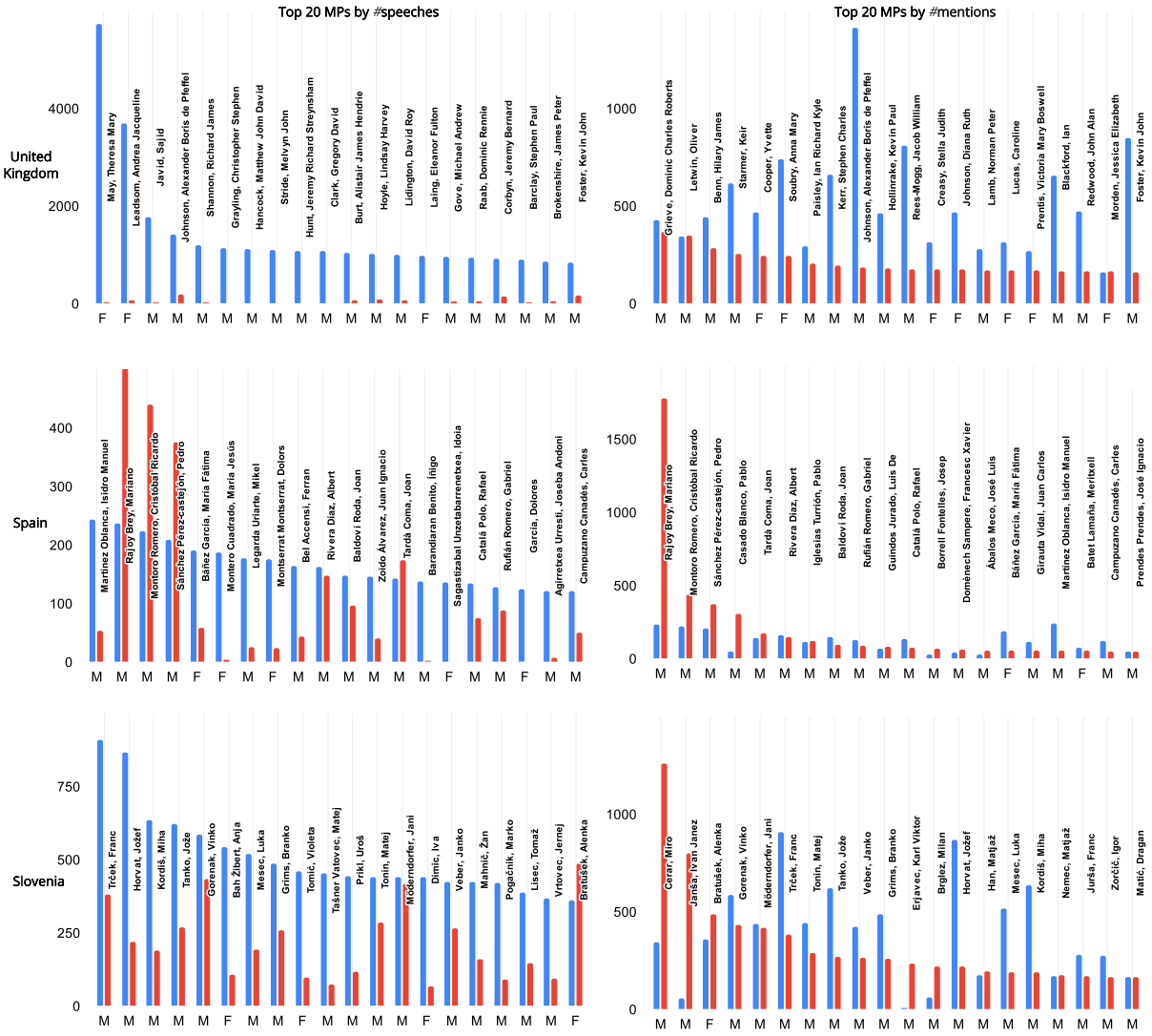

Figure 1: **Speech and mention distributions.** Each row represent the three countries, while the first column shows the top 20 MPs sorted by the number of speeches they have given, and the second shows the top 20 MPs sorted by the times they were mentioned. x-axis simply labels the MPs gender, while the y-axis presents both speech and mention counts. For each MP, the blue bar depict the number of speeches an MP has given, while the respective red bar depict the mention counts.

Karageorgiou, 2019). The network additionally visualises how male MPs dominated the immigration discourse was in Slovenia, with female MPs' AR at 24.3% (12.3 p.p lower than their overall representation, i.e. the expected value) and PR at 16.1% (20.5 p.p lower than the expected value).

While the wave of refugees in 2015–16 affected our Slovenian results, we did not see the same phenomenon in Spain. Under the topic of immigration for Spain, we observed the lowest number of speeches per individual overall, suggesting that immigration was not as much discussed in Spain as in Slovenia and the UK. We can still find prominent politicians such as Mariano Rajoy, the Prime Minister at the time, and Pedro Sanchéz, the opposition leader and Rajoy's successor, at the centre of the Spanish immigration speech network. However, Rajoy and Sanchéz were central in all the Spanish topical speech networks; We needed to attribute this effect to some extent to the overall turmoil of the Spanish government, which cumulated in a vote of no confidence against Rajoy's government in June 2018. Afterwards, we could identify certain MPs central to the topic of immigration. They were the Minister of Interior Juan Zoido, with a high AR and PR, and one of the few dominant Spanish female MPs, Ana Surra, a Catalonian politician chairing an association bringing together foreigners in Catalonia. Overall in the Spanish example, the discussions on immigration were dominated by male speakers. The female speech AR was at 33.0% and, therefore, lower than the expected value (41.2%).

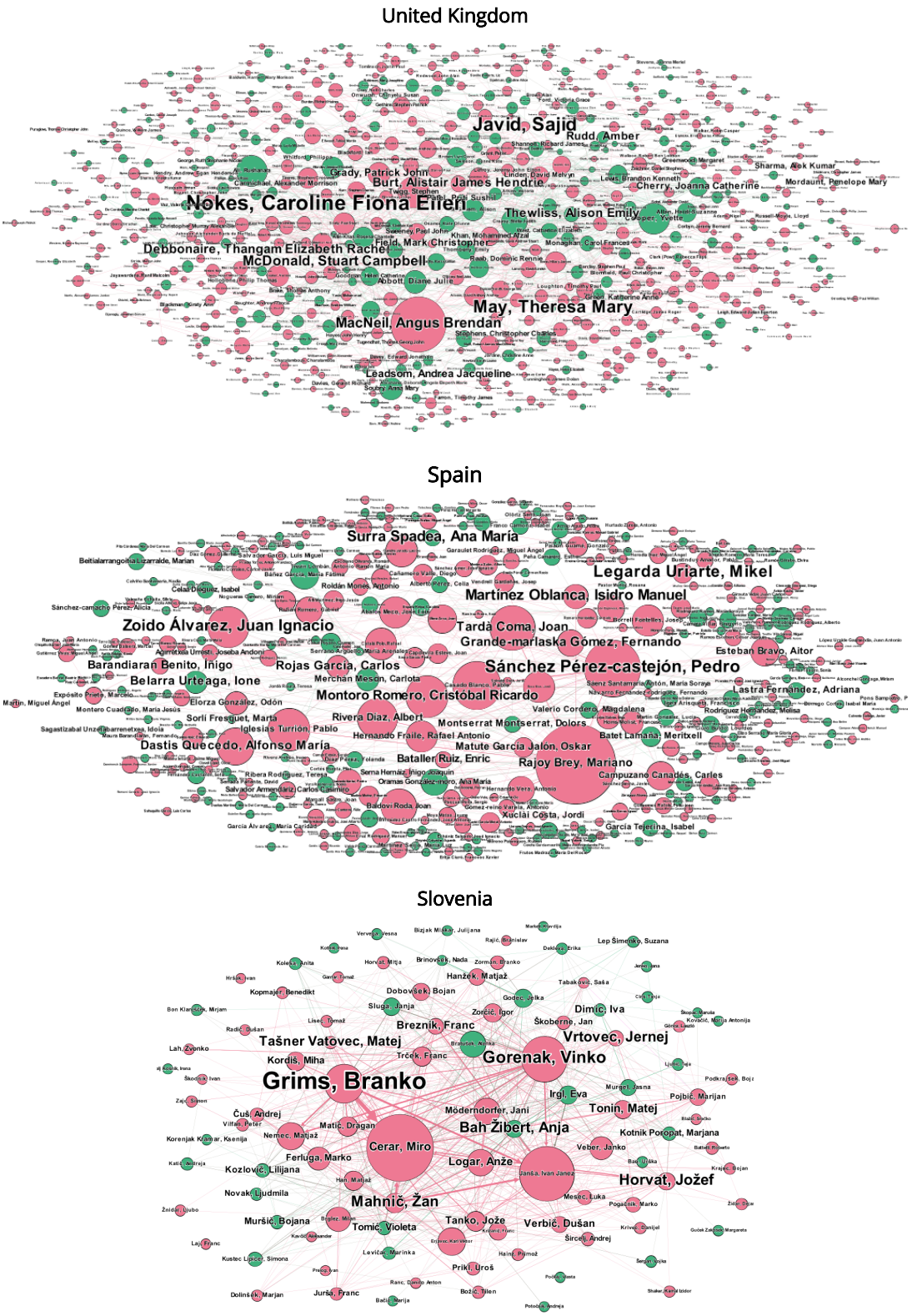

Figure 2: **Immigration co–mention networks.** The UK, Spain and Slovenia co-mention networks visualised using the Gephi (Bastian et al., 2009) tool. Color of nodes represents gender (male is magenta, female is green), name sizes are the number of speeches (Active Relevance) and node sizes are the number of mentions (Passive Relevance). For layout, we use the Fruchterman-Reignold (Fruchterman and Reingold, 1991) algorithm.

Furthermore, the female MPs' PR was significantly lower at only 10.9%.

The immigration network for the United Kingdom parliament appeared visually different from the Slovenian and Spanish examples. We determined two reasons: First, in the UK, there was a higher number of MPs (see Table 1) and, thus, a larger subcorpus of speeches on immigration (see

Table 2) than in the other two countries. Second, the stronger presence of green nodes visually represents the presence of more prominent female MPs. For the UK, we observed that the female MPs' AR was at 41.4% (9.0 p.p higher than their overall representation) and their PR at 37.6% (5.2 p.p higher than their overall representation). We, therefore, concluded that in the UK, the topic of immigration was dominated by female MPs. These prominent MPs were namely Caroline Nokes, Minister of State and Immigration at the time, and Theresa May, the former Prime Minister in the UK. May portrayed a high AR due to the PM's high frequency of speeches. However, May's PR was almost non-existent due to the British custom of addressing the Prime Minister as "Prime Minister". Our algorithm failed to detect the title as a mention, showing a critical case where our methodology falls short. Nokes portrayed the highest AR within the network, yet her PR is also marginal. The highest PR was held by Angus MacNeil of the Scottish National Party, who was on the Joint Committee for the National Security Strategy. However, we noted a high degree of self-references in MacNeil's case. Another politician with high AR was Sajid Javid, who was State Secretary and worked in housing and community.

5.2 The overall picture

For each network described, we could identify generally prominent PMs and topic-specific politicians who, e.g. held relevant ministry positions. On the one hand, we concluded that our findings do reflect the structure of each parliament and offer an insight into the power distribution of MPs in connection to certain topics. On the other hand, we established that interpreting politicians' positions within networks heavily depends on having knowledge of the country's specific political context and cannot be done based on the provided metrics alone.

Correspondingly, we discovered a critical weakness of the named entity tagging in the UK, where the custom of addressing others using titles instead of personal names distorted our results. While hardly unique to the UK, issues such as this are easily solved. Preferably, however, the matter could have been prevented with a more profound acquaintance of particular parliament traditions.

Our most important discovery, however, regards female MPs's equality within political debate. While the Spanish parliament had the highest share of female MPs (41.2%), we demonstrated that their Active Relevance (AR) in almost all topics (except for healthcare) was lower than their overall presence and, even more concerningly, their Passive Relevance was mostly under 15%. We revealed that the mere presence of female MPs in the parliament does not warrant their participation in debates. While holding the same position, female MPs have less opportunity to make their opinions heard.

6 Conclusion

This paper investigated power distribution between members in three national parliaments in Europe: the UK, Spain, and Slovenia. We measured argumentative power through the relative amount of speeches per member, which we defined as Active Relevance (AR), and the relative amount of mentions of a member by others, which we defined as Passive Relevance (PR). Furthermore, we analyzed structural power through the gender distribution of argumentative power within selected topics.

Using statistical and social network analysis on transcribed parliamentary debates, we concluded that a parliamentary member's gender could affect their argumentative power. For Spain, we saw a comparatively high percentage of female MPs in parliament. Nevertheless, this did not cause a similarly high portion of female PM speeches or mentions. Our results pointed to the same issues Bäck et al. (2014) identified for the Swedish parliament almost a decade earlier.

Our results suggest that structural power organized by gender within a single parliament has a topical variance. Still, there was little evidence for the same impact on gender distributions of specific topics in all three countries. We found that healthcare, one of the topics predefined as "soft" (i.e. female-dominated), presented distinctly female–dominated AR in all parliaments. On the other hand, the measured AR and PR of the assumed "hard" (i.e. male-dominated) topics, energy and finance, revealed a conclusive male majority in both Slovenia and Spain.

Immigration, initially deemed "soft" and ultimately established an ambiguous topic, was qualitatively considered for each country. Consequently, we demonstrated that understanding the political context of the country in question is essential to analyzing power distributions in its parliament.

Many of the topics revealed in our work deserve further exploration. Above all, this paper displays

that multidisciplinary research can disclose meaningful results and expose designs that serve as harmful blocks in achieving equality.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools, and methods (2022-2027), and the DARIAH-SI research infrastructure. The work presented in this paper is supported by the Social Sciences & Humanities Open Cloud (SSHOC) project[2] as well as ParlaMint project[3].

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Mykola Andrushchenko, Kirsi Sandberg, Risto Turunen, Jani Marjanen, Mari Hatavara, Jussi Kurunmäki, Timo Nummenmaa, Matti Hyvärinen, Kari Teräs, Jaakko Peltonen, et al. 2022. Using parsed and annotated corpora to analyze parliamentarians' talk in finland. *Journal of the Association for Information Science and Technology*, 73(2):288–302.

Hanna Bäck, Marc Debus, and Jochen Müller. 2014. Who takes the parliamentary floor? the role of gender in speech-making in the swedish riksdag. *Political Research Quarterly*, 67(3):504–518.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

Paul Bayley. 2004. *Cross-cultural perspectives on parliamentary discourse*, volume 10. John Benjamins Publishing.

Karin Bischof and Cornelia Ilie. 2018. Democracy and discriminatory strategies in parliamentary discourse. *Journal of Language and Politics*, 17(5):585–593.

Luke Blaxill. 2013. Quantifying the language of british politics, 1880–1910. *Historical Research*, 86(232):313–341.

Darko Cherepnalkoski and Igor Mozetič. 2016. Retweet networks of the european parliament: Evaluation of the community structure. *Applied network science*, 1(1):1–20.

Teun A van Dijk. 2018. Discourse and migration. In *Qualitative research in European migration studies*, pages 227–245. Springer, Cham.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaur Barkarson, Steinór Steingrímsson, et al. 2022. The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Thomas MJ Fruchterman and Edward M Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

Jose Gonzalez. 2021. Thefuzz. https://github.com/seatgeek/thefuzz.

Cornelia Ilie. 2010. *European parliaments under scrutiny: Discourse strategies and interaction practices*, volume 38. John Benjamins Publishing.

Cornelia Ilie. 2015. Parliamentary discourse. *The International Encyclopedia of language and social interaction*, pages 1–15.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Talcott Parsons. 1963. On the concept of political power. *Proceedings of the American philosophical society*, 107(3):232–262.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Pavel Rychlỳ. 2007. Manatee/bonito-a modular corpus manager. In *RASLAN*, pages 65–70.

A. Schwan and S. Shapiro. 2011. *How to Read Foucault's Discipline and Punish*. How to Read Theory. Pluto Press.

Herbert A Simon. 1953. Notes on the observation and measurement of political power. *The Journal of Politics*, 15(4):500–516.

Vladislava. Stoyanova and Eleni Karageorgiou. 2019. *The new asylum and transit countries in Europe during and in the aftermath of the 2015/2016 crisis*, volume 13 of *International refugee law series*. Brill, Boston, USA.

Teun A Van Dijk. 2000. *On the analysis of parliamentary debates on immigration*. Citeseer.

[2] https://www.sshopencloud.eu
[3] https://www.clarin.eu/parlamint

Zeitenwenden: Detecting changes in the German political discourse

Kai-Robin Lange and **Jonas Rieger** and **Niklas Benner** and **Carsten Jentsch**
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
{kalange, rieger, benner, jentsch} @statistik.tu-dortmund.de

Abstract

From a monarchy to a democracy, to a dictatorship and back to a democracy – the German political landscape has been constantly changing ever since the first German national state was formed in 1871. After World War II, the Federal Republic of Germany was formed in 1949. Since then every plenary session of the German Bundestag was logged and even has been digitized over the course of the last few years. We analyze these texts using a time series variant of the topic model LDA to investigate which events had a lasting effect on the political discourse and how the political topics changed over time. This allows us to detect changes in word frequency (and thus key discussion points) in political discourse.

1 Introduction

"Wir erleben eine Zeitenwende" – "We are witnessing a turn of eras". This quote by Germany's chancellor Scholz (Deutscher Bundestag, 2022) was the result of a turning point of political discourse in Germany after the outbreak of the Russian-Ukrainian war of 2022. This was an obvious turning point in German politics, as not only the discourse within the parliament, but also the decision making of the German government changed. For instance, 100 billion euros are planned to be spend as additional military expenses (Deutscher Bundestag, 2022). Throughout German history, there have been many changes and turning points in political discourse, but not all of them are as clearly reported or as obvious as this one. Many of them may not have been as clearly remembered because they were gradual rather than rapid changes, or because they did not have an immediate impact on real-world politics.

In this paper, we investigate all 72 years of plenary sessions of the German Bundestag to grasp the development of political discourse in Germany. We analyze these sessions by interpreting them as a time series of textual data for which we use a change detection method proposed by Rieger et al. (2022). For this, we use a rolling version of the topic model latent Dirichlet allocation (LDA), which is designed to construct topics that are coherent over time and allows for changing vocabulary. Within the resulting topics, we detect changes by analyzing the actual word usage in topics compared to theoretically expected word usage if no change occurs, which is determined using resampling. With this method, we are also able to differentiate between short-term changes and persistent ones. While "Zeitenwende" is a broad term, we refer to it as a persistent change in the way of how a political topic is discussed. This change in discussion can stem from differing speech patterns or from changing contents of said topic. The former is rarely detected as a change, as a changing speech pattern usually develops slowly. If it does change suddenly, this is, with minor exceptions, due to a change in formality, which is of no interest for our analysis as it does not affect the content of the discussions. The latter however is interesting for our analysis, as it symbolizes that the topic has changed for good due to economic, cultural or diplomatic events or developments. While this change may not affect the entirety of the political spectrum, we consider it to be a Zeitenwende, if it persistently changes one of the 30 most common and important topics of German political discussions, which we analyze using our topic model.

Data sets as large as the German parliament discussions are too large to be read and interpreted all manually. While qualitative expert analysis is needed to analyze German politics over the last decades, a quantitative text analysis can help in this regard by providing experts with ideas for what to look at and by verifying their results with an empirical basis. Our analysis aims to do just that,

as the model used allows for changes of different magnitudes to be detected by simply adjusting a single parameter. Experts can use these findings to back up their qualitative results by pointing out the importance and long lasting effect of a change for the political discussions at the time or gather ideas for changes in uncommon topics. By tuning the so-called "mixture"-parameter of this model, even rather niche changes can be found to interpret and to compare them based on the tuning needed to detect them (a high tuning parameter indicates a major change, a smaller one indicates a niche change).

Walter et al. (2021) use a similar data set to analyze ideological shifts throughout German history. This data set, DeuParl (Kirschner et al., 2021), does contain data from plenary sessions since 1867 to 2020. The data before 1949 are however less structured and contain clusters of incoherent text due to being automatically created by scanning old documents, which is why we do not use them in this analysis. While cleaning the data set is a task on its own, political analyses such as this one could greatly benefit from a "clean" version of this Reichstag data set as it enables to analyze German politics for an even larger period of time.

In a complementary approach, Jentsch et al. (2020, 2021) propose a (time-varying) Poisson reduced rank model for party manifestos to extract information on the evolution of party positions and of political debates over time.

2 Change detection via rolling modeling

We make use of a rolling version of the classical LDA (Blei et al., 2003) estimated via Gibbs sampling (Griffiths and Steyvers, 2004). This method is referred to as RollingLDA (Rieger et al., 2021) and allows new data to be added without manipulating the LDA assignments of the previous model. For this, a more reliable version of the classical LDA is used up to a date `init`. Then, according to a user-specified periodicity, minibatches of documents (`chunks`) are modeled using the data available up to that point. Moreover, the model's knowledge of previous documents is constrained in that it only uses the LDA assignments from a given time period (`memory`) to initialize the modeling of the new texts. Newly occurring vocabulary is added to the model vocabulary and subsequently considered as soon as it occurs more than five times in a minibatch. This flexibility enables the model to adapt for mutations of topics in the form of gradual or abrupt changes in word frequencies.

The minibatches are numbered in ascending order starting with the initialization batch: $t = 0, \ldots, T$. Then, using the change detection algorithm by Rieger et al. (2022), we get our set of detected changes over time by

$$C_k = \left\{ t \mid \cos\left(n_{k|t}, n_{k|(t-z_k^t):(t-1)}\right) < q_k^t \right\},$$

where $0 < t \leq T$ refers to a specific minibatch and $k \in \{1, \ldots K\}$ to one topic. As proposed by Rieger et al. (2022), $q_k^t \in [0, 1]$ denotes the 0.01 quantile of the set of cosine similarities when $n_{k|t}$ is replaced by $\tilde{n}_{k|t}^r, r = 1, \ldots, 500$, where $\tilde{n}_{k|t}^r$ denotes a resampled frequency vector under expected change and $n_{k|t}$ the observed vocabulary frequencies for each topic; analogously $n_{k|(t-z_k^t):(t-1)}$ refers to the sum of the count vectors from time points $t - z_k^t$ until $t - 1$. The algorithm has two parameters: the maximum length of the reference period to compare to, $z_k^{\max}$, and the intensity of the expected change under normal conditions p. Using the mixture-parameter $p \in [0, 1]$, which can be tuned based on how substantial the detected changes should be, the intensity of the expected change is considered in the determination of this estimator by

$$\tilde{\phi}_k^{(t)} = (1 - p)\, \hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} + p\, \hat{\phi}_{k,v}^{(t)}.$$

Depending on the choice of this parameter, we are able to gradually alter the magnitude of change needed to be detected by the model. While a large p only displays the most impactful changes, which are likely widely known, a smaller value for p allows for experts on this topic to identify more niche changes.

3 Evaluation

3.1 Data set

To analyze the German political landscape, we use the protocols of plenary sessions of the German Bundestag. These were collected over the course of 72 years, starting from the first plenary session of the Federal Republic of Germany on the 7th of September 1949 until the the 3rd of June 2022 in the 20th legislative period. Each protocol can be downloaded from the website of the German Bundestag (Deutscher Bundestag, 2016) and is provided in an XML-format, which contains, among other things, the date and entire plenary discussion

in a text format. As one plenary session might contain multiple topics and points of discussion, we split these texts into smaller texts. Because there are a total of 4345 sessions we aim to split these texts automatically instead of manually and do so by splitting them into individual speeches using regular expressions. We also deleted the attachments and registers, as well as heckling and comments. This is an ongoing work but already provides better results than splitting the texts any arbitrary number of tokens or using the original plenary sessions as single documents. In total, the 4345 plenary sessions are split into 335 065 documents. The distribution of documents by legislative period is displayed in the appendix in Table 1. The chunks of RollingLDA are adjusted to match the legislative periods, where each period is split into eight chunks (approximately two chunks per year).

3.2 Study design

For this study, we examined the different topic numbers $K = 20, \dots, 35$ each with $\alpha = \eta = 1/K$. For the RollingLDA we used the first legislative period as the initialization of the model. Starting from this, we modeled semi-annual minibatches, each using the last two years as memory. We applied the change detection algorithm with $p = 0.90, 0.91, \dots, 0.95$ and $z_{\max} = 4$, i.e., for the detection of changes, a maximum of the previous 4 minibatches ($\sim$2 years $\approx$ `memory`) are taken as the reference period. If a change is detected for topic k at time t, z_k^{t+1} is set to 1, else to $\min\{z_k^t + 1, z_{\max}\}$.

3.3 Results

Upon inspection of the results for the different parameters, we choose to present the findings for $K = 30$ and $p = 0.94$ in detail, yielding an interpretable number of detected changes while providing logical topics which can be analyzed separately from another and consistently over time. The following results serve as a proof of concept, as for a more fine-grained analysis in the future, a lower value of p can be used. This way, the model will detect changes with a lesser impact on the topic, which will enable experts on German politics to identify changes that had an impact on German politics but may not be as well-known as the results we present here. All detected changes, corresponding top words, our interpretations and the results for other parameters can be accessed via the associated GitHub repository (K-RLange/Zeitenwenden).

The changes are displayed in Figure 1. The blue and red curves represent the observed similarities and the simulated quantile similarities, respectively. Each time the blue is below the red line, a change is detected as a gray vertical line.

The topics can be separated into political topics, which contain information about the current political discussions, and formality-topics which contain the names and titles of the parliament members as well as key words for common procedures, such as the voting process when deciding about a bill. While changes can be detected in either type of topics, changes in formality-topics will most likely not contain any information about the current political situation or discussion but rather about common political procedures or who is a current member of the parliament. Topic 9 for instance is a topic that is almost completely consisting of the names of parliament members. All 7 changes are detected at the start of a new legislative period, which is reasonable as new politicians join the parliament, but is not interesting for the sake of our analysis. Similarly, topics 4, 7 and 22 yield multiple changes that can be explained by a change in procedure or a different style of logging. Thus, we focus on the remaining 26 topics when looking for Zeitenwenden that were rooted in the topics of political discussions. We are able to link 22 of our 25 detected changes in relevant topics to interprable events. The remaining changes are caused by events that we were not able to interpret in retrospect.

Our model is able to detect some obvious events which affected political discourse such as the Russian-Ukrainian war (2022, topic 21, Deutscher Bundestag, 2022), the Covid 19 pandemic (2020-21, two changes in topic 17, Organization, 2020), the European financial crisis in 2008 (topic 15, Gode, 2021), the introduction of the Euro as Germany's currency (2002, topic 28, Directorate-General for Communication, 2022), the Kosovo-war (1999, topic 21, Beaumont and Wintour, 1999), the German Reunification (1989-91, topics 1, 12, 15, Schmemann, 1989), the founding of the Bundeswehr (1955, topic 16, Bundeswehr, 2022) and the Saar-referendum (1955, topic 3, Jaeckels, 2020). Events like these had a long lasting impact on German society and politics and could be called "Zeitenwenden". Interpreting the context of change is particularly easy, as the RollingLDA-model provides us with information about the overall topic of the change consistently over time. The Kosovo and Russian-Ukrainian war are for instance both

Figure 1: Observed similarity (blue), thresholds q_k^t (red) and detected changes C_k (vertical lines, gray) over the observation period for all topics $k \in \{1, \ldots, 30\}$.

detected in topic 21, which can be interpreted as the "war"-topic. To identify the exact reason for the change, we analyze the impact of each word using leave-one-out word impacts. Such word impact graphs are displayed in Figure 2 and Figure 3 for both mentioned wars. While most changes are caused by words being used more frequently due to a new event (blue bars), some are also caused by words that are used significantly less (red bars). The financial crisis of 2008 is a case in which the change is caused both by a change of focus within an event, as "ikb" was mentioned far less frequency, while words such as "krise" (crisis) started to emerge (see Figure 5).

While these were major changes which had a lasting impact on Germany, there are several smaller changes that are detected as well, such as the Bonn-Copenhagen declarations in 1955 recognizing the danish minorities in Schleswig-Holstein (topic 21, Federal Foreign Office, 2015), the removal of the statute of limitations on murder (1979, topic 19, Schmid, 2017) and the tax reform of 1998 (topic 26, Tagesschau, 2010).

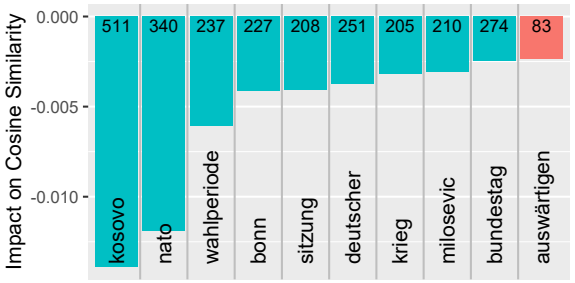

Figure 2: Leave-one-out word impacts for topic 21 (1998-99), caused by the Kosovo war.

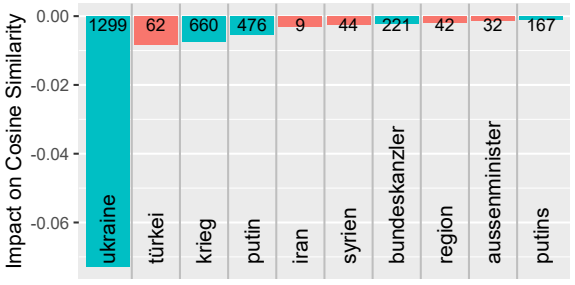

Figure 3: Leave-one-out word impacts for topic 21 (2021-22), caused by the Russian-Ukrainian war.

Our results also enable us to reflect on the relationship between both German states BRD and DDR as well as BRD and western powers over the course of 40 years by interpreting the corresponding changes of major discussion points. West-Germany began major partnerships with western countries, such as the EGKS, a German-French cooperation that was founded according to the "Schuman-Plan" making the coal and steel-industries of both countries a European rather than a national matter (topic 3, Zandonella, 2021). The Bonn-Paris conventions were also a result of both a closer connection towards western powers such as France and the troubled relationship between West-Germany and the eastern block (including East-Germany), as West-Germany became a member of NATO in 1955 (topic 1, Küsters, 2015) and introduced a mandatory conscription in 1956 (topic 27, Bundeszentrale für politische Bildung, 2016). All of this lead to the second Berlin-crisis in 1958, in which the Soviet Union demanded West-Berlin to become a free city rather than a part of West-Germany (topic 25, Barker, 1963). Still, the NATO was not left unquestioned though, as the piece demonstrations in Bonn in 1981 against the NATO Double-Track Decision were a major discussion point in the Bundestag (topic 15, Der Spiegel, 1981). In 1990 West- and East-Germany unified. This is detected in several topics (1, 12, 15), as it was an long process which had a lasting impact in almost every political sectors, such as financial politics, inner politics, outer politics and many more. In 1991, a distinction was made between the "Neue Bundesländer" and "Alte Bundesländer", denoting the parts of former East- and West-Germany after the unification. This was important as the parts of former East-Germany needed additional financial help to stabilize and reach the economical level of the western parts (topic 12). Ultimately, the usage of the word "DDR" decreases heavily in 1991 after both states had dissolved (topic 1, see Figure 4).

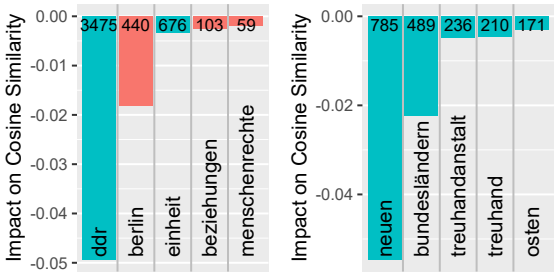

Figure 4: Leave-one-out word impacts for topics 1 (1989-90) and 12 (1990-91) concerning East-Germany.

4 Summary

To identify turning points in German political discourse, we analyzed plenary sessions of the German Bundestag from 1949 to 2021 using a change detection algorithm. This algorithm is based on a rolling version of the topic model LDA to create topics that are comparable across time. The changes detected reflect a significant change in the word distribution of the topics.

Our algorithm detects several meaningful changes over the course of the last 68 years of plenary discussions, such as key moments of the relationship between West- and East-Germany as well as world political events like the Russian-Ukrainian war, the Covid 19 pandemic and the financial crisis of 2008.

While these changes are identifiable as "true changes", we do not know how many changes we missed, as major political discussions in the 21st century such as the refugee crisis in 2014 are not detected. This might be caused by a mixture-parameter that was chosen too restrictively or by the inability of the algorithm used to detect changes in topic distribution (see Figure 6), as it is based on word distribution. Thus, topics that are suddenly a lot more relevant are not detected if the vocabulary used did not change. Identifying both would improve this analysis. Along with adjusting the mixture-parameter, this may enable a detailed analysis of Germany politics for experts on this topic. This can be further amplified by cleaning and using plenary sessions from 1867 to 1945, of East-Germany and of German state parliaments in addition to the Bundestag data set that we used here, as this would enable us to cover a broader spectrum of Germany's political discourse and history.

Acknowledgments

The present study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Elisabeth Barker. 1963. The berlin crisis 1958–1962. *International Affairs*, 39(1):59–73.

Peter Beaumont and Patrick Wintour. 1999. Kosovo: the untold story. *The Guardian*.

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

Bundeswehr. 2022. The establishment of the Bundeswehr.

Bundeszentrale für politische Bildung. 2016. 1956: Das wehrpflichtgesetz tritt in kraft.

Der Spiegel. 1981. Schmeiß die Atomwaffen in die Gracht.

Deutscher Bundestag. 2016. Deutscher Bundestag - Open Data.

Deutscher Bundestag. 2022. Deutscher Bundestag - Bundeskanzler Olaf Scholz: Wir erleben eine Zeitenwende.

Directorate-General for Communication. 2022. History and purpose of the euro.

Federal Foreign Office. 2015. Joint german-danish declaration on the 60th anniversary of the bonn-copenhagen declarations.

Solveig Gode. 2021. Bilanz der Finanzkrise in Deutschland: Die Rolle von Georg Funke, dem Gesicht der Finanzkrise“, und der Hypo Real Estate Bank.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Ulrich Jaeckels. 2020. Abstimmung 1955! - Ja oder Nein?

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2020. Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142:106813.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2021. Poisson reduced-rank models with an application to political text data. *Biometrika*, 108(2):455–468.

Celina Kirschner, Tobias Walter, Steffen Eger, Goran Glavas, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Deuparl.

Hanns Jürgen Küsters. 2015. Inkrafttreten der Pariser Verträge.

World Health Organization. 2020. Who director-general’s opening remarks at the media briefing on covid-19.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2021. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, pages 2337–2347. ACL.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. 2022. Dynamic change detection in topics based on rolling LDAs. In *Proceedings of the Text2Story’22 Workshop*, volume 3117 of *CEUR-WS*, pages 5–13.

Serge Schmemann. 1989. Clamor in the east: Germans’ special times; reunification next? *The New York Times*.

Sandra Schmid. 2017. Deutscher Bundestag - Historische Debatten (4): Verjährung von NS-Verbrechen.

Tagesschau. 2010. Jahresrückblick 1999: Sparpaket und Steuerreform.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš;, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–60. IEEE.

Bruno Zandonella. 2021. Schuman-plan. In *pocket europa. EU-Begriffe und Länderdaten*. Bundeszentrale für politische Bildung.

A Additional Material

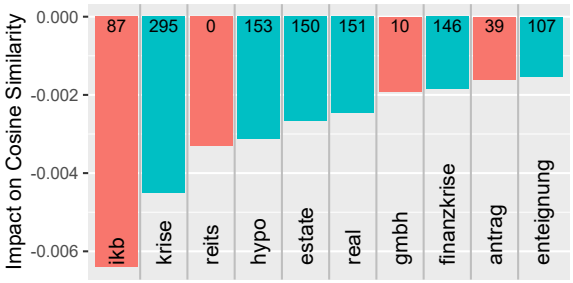

Figure 5: Leave-one-out word impacts for topic 15 (2008-09), caused by the financial crisis.

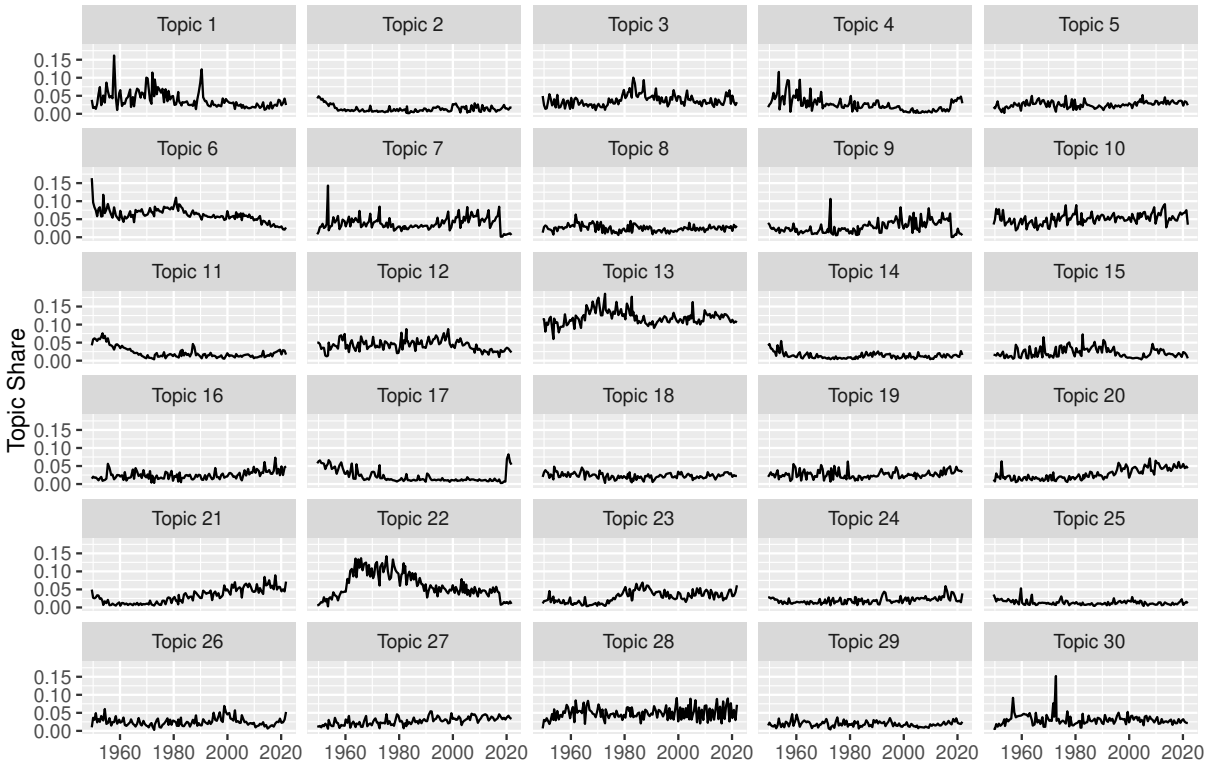

Figure 6: Topic shares per chunk: Relative number of assignments to the specific topic for a given time period.

Table 1: Approximate number of documents in relevant topics (see Section 3.3) and detected changes for each legislative period.

Period	Start date	Documents	Changes
1	1949-09-07	16868	NA
2	1953-10-06	7963	7
3	1957-10-15	7365	3
4	1961-10-17	10900	0
5	1965-10-19	17265	0
6	1969-10-20	16540	0
7	1972-12-13	19917	0
8	1976-12-14	18199	1
9	1980-11-04	10954	2
10	1983-03-29	20812	1
11	1987-02-18	20087	1
12	1990-12-20	19780	3
13	1994-11-10	20999	0
14	1998-10-26	18582	2
15	2002-10-17	13252	1
16	2005-10-18	19867	1
17	2009-10-27	25909	0
18	2013-10-22	20752	0
19	2017-10-24	25163	2
20	2021-10-26	3891	1

Comparing the coverage of the "marriage for all" vote on Twitter and in the newspapers

Maud Reveilhac
Lausanne University
Institute of Social Science
maud.reveilhac@unil.ch

Gerold Schneider
Zurich University
Department of Computational Linguistic
gschneid@ifi.uzh.ch

Abstract

This paper investigates the differences in the Twitter and newspapers coverage of the "marriage for all" popular vote that took place in Switzerland in 2021. More precisely, we ask the following questions: How salient were discussions about the marriage for all on Twitter and in the newspapers? What major arguments were mobilized in both media? How were these arguments received (i.e. retweets, likes, replies)? We extracted publicly available tweets from users involved in the debate and news articles containing specific keywords. These text data have been automatically analyzed to find major views and topics of discussions using keyword and collocation analyses, as well as topic modelling. Results show that criticism of marriage for all is clearly in the minority, but there is strong polarization over whether same-sex couples should be allowed to adopt or have children through sperm donation.

1 Introduction

On 26 September 2021, Switzerland held a popular vote to insert a provision for "marriage for all" (henceforth MFA) into its constitution. The current study explores the public debate surrounding this major event on Twitter and in the traditional newspapers. In particular, we investigate how these media appeal to different views and topics of the debate. Newspapers are a valuable source of information for citizens to understand the representations and claims related to the important societal matters, such as the MFA (Eisenegger, 2019). However, traditional news media influence is just one of the numerous sources of information in the new media environment. Individuals increasingly produce, rather than passively consume, media content, especially on social media platforms, such as Twitter, where social issues are extensively debated (Allsop, 2016). In our study, Twitter data complement the scope of mass media data with insights into individuals' immediate and personal views about the MFA. Our article seeks to uncover the differences in Twitter and newspapers coverage of the MFA. To do so, we rely on Switzerland as a case of study because the Swiss population had the opportunity to vote on this major societal issue on September 26th 2021, which passed with 64% of the votes. The public debate was emotional and underscored polarizing subjects, such as the adoption rights and surrogate motherhood. The role of social media was particularly interesting during the voting campaign leading to strong polarization of the arguments. In order to discover the main views and ideologies in our collections of tweets and articles, we first rely on keywords and collocation analyses, thus drawing from a corpus-assisted framework (McEnery and Baker, 2015), and second, topic modelling. Examples of text genres analysed with topic modelling include, for instance, newspaper text (Blei, 2012), micro-blogs (Surian et al., 2016) and open-ended survey questions (Baumer et al., 2017). This enables us to assess the prevalence of topics across media sources (Twitter and newspaper articles). We show how these two analytical steps can be applied to a collection of tweets and newspaper articles for better understanding topics and arguments surrounding the public debate about the MFA.

2 Theoretical background

Along with conducting surveys of public attitudes toward societal issues and to activist content available on social media, the use of textual data is increasingly central to uncover themes and activities surrounding important political events. In the case of the MFA, O'Connor (2017) showed the importance to compare several data sources to obtain a more nuanced picture of same-sex marriage repre-

sentations. The authors studied the role played by appeals to nature in the 2015 Ireland referendum to legalize same-sex marriage. Through content analysis of newspaper and Twitter discussion of the referendum, this study showed that appeals to nature occurred in a minority of media discussion of the referendum, but were more prominent in material produced by anti-marriage equality commentators, while predominantly occurring in relation to parenthood, traditional marriage, gender, and homosexuality.

Studies about MFA substantively benefited from the reliance on methods developed in computational sciences to investigate social phenomena. Indeed, previous studies focused on media representations of same-sex/equal marriage debates in different national contexts. For instance, drawing upon corpus linguistics and critical discourse analysis, Kania (2019) investigated two key periods from 2000 to 2017 about the German press coverage of marriage equality legislation. Relying on keywords and collocation analyses, the author evidenced the different discourses drawn upon by the German media, notably in terms of pro- or anti-marriage equality voices, and demonstrated how colloquialisms such as *Homo-Ehe* (homo-marriage) came to be accepted and used across media texts as legislation changed.

Unsupervised classification methods are useful to extract relevant content about the MFA. For instance, Hemmatian et al. (2019) relied on comments on Reddit from January 2006 until September 2017 and used a topic model to investigate how the changes in the framing of same-sex marriage in public discourse relate to changes in public opinion. They show that the contributions of certain protected-values-based topics to the debate (religious arguments and freedom of opinion) increased prior to the emergence of a public consensus in support of same-sex marriage found in surveys, and declined afterward, in contrast to the discussion of certain consequentialist topics (the impact of politicians' stance and same-sex marriage as a matter of policy).

3 Data collection and methods of analysis

3.1 Extraction from tweets and newspapers

Given the Swiss popular vote on the MFA in 2021, we decided to focus on the content surrounding this public debate that can be found online (e.g. on Twitter) and offline (e.g. national and regional newspaper articles). With the different media sources, we have a collection of long (articles) and short (tweets) documents that must be rearranged to make them comparable in size. We therefore decided to define a document as a paragraph with respect to articles and as the sum of a given user's messages regarding Twitter. Figure 1 displays the filtered number of collected documents over time in relation to major events.

For the newspaper articles, we extracted all articles containing the keywords *Ehe für alle* and *mariage pour tous* (N=2,705 articles) which are then split into paragraphs (n=37,020). The included newspapers are all daily Swiss news from Swissdox[1]. For conducting topic modelling, we kept only paragraphs that contained the keywords (we used the following lowercase queries: *ehe.*fuer.*alle, ehe.*für.*alle, heirat.*, ehe, homo.*, lgbt.*, lesbisch.*, gleichgeschlecht.*, gay.*, leihmutter.*, fortpflanz.*, kinder.*, famil.*, samenspende.*, partnerschaft.*, adopti.**). The keywords are manually cleaned for a list of candidate terms that appeared in the top terms of the Twitter documents. We selected only the news articles written in German, that contain at least five words, and that were written between January 1 and October 1 2021 (12,335 paragraphs from 2,367 articles).

For Twitter, we collected tweets from the main committees involved in the popular vote and from other very involved users who are either members of the committees or political actors who declared being supportive or opposing the vote. Tables 1 and 2 in the Appendix provides the detailed Twitter handles supporting and opposing the vote. Additionally, we collected tweets from the followers of the committees selecting tweets that focus on the MFA using a list of search queries (keywords and hahstags): *ehe, ehe für alle, ehe fur alle, ehefueralle, ehefüralle, jafüralle, jaichwill, ouijeleveux, loveislove, lovewins, mariage, mariage pour tous, mariagepourtous, mariagepourtoutes, mariagepourtoutesettous, rainbowswitzerland, loveisliberal, lgbt.**. This leaves us with 930 followers of the supporting committees and 70 of the opposing committee. Finally, we used only the hashtags (*ehefueralle, ehefüralle, jafüralle, jaichwill, ouijeleveux, loveislove, lovewins, mariagepourtous, mariagepourtoutes, mariagepourtoutesettous, rainbowswitzerland, loveisliberal*) pointing to the vote and collected tweets from 1,213 other users that

[1] https://swissdox.linguistik.uzh.ch

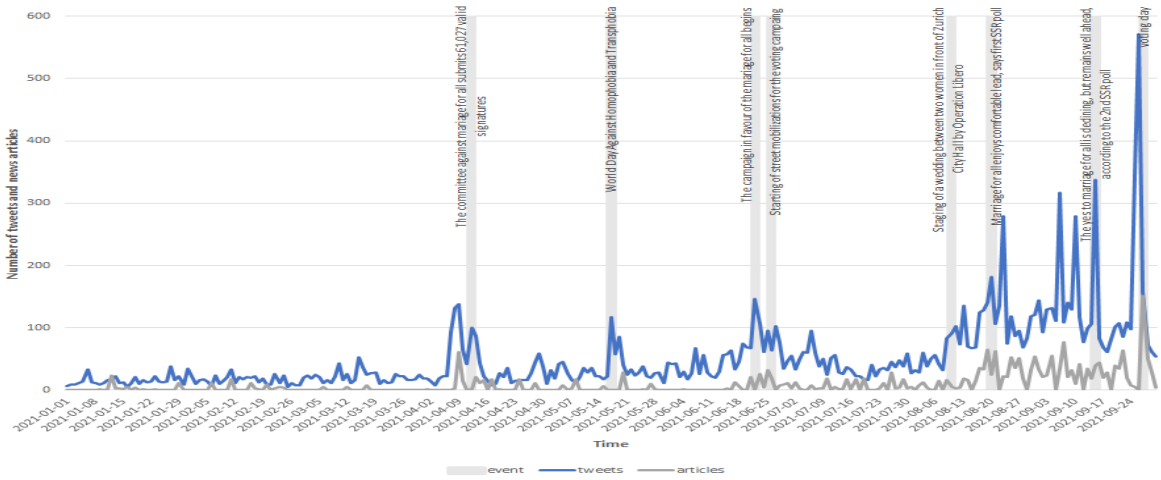

Figure 1: Distribution of tweets and articles over time.

match these queries.For the subsequent analyses, we grouped Twitter users into a general group (containing actors in Table 1 and the national committees' followers, as well as the other Twitter users) and an unsupportive Twitter group (containing actors in Table 2 and the national committee's followers) users towards the MFA. We kept only tweets containing at least five words and that were written between January 1 and October 1 2021 (N=13,983; 8,122 were written originally in German and the remaining tweets were translated into German using *GoogleTranslate*). Our Twitter corpus is thus a good approximation of the actual Twitter discussion on the MFA.

3.2 Keyword and collocation analyses

Our first research step is situated within corpus-assisted analysis (McEnery and Baker, 2015) and makes use of quantitative tools in order to uncover salient terms in newspaper and Twitter discourse on the Swiss MFA debate. We use statistical tests in order to identify words that (co)occur with a high frequency. We identify statistically salient keywords and collocations, relying on the library *tidylo* from Schnoebelen et al. (2022) in the *R* programming language. This first analytical step, in turn, helps to inform the interpretation of existing topics (see next subsection).

3.3 Topic modelling

The logic underlying topic modelling is based on word co-occurrence, that is, words that frequently appear together indicate a recurring topic (Blei, 2012). After conducting several models, we found that the best number of topics was 20 (one topic, the topic 8, was excluded from the analyses as it groups untranslated French terms). This is supported by reading the top-words associated with each topics and by reading the top-documents mostly associated with each topic.

We rely on the topic probabilities by document and use the most prevalent topic for each document. We also use the media type (tweets or articles) and the stance of the users (general or unsupportive) as a meta-information to build the semantic space underlying the MFA discussions. Before applying this model, we also conducted several pre-processing steps. Namely, we removed stop-words (e.g. *and, but, I, we, etc.*), replaced umlauts and accentuated letters, split concatenated expressions (e.g. *GenderEquality* becomes *gender equality*), and lowercased the text.

4 Results

4.1 Detected keywords and collocates

Tables 3 and 4 provide the top terms for the keyword and the collocation analyses.

For the identification of keywords, the terms found in newspapers and on Twitter were ranked according to effect size, based on Log Ratio, which indicates how big the difference in the statistical significance between corpora (namely, newspaper, Twitter supportive, and Twitter unsupportive) is for the identified keywords. We decided to focus on the top 3 keywords for each source. An overview of the results can be found in Table 3. It shows that the keywords refer to specific categories (see column 1) covering family life, sexual identities, involved organizations or political actors and stakeholders, and legislative issues.

A collocation analysis was done for the term marriage (in German: *ehe*). Collocations allow for exploring lexical associations, providing prototypical contexts and associations of a specific word (Firth, 1957). Each corpus was analyzed separately (namely, newspaper, Twitter supportive, Twitter unsupportive). Based on the top 5 results for each source (with a frequency of 5 being the group cut-off point and the selection including only nouns and adjectives), collocates were compared and categorized as either common or salient in a corpus. Table 4 displays the results and shows that the shared collocates are about the voting campaign (e.g. *referendum*) and discussions about the conception of family (e.g. *family, adoption, same-sex*). The collocates for the newspaper articles focus on the actors involved in the campaign (e.g. *opponent, befürworter*), while the collocates on Twitter relate to the marriage definition (e.g. *zivile*), the opening to same-sex couples (e.g. gleichstellung), as well as the risks (e.g. *vergewaltigung*). The collocates from unsupportive users on Twitter focus more specifically on the juridical aspects of the debate and on religious questions.

A main finding of the analysis of salient keywords is the strong contrast between framing the MFA in terms of a legal institution for same-sex couples (in terms of children rights, legality, and political process) and a more "engaged" framing (in terms of equal rights, marriage definition, and the opening to procreation and adoption rights). A main finding of the analysis of collocates is the common agenda shared by the traditional news media and Twitter discussions in terms of the definition of the family roles and the campaign organization. Overall, these findings provide converging evidence on how the marriage equality debate was represented in both media.

4.2 Detected topics and other users' reactions

We relied on the proportion of each topic across the documents, and we mean aggregated these proportion by media source as well as by the stance of the Twitter users (general and unsupportive). Figure 2 displays the prevalence of topics by ordering the topics according to the prevalence for the unsupportive Twitter group. It also shows the mean number of reactions (in terms of retweets, likes, and replies) received by each topic on Twitter (the metrics for measuring the reactions have been updated on June 2022).

We find that views of the traditional family (e.g. rainbow family and sperm donation as a criticisms) and religious arguments (e.g. religious aspects such as the calls to "nature") are important topics promoted by the unsupportive users on Twitter, which also indicating that these topics were the majors points of disagreement between both camps. These topics triggered on average little reactions from other users, thus showing that unsupportive users also received less attention on Twitter. In contrast, the topics that generated the most reactions are salient topics in the traditional news in terms of generic policy issues (e.g. arguments of the opposing and supporting camps, societal and progressist views), among which a significant contributor to the discourse is "LGBT-related anecdotes" that form the "rainbow family" topic in the traditional news, thus presumably due to increasing acceptance of LGBT experiences in media discourse. Another topic that received high attention from other users relates to the LGBT community in the tweets of supportive Twitter users.

5 Conclusions

The analyses suggest that there was a generalized support for the MFA in traditional newspapers and on Twitter. Anti-marriage equality arguments and topics (e.g. marriage definition, religious aspects, biological attributions, and traditional family) were hardly echoed in the press. However, this could also point to the difficulty to collect anti-marriage views, despite the fact that our data collection strategy also attempted to integrate unsupportive views (especially, via Twitter). A key strength of our data collection strategy is its real-world relevance. Including different media enables us to convey complementary findings related to the public debate about the MFA, notably because newspaper articles may encompass an elite bias that could be nuanced by the more spontaneous social media content.

Our case study does not come without limitations. First, our social media data come from only one specific platform, namely Twitter. However, Twitter is known to be more relied upon to express political views compared to other social media (e.g. Facebook). Second, our sample of social media data is skewed towards supportive opinions of the MFA. However, this trend is also reflected by representative opinion surveys, where only a minority of citizens were against the vote. The unsupportive people stemmed generally from the (extreme) right

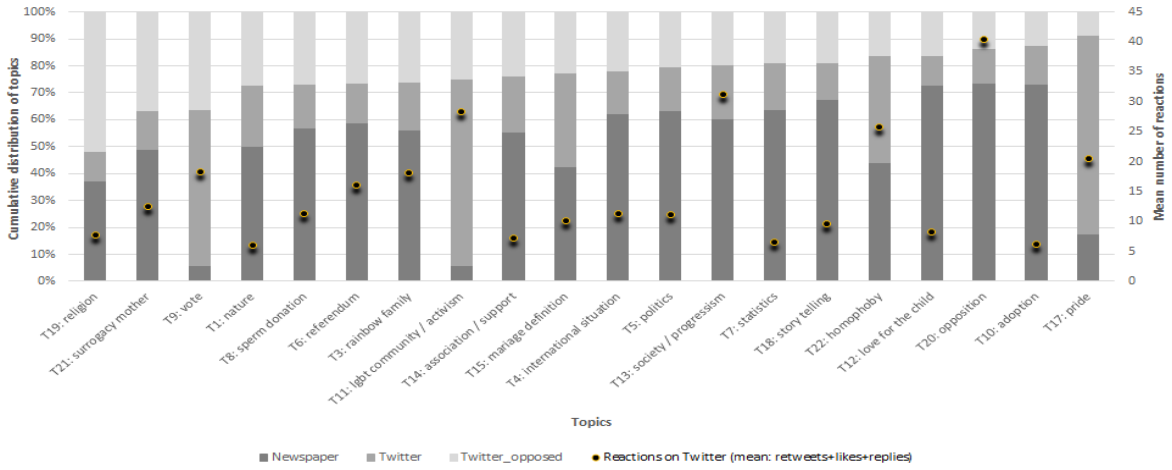

Figure 2: Distribution of topics over media types and Twitter user groups, including also the mean number of reactions on Twitter.

of the political spectrum, were people who distrust the government, were members with strongly religious backgrounds, and were overwhelmingly rejecting the same-sex parenting argument (Golder et al., 2021). Third, we focused on the text as the main source of information to conduct our analyses. However, future studies could also work with pictures that illustrate news articles and that are posted online. Images could convey additional insights about appeals to values and representations manifest in everyday discourse. For instance, we encountered Twitter posts that contained pictures of a crying baby exclaiming, "I don't have a mom".

Acknowledgments

The authors have no conflict of interest to disclose. The first author has planned the study and conducted the analyses. Both authors have contributed to the writing of the paper. We also would like to recognize the support that we received from Dominik Stammbach from ETH Zurich.

References

Bradley Allsop. 2016. Social media and activism: A literature review. *Social Psychology Review*, 18(2):35–40.

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

David Blei. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1):8–11.

Mark Eisenegger. 2019. *Main findings. The Quality of the Media – Switzerland. Yearbook 2019.* Zürich: Forschungszentrum Öffentlichkeit und Gesellschaft (fög).

John R. Firth. 1957. *Papers in Linguistics: 1934-1951*. Oxford.

Lukas Golder, Martina Mousson, Cloé Jans, Thomas Burgunder, Daniel Bohn, Roland Rey, and Corina Schena. 2021. *1ère enquête "SRG Trend" relative à la votation du 26 septembre 2021*. Available at: https://cockpit.gfsbern.ch/fr/cockpit/srg_trend_032021_fr-3-3.

Babak Hemmatian, Sabina J. Sloman, Uriel Cohen Priva, and Steven A. Sloman. 2019. Think of the consequences: A decade of discourse about same-sex marriage. *Behavior Research Methods*, 51(4):1565–1585.

Ursula Kania. 2019. Marriage for all ("ehe fuer alle")?! a corpus-assisted discourse analysis of the marriage equality debate in germany. *Critical Discourse Studies*, 17(2):138–155.

Anthony McEnery and Paul Baker. 2015. *Corpora and discourse studies: Integrating discourse and corpora*. Springer.

Cliodhna O'Connor. 2017. "appeals to nature" in marriage equality debates: A content analysis of newspaper and social media discourse. *British Journal of Social Psychology*, 56(3):493–514.

Tyler Schnoebelen, Julia Silge, and Alex Hayes. 2022. *Package "tidylo"*. Available at: https://cran.rapporter.net/web/packages/tidylo/tidylo.pdf.

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

6 Appendices

Table 1: List of Twitter handels supporting the marriage for all

Categories	Twitter.handels
National committee webpage	@ehefueralle_JA
	@JA_fuer_alle
	@mariage_OUI
National committee	@d_stolz
	@janphmueller
	@M_Erhardt
	@MariaVonKaenel
Member organisations	@LGBTfamilie
	@LOS_Schweiz
	@Network_CH
	@pinkcross_ch
	@RegenbogenZH
	@wybernet
Organisations	@Amnesty_Schweiz
	@Amnesty_Suisse
	@AvenirSocial
	@be_you_network
	@campaxorg
	@communelausanne
	@Dialogai
	@EtatdeVaud
	@FeministeGreve
	@FrauenstreikCH
	@Grevefeministe
	@HAB_lgbt
	@HabsBasel
	@haz_queer
	@infoTGNS
	@LilithLausanne
	@maennerch
	@MPT_CH
	@ncbischweiz
	@operationlibero
	@ProFamiliaCH
	@ProFamiliaVaud
	@projuventute
	@Queer_ch
	@queeramnesty
	@QueerBienne
	@queeroffice
	@queerstudents
	@SAJV_CSAJ_FSAG
	@stopsuicide1
	@VG_VoGay
	@VilleDeGeneve
	@Zonaprotetta
Parties	@FDP_Liberalen
	@GrueneCH
	@grunliberale
	@LesVertsSuisses
	@PLR_Suisse
	@PSSuisse
	@spschweiz
	@vertliberaux
	@SVP_Ehefueralle
Engaged politicians	@Caroline_Dayer
	@CLuisierBrodard
	@CottierDamien
	@FlueckigerYves
	@johanrochel
	@MathiasReynard
Member of SVP_Ehefueralle	@_MartinaBircher
	@larsguggisberg
	@MFrauchigerSVP

Table 2: List of Twitter handels opposing the marriage for all

Categories	Twitter.handels
National committee webpage	@Ehefueralle_NO
National committees / parties	@EDUSchweiz
	@Mitte_Centre
	@SVPch
	@UDCch
	@UDFVaud
Politicians	@dehaudt
	@EBertinat
	@GregoryLogean
	@marcochiesa74
	@MarcoRomanoPPD
	@martin_haab
	@monika_rueegger
	@PieroMarchesi1
	@roduitbenjamin
	@SchlaepferThe
	@udcvr64
	@verenaherzog
	@YvesNidegger

Table 3: List of top terms for the keyword analysis (TW_supp: Twitter supportive, TW_opp: Twitter unsupportive, news: newspaper)

Classification	Top terms (translation) - group
political actors	judf (judf) - TW_opp
	svp (svp) - TW_opp
	gegner (opponent) - TW_opp
	operationlibero (operationlibero) - TW_supp
	grunliberale (grunliberale) - TW_supp
	evppev (evppev) - TW_supp
	jugendkomitee (youth committee) - news
	parteipräsident (party president) - news)
	justizministerin (Justice Minister) - news
organisation	kirche (church) - TW_opp
	gesellschaft (society) - TW_opp
	gott (God) - TW_opp
	ilga (ilga) - TW_supp
	pinkcross (pinkcross) - TW_supp
	zurichpride (zurichpride) - TW_supp
	firmen (companies) - news
	syngenta (syngenta) - news
	mediensprecher (media spokesman) - news
legislation	recht (right) - TW_opp
	vorlage (proposal) - TW_opp
	referendum (referendum) - TW_opp
	homophobia (homophobia) - TW_supp
	hassverbrechen (hate crime) - TW_supp
	gleichheit (equality) - TW_supp
	schöpfung (creation) - news
	verfassungsebene (constitutional level) - news
	abklärungen (clarifications) - news
identities	lgbt (lgbt) - TW_opp
	gleichgeschlechtliche (same-sex) - TW_opp
	lesbische (lesbian) - TW_opp
	lgbt (lgbt) - TW_supp
	equality (equality) - TW_supp
	vielfalt (diversity) - TW_supp
	homosexuel (homosexual) - news
	weibchen (female) - news
	männchen (male) - news
generic	menschen (people) - TW_opp
	liebe (love) - TW_opp
	leben (life) - TW_opp
	together (together) - TW_supp
	geschichte (history) - TW_supp
	stolz (proud) - TW_supp
	urnengang (vote) - news
	gesellschaftspolitischen (socio -political) - news
	möglichkeit (opportunity) - news
family	kinder (children) - TW_opp
	paare (couples) - TW_opp
	samenspende (sperm donation) - TW_opp
	lgbtfamilie (lgbt family) - TW_supp
	spende (sperm) donation) - TW_supp
	hochzeit (marriage) - TW_supp
	erzeuger (genitor) - news
	gottgewollt (god-sent) - news
	schwanger (pregnant) - news

Table 4: List of top terms for the collocation analysis with the term marriage

Source	Top terms (translation, frequency)
common	abstimmung (poll, 233)
	begriff (expression, 50)
	adoption (adoption, 20)
	familie (family, 50)
	gleichgeschlechtliche (same-sex, 243)
newspaper	abstimmungskampf (voting campaign, 80)
	befürworter (supporter, 63)
	ausgabe (output, 41)
	urne (urn, 52)
	gegner (opponent, 235)
Twitter	abstimmen (decide by vote, 25)
	felseltern (same-sex parent, 13)
	gleichstellung (equality, 34)
	vergewaltigung (violence, 14)
	zivile (civilian, 40)
Twitter unsupportive	gesetz (law, 5)
	gott (God, 6)
	gottes (God, 5)
	heilig (holy, 5)
	polygame (polygame, 5)

Uncovering Policy Uncertainty Using Semantic Search Models

Sami Diaf and Florian Schütze
University of Hamburg, Germany
{sami.diaf, florian.schuetze}@uni-hamburg.de

Studying uncertainty contained in collections of documents has been a major task for political researchers and economists who aimed at measuring this degree using exclusively automated text analysis tools, as for sentiment analysis and topic models, to feed further inferences or test hypotheses. Such bag-of-word applications constraint the analysis and cannot render a clear picture of uncertainty drivers and their persistence, even if semi-supervised strategies may offer substantial improvements at the topic level. This work proposes a semantic search strategy, using Top2vec algorithm, to identify latent sources of uncertainty by uncovering a coherent topic structure whose representations will be used to get uncertainty prevalence and its persistence within documents and debates. As opposed to aggregate-level measurements, this strategy is suited to study per speaker debates at central banks where uncertainty is considered a forward guidance tool and a key strategy when devising monetary policy actions. Applied to FOMC transcripts, the resulting semantic space yields non-overlapping topic vectors indicating a dominance of economic discussions and forecasters‘ jargon in uncertainty formation within committee meetings, while risks concerns are bounded to financial markets using an investment jargon. Moreover, results demonstrate the importance of experts' contributions in steering the economic debate, hence coloring uncertainty with words not found in traditional uncertainty wordlists and diffusing a significant persistence to uncertainty prevalence within debates.

Moderation Mining on Social Media

Julian Dehne and Valentin Gold
Center of Methods in Social Sciences, University of Göttingen, Germany
julian.dehne@uni-goettingen.de, valentin.gold@sowi.uni-goettingen.de

We propose and compare three methods to identify content moderation on social media. In particular, we identify Tweets in which users take the lead in moderating their peers through posting their own reflections of the conversations. The data might then be used to train an AI intervention system to actively moderate social media content in the wild.

For many years, content moderation on social media focussed only on removing uncivil online behavior and verbal aggression. Recent developments in the field of AI have open up possibilities to go beyond deleting content, but also to automatically moderate conversations. Many of these approaches focus on hate speech only (e.g. Hangartner et al. (2021), Yildirim et al. (2021)) – a scenario when escalation already happened and suspension warnings are the only suitable tool to shape user behavior. Moderation, however, includes a wider set of strategies also taking "softer" modes of escalation into account: For instance, when conversations get off track and, at least in the perspective of some users, escalation is about to start, these users might spontaneously take over a moderating role trying to shape the behavior of their peers (Veglis, 2014, 143–144). It is the aim of this project to mine tweets for a large set of soft moderation modes.

Our theoretical approach builds upon two strands of literature: First, we build upon a functional theory of moderation (Edwards, 2002) separating moderation strategies in either strategic, conditioning, or process functions. The process functions, for instance, include sets of strategies to manage the conversation in view of its interactional goal, its agenda, and its schedule. However, not all of these strategies apply to the context of social media and, even more important, the context in which these strategies are employed is not explicitly stated. Hence, in a second step, the identified subset of moderation functions is separated into two dimensions: On the one hand, we separate between unitary and adversarial conflicts (Black et al., 2011); on the other hand, a moderation strategy might either foster dealing with conflicts that center around emotions or issues respectively. Using these combinations of dispositions as dimensions, we derive at a theoretical model for social media moderation.

To test the theoretical model, we propose and compare three different methodological approaches of mining moderating tweets. First, we propose a qualitative inductive approach: Based on the theoretical model, tweets are labeled as either moderating or not. Second, we use a computational algorithmic-based procedure to detect candidates of moderating statements. The algorithm is derived from the implicit and explicit theoretical expectations of when moderation is about to happen in a social media conversation. Similar to the first approach, the candidates are labeled accordingly. Our final qualitative deductive approach applies a dictionary of moderating terms and phrases. These keywords are then reviewed for their distinctiveness. In principle, each of these approaches have their merits and might be used to train a classifier that identifies moderating tweets. However, the applicability for social media conversations differs widely between these three procedures.

1. Bibliographical References

Black, L., Burkhalter, S., Gastil, J., and Stromer-Galley, J., (2011). *Methods for analyzing and measuring group deliberation*, pages 323–345. Taylor and Francis, January.

Edwards, A. (2002). The moderator as an emerging democratic intermediairy: The role of the moderator in internet discussions about public issues. *Information Polity*, 1(7):3–20.

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., and Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Veglis, A. (2014). Moderation techniques for social media content. In Gabriele Meiselwitz, editor, *Social Computing and Social Media*, pages 137–148, Cham. Springer International Publishing.

Yildirim, M. M., Nagler, J., Bonneau, R., and Tucker, J. A. (2021). Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics*, page 1–13.

Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller and Tim Landgraf
Free University of Berlin, Germany
solopov97@zedat.fu-berlin.de, {oana-iuliana.popescu, c.benzmueller, tim.landgraf}@fu-berlin.de

The full-scale conflict between the Russian Federation and Ukraine generated an unprecedented amount of news articles and social media data reflecting the opposing ideologies and narratives. These polarized campaigns have led to mutual accusations of misinformation and fake news, shaping an atmosphere of confusion and mistrust for readers all over the world. In this study, we analyze how the media affected and mirrored public opinion during the first month of war using news articles and Telegram news channels in Ukrainian, Russian, Romanian, French and English. We propose and compare two methods of multilingual automated pro-Kremlin propaganda identification, based on Transformers and linguistic features. We analyse advantages and disadvantages of both methods, their adaptability to new genres and languages, and ethical considerations of their usage for content moderation. Our analysis indicates that there are strong similarities in terms of rhetoric strategies in the pro-Kremlin media in both Ukraine and Russia. While being relatively neutral according to surface structure, pro-Kremlin sources use artificially modified vocabulary to reshape important geopolitical notions. They also have, to a lesser degree, similarities with the Romanian news flagged as fake news, suggesting that propaganda may be adapted to each country and language in particular. Both Ukrainian and Russian sources lean towards strongly opinionated news, pointing towards the use of war propaganda in order to achieve strategic goals. With this work, we aim to lay the foundation for further development of moderation tools tailored to the current conflict to help local human moderators and common users in these countries.

Contagious Populist Radical Right: The Role of Issue Salience for the Electoral Success in EP Elections

Sara Schmitt, Uwe Remer and Raphael Heiberger
University of Stuttgart
{sara.schmitt, uwe.remer, raphael.heiberger}@sowi.uni-stuttgart.de

A rich literature has revealed that mainstream parties gain from issue shifts towards topics owned by niche parties. Most studies use party manifestos in a cross-national context. So far, however, the direct communications of politicians and processes leading to issue ownership have been mostly disregarded. In addition, the effect of issue shifts in the European Parliament have been much less understood, although the European Parliament has recently seen a dramatic rise of radical right and populist right-wing parties (RRPP). We address these research gaps by focusing on Twitter posts of all Members of the European Parliament between 2014 and 2019 (over 3 million Tweets). Moving beyond manifesto data, we utilize the huge repositories of social media in a quantitative manner to study the salience shifts of parties in reaction to niche competitors. Structural Topic Models allow us to trace politicians communicative behavior dynamically and to explore empirically the issues RRPP occupy in the European Parliament. In particular, we reveal which parties "appeal broadly", which issues are owned by RRPP and how other parties might adopt issues owned by RRPP. By using Negative Binomial Regressions, we then link the patterns of issue diversity and the level of issue contagion to the outcomes of the European Parliament election in 2019. Our results show that adopting topics owned by RRPP has indeed a positive effect on election results of mainstream and left-wing parties. A closer look reveals, however, that it is moderated by how promptly parties adopt to these issues and whether they are in line with the party's ideological orientation. Furthermore, we observe strong negative effects if RRPP concentrate on typical right-wing issues.

Extracting Fuzzy Concepts from Online Job Advertisements in German

Johanna Binnewitt and Kai Krüger
BIBB – Bundesinstitut für Berufsbildung
{johanna.binnewitt, kai.krueger}@bibb.de

Online Job Advertisements (OJAs) are a valuable data source for various research communities and disciplines, including labor market analysis in social sciences. Their broad online availability allows monitoring labor market almost in real time. For this purpose, various Natural Language Processing (NLP) methods from the field of information extraction (IE) can be applied to extract information about required competences, used work tools or incentives offered by companies. Here, one challenge is to link existing theory-driven concepts with expressions in the text. One reason is that these theoretical concepts are fuzzy and complex (Deist and Winterton, 2005). Another reason is the linguistic variety in OJAs: recruiters often use creative wording to differentiate themselves from other companies (Engstrom et al., 2017), or corporate policy aspects can lead to the same content being expressed differently. Likewise, information in the text are formulated with varying degrees of implicitness. This leads to huge variations in span length and in the (conceptual and linguistic) complexity of expressed concepts. Consider the following example: (1) Willing to present balance sheets to stakeholders. (2) Proven extensive experience in presenting scientific results at relevant conferences. Both examples contain (inflected forms of) the keyword "to present". But we need to decide if extracting this keyword is sufficient or if we should try to preserve information like the requirement level, the attitude and the target audience. Decisions about what span to extract are (or should be) influenced by several factors, like the goal of the downstream analysis or more pragmatic reasons, such as whether the NLP model can handle the concept identification. In our research, we aim to identify further reasons why operationalizing IE is difficult and find ways to overcome these difficulties.

We argue that transparency of datasets and annotation guidelines can help with this, not only for OJA analysis, but in all areas where fuzzy concepts are examined using big data NLP approaches. Knowing which decisions a research team has made will allow other researchers to connect their research to it. For OJA analysis, this has often not been the case (Zhang et al., 2022).

More broadly, there are two contradicting trends observable within the NLP community. On the one hand, recent advances in NLP such as the introduction of transformers (Devlin et al., 2018) or innovative strategies such as prompt based NLP (Liu et al., 2021) allow researchers to build powerful NLP models for their custom problems with a decreasing amount of required labeled data. However, due to lack of transparency, standardized annotation guidelines and public datasets, research teams conduct incompatible research on the same subject. On the other hand, rather than exploring new fields, many NLP researchers focus on finding tiny im- provements over long-established datasets (Church and Kordoni, 2022). Here, datasets are publicly established. However, the overall contribution to the literature is questionable (ibid.).

We believe that gathering and annotating good datasets for unexplored problems is key to further facilitate NLP research, especially in areas such as Computational Political and Social Sciences or Digital Humanities. We advocate valuing papers that provide high transparency regarding concept definitions and source code rather than just reporting metrics. Consequently, this project aims to develop and publish annotation guidelines for IE in German-language OJAs. We also plan to publish annotated data if possible regarding data protection. We then plan to build NLP models that extract these concepts from a large corpus to eventually analyze the results in the context of labor market research.

1. Bibliographical References

Church, K. W. and Kordoni, V. (2022). Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2):249–269.

Deist, F. D. L. and Winterton, J. (2005). What is competence? *Human Resource Development International*, 8(1):27–46.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Engstrom, C. L., Petre, J. T., and Petre, E. A. (2017). Rhetorical analysis of fast-growth businesses' job advertisements: Implications for job search. *Business and Professional Communication Quarterly*, 80(3):336–364.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Zhang, M., Jensen, K. N., Sonniks, S. D., and Plank, B. (2022). Skillspan: Hard and soft skill extraction from english job postings.

Putin's world through a corpus of his speeches

Natalia Levshina
Max Planck Institute for Psycholinguistics
natalia.levshina@mpi.nl

This paper presents a corpus of Vladimir Putin's speeches and public addresses from 2012 until now. The corpus, which is based on the official transcripts, contains information about the event and the speakers. Its use is illustrated by a case study of proper names describing countries and important international organizations in Putin's speech. First, a frequency analysis of the proper names is performed, which reveals the changes in Russia's international relationships over the years. Second, a RuBERT-based sentiment analysis is discussed, which shows which of the names are used in positive or negative contexts. All this allows us to identify Putin's tactics and strategies in building the "multipolar" world.

In the view of Russia's "special military operation" in Ukraine and fossil fuel wars, it is more important to analyze Russian propaganda than ever. The main Russian propagandist is Vladimir Putin himself, whose opinions are transmitted (and often amplified) at all levels of the "power vertical" in Russia. Having a corpus of Putin's speeches can help political scientists to investigate the nature and dynamics of Russian propaganda. In this paper I present such a corpus, which has been scraped from the website kremlin.ru. At the moment, it includes transcripts from 2012, when Putin became the President of Russia after his placeholder Dmitrij Medvedev, until June 2022. I use the corpus for a case study of names of different countries and international organizations. First, I present a frequency analysis of the proper names. In the second part, I apply sentiment analysis based on a BERT language model to analyze the sentiments of the contexts in which these proper names were mentioned. The findings help to expose some of Putin's propaganda tactics and can be of interest to politologists, journalists, policy-makers and other stakeholders.

Aspect-based emotion analysis of Hungarian parliamentary speeches

István Üveges[1,2], Veronika Vincze[3], Orsolya Ring[2], Csenge Guba[1]

[1]University of Szeged, Doctoral School in Linguistics
[2]Centre for Social Sciences, Budapest
[3]ELKH-SZTE Research Group on Artificial Intelligence, Szeged
Uveges.Istvan@tk.hu
vinczev@inf.u-szeged.hu
ring.orsolya@tk.hu
csenge.guba@gmail.com

Abstract

Our study is the first to investigate the possibilities of detecting emotions in Hungarian political speeches. Here we introduce (1) a manually annotated emotion data set of 1008 parliamentary speeches, (2) an emotion annotation framework that uses inductive approaches to identify emotions and their aspects in the corpus (3) a the state-of-the art tool for Aspect Based Sentiment Analysis (ABSA) by exploiting a Hungarian BERT model.

1 Introduction

The emotion analysis of political speeches is one of the most challenging tasks in Natural Language Processing. Most existing tools are available for English texts and require adaptation to produce valid results – especially for morphologically rich languages such as Hungarian (Jang and Shin, 2010; Mladenović et al., 2016). With the rise in significance of online sources for political texts a need emerged for a new, computer-automated method for large-scale analysis which can provide modern political science with the incredibly beneficial empirical data that comes from emotion analysis. While sentiment analysis usually relies on the labels positive/negative/neutral, emotion analysis is based on a more sophisticated and complex categorisation that allows for a better understanding of the underlying emotional content (Liu et al., 2010). Politicians use a mix of emotions to influence their audience, hence the emotional analysis of political speeches becomes more challenging. In this way, it might not be sufficient to rely on a basic sentence-level emotion analysis method, rather, a more fine-grained and aspect-based analysis could be applied here. With aspect-based classification, emotions in the text can be more accurately identified as we are able to provide the linkage of emotion to their specific objects (Liu et al., 2010; Gao et al., 2019; Song et al., 2020). By the task of emotion analysis we mean emotion detection which refers to both the task of detecting if a text conveys any type of emotion or not and the task of classifying existing emotions into a defined set. We examine not only the content of the emotion but also the aspect of the given emotion in relation to which emotion is expressed (Liu et al., 2012a; Gold et al., 2018).

There have not been any recent Hungarian sentiment or emotion dataset consisting of political texts, we are aware of only two publicly available sentiment corpora. The OpinHuBank corpus (Miháltz, 2013) is a freely available sentiment corpus for research and development purposes. The other Hungarian sentiment corpus consists of Hungarian opinion texts written about different types of products (Szabó et al., 2016). Recognizing this research gap, we have built an emotion dataset at the aspect level from Hungarian parliamentary speeches for use in different machine learning approaches.

The main contributions of our paper are summarised as follows:

1. We design a new emotion annotation framework that uses inductive approaches to identify emotions in the corpus and can be aggregated into Plutchik emotion categories (Plutchik, 1982).

2. We present an emotion corpus of Hungarian parliamentary speeches, manually annotated at the aspect level with 756,672 tokens.

3. We perform a huBert-based (Nemeskey, 2020) machine learning experiment for aspect-based classification of annotated emotion categories.

2 Related work

2.1 Emotion analysis

In reviewing the main branches in the literature, it is important to distinguish between sentiment analysis in the strict sense and the much broader emotion analysis: the latter provides a much more

diversified interpretation of the emotional contents of texts (Marcus, 2000).

While the former interprets the concept of a sentiment value only in terms of being positive, negative, or neutral, the latter focuses on the recognition of specific emotions (e.g. anger, fear, joy, etc.). While the two terms are often used synonymously, emotion analysis is more fine-grained with regard to the applied category system.

Probably the most well-known emotion category systems are those of Ekman (Ekman and Wallace V. Friesen, 1982) and Plutchik (Plutchik, 1964, 1980, 1982). Ekman, who studied the cross-culturally uniform nature of human facial expressions, has set up 6 distinct categories for classifying emotions: anger, disgust, fear, happiness, sadness and surprise. Plutchik also developed a somewhat contrary evolutionary psychological theory of emotions, and defined 8 categories to classify them: anticipation, - surprise, joy - sadness, acceptance - disgust and anger - fear.

In psychological literature the concept of emotion raises a number of questions and there is no formal criteria for what is and what is not an emotion (Chapman and Nakamura, 1998; Cabanac, 2002; Griffiths, 2008). The empirical analysis of emotions has uncovered a complexity of concepts (Lakoff and Kovecses, 1987), the entanglement of meanings with the specifics of local culture (Wierzbicka, 1999), and also a lack of exact equivalents of special emotional expressions in different languages (Russell, 2003). Emotions are considered as aspects of complex, interactional systems of an organism which means that there are various relationships between them (Plutchik, 1982). These properties are difficult to specify and analyse with text mining methods.

The other fundamental question is what carries emotion at the content level. There have been a number of previous research (Feng et al., 2013; Loukachevitch and Levchik, 2016) that constructed sentiment and emotion lexicons with connotative sentiment value rather than explicit sentiments exclusively. For instance, *award* and *promotion* have positive connotations and *unemployment* and *terrorism* have negative ones. As Loukachevitch and Levchik state, "*non-opinionated words with connotations usually convey information about negative or positive phenomena (facts) in social life*" (Loukachevitch and Levchik, 2016), making an automatic analysis of these connotative semantic contents challenging.

2.2 Emotions in political communication

Political communication encourages political action by eliciting emotional impact and propagating different ideas. As a result of the technical and social changes of the past decades, the number of participants in the communication and the available channels have increased, which had an impact on the nature and intensity of political communication. Political actors respond to the appreciation of their role by professionalising their communication. Political speeches are well-designed actions with a goal of not only informing but persuading the audience. An important area for political communication is the parliament, where elected representatives discuss submitted bills and other matters of national importance. During parliamentary debates, various topics arise, arguments and counter-arguments collide and through them a political agenda is formed, which then thematises public debates (Bene and Nábelek, 2019). Research on the expression of emotion in political communication has been increasingly emphasised in recent years, both in international and Hungarian social science research (Szabó, 2020; Crigler and Just, 2012; Wagner and Morisi, 2019; Settle, 2020; Richards, 2004; Haselmayer and Jenny, 2017). These studies primarily analyse the speeches of politicians in the media and on social media (Aparicio et al., 2021; Wang et al., 2021; Rufai and Bunce, 2020) but the analysis of the emotional charge of political and especially parliamentary speeches and their aspects with NLP tools is a novel idea (Gold et al., 2018; Jafarian et al., 2021), especially for Hungarian.

2.3 Aspect Based Sentiment Analysis

Sentiment and emotion analysis can be used to investigate the general polarity of a text or sentence as well as the emotions it conveys, but it is often insufficient to obtain practically useful data. The main reason behind this is that sentences often do not express just a single sentiment or emotion but many of them. For example, there are frequent cases where two clauses referring to two properties of an object have completely different sentiment values. Such cases are by default difficult to deal with in classical sentiment analysis procedures, which are not able to detect if negative and positive sentiment values do not refer to the same entity or the same aspect of the same entity.

A potential solution can be the use of Aspect

Based Sentiment Analysis (ABSA). In the most general approach, ABSA systems are designed to identify aspects of the text in relation to which sentiment is expressed and to determine the sentiment value for each aspect. An aspect can be any kind of entity in the real world such as personal names, companies (traditional targets of Named Entity Recognition – NER) or personal pronouns referring to them, any kind of properties (of a product, for instance), etc. Therefore, by using ABSA solutions, the main goal is not just to identify a sentiment value for a textual unit but also to find the appropriate entities to which the given sentiment is connected (Liu et al., 2012b; Zhang et al., 2022).

3 The corpus

For annotation, we selected Hungarian parliamentary pre-agenda speeches delivered by Members of Parliament from the period of 2014-2018. Transcripts of these speeches are publicly available at the official website of the Hungarian National Assembly (parlament.hu), so we have scraped their texts automatically. During this period, 1008 speeches were delivered, all of which are included in our corpus. Topics of these speeches cover various themes like health, education, and social issues.

Pre-agenda speeches are presented in the Hungarian legislature at the beginning of each parliamentary session. They are generally held by frontbencher MPs and members of the government. Their topic can be freely selected by the presenter. Generally, they are followed by a short debate.

Although the texts of our corpus are spoken language data, their style is official, it differs from the spoken language corpora available in Hungarian, which contain spontaneous speech and/or have an informal style (Vincze et al., 2021; Szabó et al., 2021). They contain many addressing terms and thanks (*Dear House*, *Thank you for giving me the floor*), they use almost exclusively formal speech, however, the transcripts do not contain the hesitations, small breaks, false beginnings that are typical of live speech.

3.1 The category system

The system used in the present study is somewhat different from but in many ways builds on the earlier systems mentioned before. We divided Plutchik's categories into further sub-categories, named "emotion topics". The relation between Plutchik's system and the one used in the current corpus annotation can be seen in Table 1.

The reason for subcategorising Plutchik's system was that our previous experience has shown that annotators were able to make decisions about the classification of textual units more easily and with better inter-annotator agreement when each emotion category was broken down into "components". More precisely, when developing the annotation principles, the concept of emotion topics was defined as being more general then "basic" emotions (cf. Plutchik's and Ekman's system) but narrower than just simple topic tagging, being somewhere between the two (Ring et al., 2022). Here, topic tagging is understood as the simple designation of the term that carries the most important message of a given textual unit.

Although the emotion topics themselves do not necessarily refer to emotions literally (see the example below) the events they describe certainly evoke an emotion.

> A man just fell down at the bus stop with a stroke, but people nearby could save his life by giving him first aid.

Here, the first part of the sentence describes an accident but the second part describes an act of help, which is related to the "traditional" concept of emotion *trust*.

Emotions in Plutchik's categorisation may include more than one emotion topic in our classification. For instance, sadness covers the following emotion topics in our subcategorisation: suffering, sorrow and catastrophe/accident, shown on the sentences below.

> Sok százezer roma ember egészen kilátástalan helyzetben van, elesett, kiszolgáltatott embertársaink. 'Many hundreds of thousands of Roma people are in a totally desperate situation, our deprived and vulnerable fellows.'
>
> A múzeum dolgozói feladták a reményt, hogy az ellopott festmény visszakerül hozzájuk. 'The employees of the museum gave up all hope that they can get back the stolen painting.'
>
> 8 ember meghalt hétfőn, amikor egy 48 utast szállító hajó elsüllyedt. '8 people died on Monday when a ship carrying 48 passengers sank.'

Although all of the above examples induce sadness in the reader, the reason for sadness is always somewhat different. We argue that in a political context it is important to emphasise whether a situation originated from the actions or lack of actions of someone (in this context, mostly the governing parties) or not, as MPs would often like to convince the audience that those behind a negative event should take responsibility. In the first example, the speaker may feel sorry for the poverty of Roma people but also wants to blame the governing parties for not caring about them. In the second example we hear about a crime (i.e. a stolen painting), but there is no indication that someone is to be blamed for it. Finally, in the third example, a catastrophe is reported, which has definitely nothing to do with any political actions or human interventions, however, the speaker and the reader may still feel sad about it.

Table 1 shows the categorization scheme applied in our study.

3.2 Annotation process

Five linguists carried out the annotation under the supervision of a master annotator. They marked the emotion topics of the text at the level of clauses. In addition, they also marked the keywords that evoked the emotion topic in question and the arguments of the emotion.

As can be seen in Table 3, there are large differences in the frequency of the categories. The top four categories are improvement, suffering, conflict and contempt, which cover over 90% of the data. This is probably due to the nature of the data, i.e. in parliamentary speeches, MPs usually discuss issues what need improvement and are currently in a bad state (deterioration) or have been successfully improved lately (improvement). On the other hand, it can be observed that a significant part of the speeches contains lots of instances of conflict and contempt, which can be explained by the characteristics of political debates: members of politically opposed parties may often speak critically of other parties' members and their activities.

In order to assure the quality of the corpus, about 25% of the files were double-annotated, and inter-annotator agreement rates were calculated for these files. The aggregated value of agreement is listed in Table 2 (in terms of Cohen's Kappa). A higher agreement rate could be achieved for frequent categories, but the general agreement was 0.41 Cohen's Kappa, which means moderate agreement. However this value is far from perfect, it is worth mentioning that emotion annotation task is usually challenging for the human annotators, and a number around 0.4 - 0.5 in terms of Cohen's κ is quite average in emotion annotation tasks (Chen et al., 2021).

4 Machine learning experiments

In this section, we present our machine learning experiments on identifying emotions in Hungarian parliamentary speeches, with regard to aspects of emotion.

4.1 Data used

Since the corpus was still under construction at the time of submission, it was not possible to use the full dataset in our investigations. The used dataset we selected from our corpus contained 260,789 tokens and 8820 sentences from the 618 speeches out of the total 1008.

Moreover, we further needed to filter the collected data for the currect experiment (for the exact reasons, see 4.3). The basic statistical data of the used dataset can be seen in Table 3.

4.2 Methods

To give a solution for aspect based sentiment analysis of the given corpus, so to find not only the emotion expressed by a sentence but also the real world object it refers to, we used a pytorch-based implementation. This project (described in detail in (Tang et al., 2016)) was originally created for solving SemEval 2014 - aspect-based sentiment analysis (4.) (Pontiki et al., 2014). The repository contains both non-BERT-based and BERT-based solutions. From this latter group, the BERT-SPC, also originally was developed for solving SemEval 2014, subtask SB2, in which the sentiment value for the aspect had to be identified.

It should be noted that for SemEval SB2, there were originally only 4 categories to choose from, divided into 'positive', 'negative', 'conflict' and 'neutral' labels, with the 'conflict' label used when both positive and negative emotions were expressed by a given aspect term. Given that 12 category labels are available in the present case, the preliminary expectation was that the task would be performed with lower efficiency.

Since the original ABSA-PyTorch implementation was designed for English, the task required

Category Nr.	Related concepts	Emotion topic	In Plutchik's system	Sentiment
1	fear, threat, intimidation, dread, anxiety	**Fear**	Fear	Negative
2	suffering, deprivation, misery, poverty, torment, failure, negative change	**Suffering**	Sadness	
7	sorrow, despair, hopelessness, melancholy	**Sorrow**		
10	misfortune, catastrophe	**Misfortune**		
3	crime, terror, assassination, persecution, cruelty, organized crime, vandalism, intentional harm, violence	**Crime**	Anger	
9	anger, fury, hatred	**Anger**		
5	conflict, confusion, conflict of interest, revenge, punishment	**Conflict**	Disgust	
6	contempt, mockery	**Contempt**		
4	improvement, relief, development, success, positive change	**Improvement**	Success	Positive
8	joy, enjoyment, merriment, serenity, love, acceptance, tolerance	**Joy**	Joy	
11	assistance, rescue, relief, healing, care, deliverance	**Assistance**	Trust	
12	justice, investigation	**Justice**		

Table 1: Emotional topics and their equivalents in Plutchik's category system, alongside the relevant sentiment value.

	A1	A2	A3	A4	A5
A1		0.328	0.323	0.498	0.572
A2	0.328		0.466	0.451	0.328
A3	0.323	0.466		0.288	0.272
A4	0.498	0.451	0.288		0.607
A5	0.572	0.498	0.323	0.607	
Average	0.43	0.415	0.342	0.448	0.402

Table 2: Inter-annotator agreement (Cohen κ) between annotators (A1 - A5). Average: the average agreement of the annotator with all the others.

Category	Token	Sentence
Improvement	87854	3077
Suffering	61364	1988
Conflict	60206	1884
Contempt	28961	1092
Crime	7752	260
Joy	4985	191
Justice	4603	138
Fear	1855	63
Assistance	1693	62
Sorrow	952	36
Misfortune	321	18
Anger	243	11

Table 3: Basic statistics of the dataset, differentiated by emotion topics

a solution using the `huBERT` (Nemeskey, 2020) Hungarian model[1]. This was carried out with the use of a github repository[2] created for preliminary investigations about the possibilities of an ABSA task in Hungarian.

4.3 Corpus preparation

The annotation was carried out with the Tagtog[3] online annotation tool, which gives a `.json` file as a result. At first, these files had to be processed to get the labelled aspects and emotion categories for each original `.txt` file. Since the text was presented in an unsegmented way, the next step was sentence segmentation. For this purpose, we used the language model[4] developed for the `huspacy`(Orosz et al., 2022) natural language processing toolkit for Hungarian.

Since manual annotation inevitably comes with errors, a smaller set of sentences was subjected to a manual check in order to filter out as much noise or typical errors from the training data as possible. This kind of review revealed that a typical problem was that annotators marked too many tokens as aspects. An aspect term should always refer to an entity of the real world, or to a property of such an entity. Therefore adverbs, like *tegnap* ('yesterday') or *talán* ('maybe') on their own can hardly

[1] https://huggingface.co/SZTAKI-HLT/hubert-base-cc

[2] https://github.com/PasztorAkos/ABSA_Pytorch_HUN_Sent_An

[3] gttps://tagtog.net/

[4] https://huggingface.co/huspacy/hu_core_news_lg

be considered aspects. To minimize the number of misclassified aspect terms, we used a simple heuristical approach: all words that had been annotated as an aspect but did not contain at least one token tagged as a noun, pronoun or proper name were deleted from the dataset. For POS-tagging we used `huspacy` and the `hu_core_news_lg` language model again.

Our data was somewhat special compared to usual datasets, since every sentence could be present with more than one aspect terms with a emotion value for each. Such sentences were added to the dataset as sets (so the same sentences with different aspect term were grouped together). After that, train-test selection was performed on these sets; we paid attention to the fact that a given sentence might contain several aspects with the same ET, which can cause redundancy in the data that should be handled. We carried out train-test selection in a way that one sentence (with all of its aspects) occurred either in the training set or in the test set. This was necessary because the emotion value was often the same for all aspects in a sentence, and the presence of such a sentence in both the train- and test-set would have significantly distorted the final classification results.

5 Results and discussion

By default, the model training runs for up to 20 epochs before it is finished automatically. In the case of our particular dataset, training was finished after 6 epochs of learning with an early stop.

Table 4 shows the values of the main metrics measured for our emotion topics during each epoch. By examining the data, it is clear that the highest scores over the 6 epochs were typically achieved during the first few epochs (this trend is particularly striking for F1 scores, where the first epoch was the most successful for 7 out of the 12 emotion topics). These trends suggest that under the current system of categories and with the present amount of training data, additional training does not contribute to better prediction results. Figure 1 illustrates this trend with a weighted average of the metrics for our 12 emotion categories.

To get a better insight into the results, we compared them to the proportion of sentences annotated to each emotion topic in our current corpus. The corresponding values are presented in Table 5 as follows: emotion topic, the number of sentences annotated in the test set to the given emotion topic,

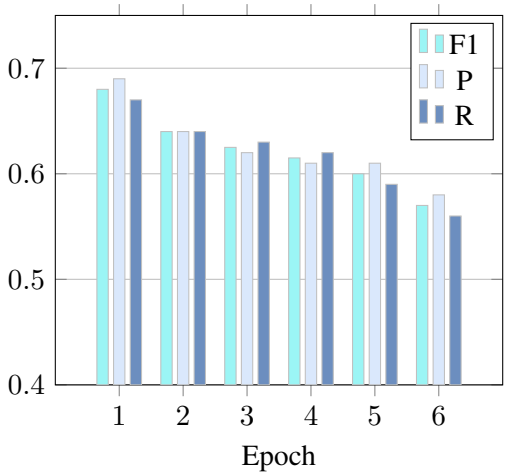

Figure 1: Precision, Recall and F1 per epoch (Average for 12 Emotion Topics)

the proportion of these sentences over the corpus, the average of the F1-values measured over the 6 epochs, the minimum and maximum values of F1 over 6 epochs and the observed standard deviation.

This clearly shows another important feature of the corpus, namely the imbalance in the data. Around 57.46% of the total number of instructive examples are sentences belonging to 2 categories (4: “misfortune”, 2: “suffering”, both of them belonging to the category “sadness” in Plutchik’s scheme), and the addition of two more categories already covers 91.22% of the total data. It is worth mentioning that the data used here do not fully overlap with the base corpus’ statistics, since if a sentence contained more than one aspect, it occurred more than once in the training/test data.

With regard to a potential later expansion of the corpus with sentences belonging to the currently underrepresented categories, we were curious to investigate whether a sufficiently large number of sentences correlates with higher F1 values. Figure 2 illustrates the relationship between the number of items in each category and the average F-scores over 6 epochs. In addition to the data points, the “best-fit line” is also shown with the corresponding confidence intervals.

By calculating the Pearson r correlation coefficient (0.775 with p = 0.003 in the present case), it can be stated with high confidence that there is a strong positive linear relationship between the number of sentences belonging to the given categories and the values of the measured metrics. Therefore it seems that the performance of the model could be improved by increasing the corpus with sen-

ET.	Epochs 1			2			3			4			5			6		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0.39	0.35	0.37	0.33	**0.37**	0.35	**0.63**	0.33	**0.43**	0.52	0.25	0.34	0.25	0.27	0.26	0.57	0.25	0.35
2	**0.73**	0.66	**0.7**	0.68	0.63	0.65	0.6	**0.73**	0.66	0.59	0.71	0.64	0.69	0.53	0.6	0.54	0.67	0.6
3	**0.62**	0.31	**0.41**	0.45	0.32	0.37	0.46	**0.36**	0.4	0.44	0.35	0.39	0.36	0.35	0.35	0.38	0.29	0.33
4	**0.86**	0.8	**0.83**	0.75	0.83	0.79	0.72	**0.86**	0.78	0.72	0.82	0.77	0.76	0.72	0.74	0.75	0.69	0.72
5	0.5	0.7	**0.58**	0.53	0.63	**0.58**	**0.54**	0.49	0.52	0.51	0.51	0.51	0.43	**0.71**	0.54	0.49	0.41	0.45
6	0.5	**0.5**	**0.5**	**0.63**	0.26	0.37	0.58	0.25	0.35	0.57	0.24	0.34	0.45	0.27	0.33	0.34	0.47	0.39
7	0.3	**0.2**	**0.24**	0.08	0.17	0.11	0.27	0.13	0.18	0.2	0.17	0.18	0.21	0.1	0.14	**0.67**	0.07	0.12
8	0.6	0.38	0.47	0.4	**0.54**	0.46	0.37	0.29	0.33	0.65	0.32	0.43	0.58	0.44	**0.5**	**0.67**	0.33	0.44
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0.36	0.29	0.32	0.16	**0.5**	0.25	**1**	0.07	0.13	0.75	0.21	**0.33**	1	0.14	0.25	0	0	0
11	**1**	**0.35**	**0.52**	0.35	0.33	0.34	**1**	0.29	0.45	0.31	0.33	0.32	0.94	0.33	0.49	0.59	0.25	0.36
12	**0.88**	0.31	0.45	0.41	0.54	0.47	0.62	0.51	0.56	0.51	**0.64**	**0.57**	0.53	0.47	0.5	0.61	0.45	0.52

Table 4: The main metrics measured in the study, broken down by epoch (ET.: Emotion Topic, based on Table 1, **emphasis**: maximum value of the given metric for the particular ET. during the 6 epochs of running)

ET.	Nr.	%	F1	Min	Max	σ
4	2564	34.94%	0.77	0.72	0.83	0.039
2	1653	22.52%	0.64	0.6	0.7	0.038
5	1570	21.39%	0.53	0.45	0.58	0.049
6	908	12.37%	0.38	0.33	0.5	0.063
3	216	2.94%	0.38	0.33	0.41	0.031
8	158	2.15%	0.44	0.33	0.5	0.058
12	114	1.55%	0.51	0.45	0.57	0.048
1	52	0.71%	0.35	0.26	0.43	0.055
11	51	0.69%	0.41	0.32	0.52	0.084
7	30	0.41%	0.16	0.11	0.24	0.048
10	14	0.19%	0.21	0	0.33	0.127
9	9	0.12%	0.00	0	0	0

Table 5: The distribution of emotion topics in the training corpus, and the measured F1-scores during training. (ET: emotion topic, Nr.: Number of sentences in test corpus, %: percentage

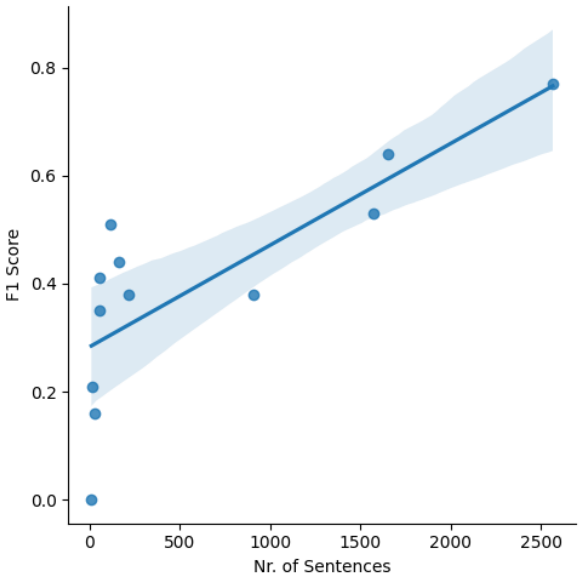

Figure 2: Correlation between emotion topics and measured average F1

tences belonging to the currently underrepresented emotion topics.

Putting the results into context is made difficult by the fact that the category system used for annotation is unique and has only been used in this project yet. With regard to emotion categories, Kis György et al. (2022) presents a similar experiment using an 8-class category system (7 emotions from (Plutchik, 1964), and a neutral category). Their implementation used context-dependent word embeddings generated from the hidden layer of the huBERT tokenizer, and then used these vectors to predict the emotion categories using a cross-validated logistic regression algorithm. The amount of training data they used (number of sentences) and the measured prediction results are presented in Table 6.

Comparing their results with the current experiment shows that despite the fact that the current corpus' annotation operates with more categories and aspect-based sentiment classification is considered an even more complex task than the classification of emotions only, the results seem to be comparable where sufficient training data was available.

Although in many cases the amount of data does not allow us to draw general conclusions, it seems that after a rebalancing of the corpus, our dataset could be suitable for training models with even better efficiency than the current one.

Emotion	Anger	Disgust	Fear	Joy
Sentence Nr.	934	3202	360	156
Weighted F1	0.61	0.61	0.25	0.32
Emotion	Neutral	Sadness	Trust	Anticipation
Sentence Nr.	1707	2645	1419	4214
Weighted F1	0.51	0.61	0.66	0.71

Table 6: (Kis György et al., 2022)'s results in Plutchik's 7 emotion categories + sentences classified as neutral.

6 Conclusions and future work

In this paper, we have presented a novel emotion annotation framework that uses inductive approaches to identify emotions and their aspects in the corpus and implemented ABSA-PyTorch and Hungarian BERT-model to classify the emotion of a given aspect.

We have described in detail the pre-processing of the data generated by the project, as well as the conversion steps required to use ABSA-PyTorch and the problems encountered, as well as their solutions. The results of the machine learning experiment performed were evaluated using traditional precision, recall, and F-value metrics, and a correlation was found between the results and the unbalanced nature of the training data.

Our results show that despite the fact that the corpus uses multiple categories the results move at a comparable level for those categories where a sufficient amount of training data was available for the model. Once the corpus is balanced, it may be suitable for training more efficient models.

In the future, we would like to work on a balanced corpus with text augmentation and investigate the possibilities of extending our annotation framework for other types of texts.

Acknowledgments

The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program.

The research was supported by the European Union's Horizon 2020 research & innovation programme under Grant Agreement no. 951832

References

Joao Tiago Aparicio, João Salema de Sequeira, and Carlos J Costa. 2021. Emotion analysis of portuguese political parties communication over the covid-19 pandemic. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Márton Bene and Fruzsina Nábelek. 2019. A politikai kommunikáció története a külföldi szakirodalomban . In Balázs Kiss, editor, *A szavakon túl. Politikai kommunikáció Magyarországon, 1990-2015*, pages 11–29. L'Harmattan Kiadó.

Michel Cabanac. 2002. What is emotion? *Behavioural processes*, 60(2):69–83.

CR Chapman and Y Nakamura. 1998. A bottom up view of emotion. In *ASSC Seminar http://server. phil. vt. edu/assc/watt/chapman1. html*.

Xin Chen, Zhen Hai, Deyu Li, Suge Wang, and Dian Wang. 2021. Jointly identifying rhetoric and implicit emotions via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1429–1434, Online. Association for Computational Linguistics.

Ann N Crigler and Marion R Just. 2012. Measuring affect, emotion and mood in political communication. *The Sage handbook of political communication*, pages 211–224.

Paul Ekman and Phoebe Ellsworth Wallace V. Friesen. 1982. What emotion categories or dimensions can observers judge from facial behavior In Paul Ekman, editor, *Emotion in the Human Face. 2nd ed.*, page 39–55. Cambridge University Press, Cambridge, UK.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with bert. *Ieee Access*, 7:154290–154299.

Darina Gold, Marie Bexte, and Torsten Zesch. 2018. Corpus of aspect-based sentiment in political debates.

Paul E Griffiths. 2008. What emotions really are. In *What Emotions Really Are*. University of Chicago Press.

Martin Haselmayer and Marcelo Jenny. 2017. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6):2623–2646.

Hamoon Jafarian, Amir Hossein Taghavi, Alireza Javaheri, and Reza Rawassizadeh. 2021. Exploiting bert to improve aspect-based sentiment analysis performance on persian language. In *2021 7th International Conference on Web Research (ICWR)*, pages 5–8. IEEE.

Hayeon Jang and Hyopil Shin. 2010. Language-specific sentiment analysis in morphologically rich languages. In *Coling 2010: Posters*, pages 498–506.

Márk Kis György, Orsolya Ring, and Miklós Sebök. 2022. A novel cost-efficient use of bert embeddings in 8-way emotion classification on a hungarian media corpus. *SocArXiv*.

George Lakoff and Zoltan Kovecses. 1987. The cognitive model of anger inherent. *American English-In Cultural Models_in_Language and Thought-Dorothy Holland and Naomi Quinn. eds*, pages 195–221.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Kang Liu, Liheng Xu, and Jun Zhao. 2012a. Opinion target extraction using word-based translation model. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1346–1356.

Kang Liu, Liheng Xu, and Jun Zhao. 2012b. Opinion target extraction using word-based translation model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1346–1356, Jeju Island, Korea. Association for Computational Linguistics.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general russian sentiment lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1171–1176.

George E Marcus. 2000. Emotions in politics. *Annual review of political science*, 3(1):221–250.

Márton Miháltz. 2013. Opinhubank: szabadon hozz áférhetö annotált korpusz magyar nyelvü véleményelemzéshez [opinhubank: a freely available annotated corpus for hungarian opinion mining]. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*.

Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620.

Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó, and Richárd Farkas. 2022. Huspacy: an industrial-strength hungarian natural language processing toolkit. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*.

Robert Plutchik. 1964. The emotions: Facts, theories and a new model. *American Journal of Psychology*, 77:518.

Robert Plutchik. 1980. Emotion, a psychoevolutionary synthesis. New York (etc.). Harper and Row.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information/sur les sciences sociales*, 21 (4-5):529–553.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Barry Richards. 2004. The emotional deficit in political communication. *Political Communication,*, 21(3):339–352.

Orsolya Ring, Martina Katalin Szabó, Bendegúz Váradi, Csenge Guba, and István Üveges. 2022. Approaches to Sentiment Analysis of Hungarian Political News at Sentence Level with Dictionary-based Method and with Machine Learning. Under review.

Sohaib R Rufai and Catey Bunce. 2020. World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis. *Journal of public health*, 42(3):510–516.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

J Settle. 2020. Moving beyond sentiment analysis: Social media and emotions in political communication. *The Oxford Handbook of Networked Communication. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190460518.013*, 20.

Youwei Song, Jiahai Wang, Zhiwei Liang, Zhiyue Liu, and Tao Jiang. 2020. Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference. *arXiv preprint arXiv:2002.04815*.

Gabriella Szabó. 2020. Emotional Communication and Participation in Politics. *Intersections. East European Journal of Society and Politics*, 6(2).

Martina Katalin Szabó, Veronika Vincze, Orsolya Ring, István Üveges, Eszter Vit, Flóra Samu, Attila Gulyás, Júlia Galántai, Zsuzsanna Szvetelszky, Eliza Hajnalka Bodor-Eranus, and Károly Takács. 2021. StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. Szegedi Tudományegyetem.

Martina Katalin Szabó, Veronika Vincze, Katalin Ilona Simkó, Viktor Varga, and Viktor Hangya. 2016. A hungarian sentiment corpus manually annotated at aspect level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Veronika Vincze, István Üveges, Martina Katalin Szabó, and Károly Takács. 2021. A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. Szegedi Tudományegyetem.

Markus Wagner and Davide Morisi. 2019. Anxiety, fear, and political decision making. In *Oxford research encyclopedia of politics*. Oxford University Press.

Yuming Wang, Stephen M Croucher, and Erika Pearson. 2021. National leaders' usage of twitter in response to covid-19: A sentiment analysis. *Frontiers in Communication*, 6:732399.

Anna Wierzbicka. 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge university press.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. arXiv:2203.01054v1.

PANEL: Theory-Driven Modelling of Complex Socio-Psychological Constructs in Text

Veronika Batzdorfer
Digital Society Observatory
GESIS
www

Valentin Gold
Center of Methods in Social Sciences
University of Göttingen
www

Camille Roth
Centre Marc Bloch
CNRS
www

Henning Wachsmuth
Computational Social Science
University of Paderborn
www

Abstract

The main goal of this workshop is to bridge the gap between researchers from political/social science and NLP/computer science and bring together researchers and ideas from the different communities, to foster collaboration and catalyze further interdisciplinary research efforts. Towards this end, we invited speakers from the different communities. The focus of the panel will be on theory-driven modelling of complex political or socio-psychological constructs in text, such as populism, polarisation or political cynicism. In particular, we want to discuss challenges for modelling such multifaceted concepts from a theoretical as well as from a machine learning point of view, and how such models can be evaluated in a meaningful way.

We hope that this will foster discussions and allow us to reflect on our different research practices, methods and tools, and will help to improve the communication between our fields.

The challenges and opportunities of defining what is political in social media

David Jurgens
University of Michigan, USA

Studies of political speech on social media typically require defining who or what is political in order to analyze political behavior. However, these definitional choices have serious implications for what can be learned, as well as who or what is left out of the analysis. In this talk, I will describe two studies studying political text in social media and how design choices in what is political can lead to different interpretations in the results. The first study asks whether political users are more toxic to each other in Reddit in order to understand whether cross-partisan discourse is driving higher incivility. The second study examines the framing in speech about immigrants in Twitter to understand how political leaning influences different associations with immigrants. Across both studies, I show that political affiliation is linked to specific behaviors—toxicity and politics do frequently co-occur and some political leaning are more likely to use certain framing about immigrants than another—but that different choices in how we define political users and speech are necessary to provide more precise and nuanced insights of political behavior. Across both studies, I am to highlight the challenges and opportunities of computationally modeling political behavior in social media.

Index of Authors