International Journal of Speech Technology
https://doi.org/10.1007/s10772-024-10142-4

Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election

Nils Constantin Hellwig[1] · **Jakob Fehle**[1] · **Markus Bink**[1] · **Thomas Schmidt**[1] · **Christian Wolff**[1]

Received: 10 August 2024 / Accepted: 1 September 2024

Abstract
We present a study in the context of computational social science that explores the topics debated in the context of the 2021 German Federal Election by using the topic modeling technique BERTopic. The corpus consists of German language tweets posted by political party accounts of the major German parties, as well as tweets by the general public mentioning the party accounts. We examined the textual content of the tweets but also included the text in images that were posted into the analysis by extracting the text using optical character recognition (OCR). Our results show that the most frequently discussed topics are party-oriented policies (including call-to-action content), climate policy and financial policy, with these topics being discussed in tweets by both, the political party accounts and tweets by accounts mentioning them. In addition, we observed that some topics were discussed consistently throughout the year, such as the COVID-19 pandemic, climate policy or digitization, while other topics, such as the return to power of the Taliban in Afghanistan or Israel were debated to a greater extent at limited time frames during the election year.

Keywords BERTopic · Topic modeling · Sentiment analysis · Natural language processing

1 Introduction

The election to the 20th German Bundestag in 2021 can be considered a historic election. After 16 years, Angela Merkel, who is a member of the party Christian Democratic Union (CDU), did not run again for the office of chancellor. The sister parties CDU and Christian Social Union (CSU), who constitute a parliamentary group in the Bundestag, received fewer votes compared to the Social Democratic Party (SPD), of which Olaf Scholz is a member, who was subsequently elected as new chancellor. After the election, the SPD formed a government with the Green Party (Bündnis 90/Die Grünen) and the Free Democratic Party (Liberals, FDP).

Polls of eligible voters showed that topics such as social security, climate, economy, labor and the management of the ongoing COVID-19 pandemic were decisive for them in deciding which party to vote for in this election.[1] Given the continued presence of state-level pandemic restrictions during the election year, campaigning on social networks played an important role. One of the most popular social networks used by many politicians to disseminate political content is Twitter.[2] Twitter enables users around the world to express themselves and interact with others through short messages called "tweets". In the election year 2021, there was still a limitation so that a tweet could have a maximum of 280

✉ Nils Constantin Hellwig
nils-constantin.hellwig@ur.de

✉ Jakob Fehle
jakob.fehle@ur.de

✉ Markus Bink
markus.bink@student.ur.de

✉ Thomas Schmidt
thomas.schmidt@ur.de

✉ Christian Wolff
christian.wolff@ur.de

[1] Media Informatics Group, University of Regensburg, Regensburg, Germany

[1] Infratest dimap on behalf of ARD: https://www.tagesschau.de/wahl/archiv/2021-09-26-BT-DE/umfrage-wahlentscheidend.shtml.

[2] As of April 2024, Twitter has been rebranded as X. However, we will use the name Twitter in this paper since the data was acquired before the rebranding and is still a common reference for the platform.

Published online: 29 October 2024

characters. This limitation encourages users to express themselves briefly and clearly. Twitter also supports the sharing of images, GIFs, and videos alongside text messages.

Due to the large amount of publicly available data, Twitter is often used in research to gauge political sentiment by analyzing both: the tweets of politicians and party accounts, as well as tweets by the general public (Budiharto & Meiliana, 2018; Costa et al., 2021; Hellwig et al., 2023; Schmidt et al., 2022). Furthermore, it is possible to gain an understanding of what political topics are being discussed at a given time. An established method for this task in natural language processing (NLP) to identify topics in text, is to apply topic modeling. Topic modeling is a statistical technique to identify latent topics or themes within a collection of documents, such as tweets (Hong & Davison, 2010). As an illustration, terms such as "covid", "vaccine" and "lockdown" might be grouped together to capture the thematic emphasis on COVID-19 pandemic-related matters. Topic modeling have been employed extensively over the past decade to uncover latent structures within a corpus, enabling a broad spectrum of applications (cf. Boyd-Graber et al., 2014). In this paper, we apply topic modeling to explore topics that were discussed on Twitter during the 2021 federal election campaign. We examined two different type of German language tweets: (1) the tweets of a fixed set of accounts of politicians and political parties and (2) tweets from users who mentioned those accounts with the @-sign to gain further insights into the general public's perspective. In addition to the text of the tweets, the images included in these were analyzed as well. The research questions are as follows:

- What are the major topics considering the tweets of the entire election year 2021?
- How do the topics addressed in the tweets of the political actors differ from the topics addressed in the tweets of general users who mention these accounts?
- How do the number of tweets for each identified topic change over the course of the election year?
- How does the sentiment of tweets assigned to specific topics evolve throughout the course of an election year?
- Which topics are addressed by both political parties and the general public?

The main contributions to the research area are as follows:

- The extension of a pre-existing corpus of 58,864 tweets by 11,817 images posted by 89 Twitter accounts of the major German political parties for the election year 2021.
- The extension of a pre-existing corpus of 707,241 tweets by 26,371 images that were included in tweets that mentioned (using @-sign) the 89 Twitter accounts of the major German parties for the election year 2021.
- Application of topic modeling via BERTopic as proposed by Grootendorst (2022) to identify topics in both corpora.
- Analysis of topics in context of sentiment analysis results throughout the year.

Resources related to this work such as programming code, visualizations and corpus information are publicly available on GitHub.[3]

2 Related work

In this section, we first elaborate on the current state-of-the-art of topic modeling, and then summarizes the research in the context of topic modeling on political Twitter.

2.1 Methods for topic modeling

Topic modeling has proven to be an effective, unsupervised method to identify common patterns and relationships in textual data (cf. Jelodar et al., 2019). Many previously introduced machine learning approaches for topic modeling are based on Latent Dirichlet Allocation (LDA) (Jelodar et al., 2019). LDA is a generative probabilistic model first introduced by Blei et al. (2001). Documents are treated as combinations of different underlying topics. Each topic is defined by a collection of words that are likely to appear together (Jelodar et al., 2019). By examining the words with the highest probabilities in each topic, LDA helps one to understand the main themes or subjects discussed in the documents (Jelodar et al., 2019). Topic models based on LDA have been applied in various fields such as medical science (Paul & Dredze, 2011), software engineering (Asuncion et al., 2010), social media analysis (Moßburger et al., 2020; Schmidt et al., 2020b), digital humanities (Schmidt et al., 2020a) and political science (Dahal et al., 2019; Karami et al., 2018; Xue et al., 2020).

A limitation of methods based on LDA is that they overlook the semantic connections between words by utilizing bag-of-words representations (Grootendorst, 2022). By disregarding the contextual information of words within a sentence, the bag-of-words approach may not effectively represent the documents (Grootendorst, 2022). To address this problem, text embedding techniques such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have gained popularity. Unlike most previous models that treat words in isolation, BERT processes words in relation to their surrounding context, enabling a deeper understanding of words and capturing their

[3] GitHub repository with supplements: https://github.com/NilsHellwig/Topic_Modeling_Twitter_German_Federal_Election_2021.

Table 1 Statistics on the entire corpus collected by Schmidt et al. (2022): Tweets by politicians and official party accounts of the respective party

Party	Political orientation	Pre-election	Post election	# Tweets	# Images (with text)	%	# Tokens	Avg. Tweets length
SPD	Center left	Government	Government	11,353	3710 (1880)	19.3	623,572	54.93
CDU/CSU	Center right	Government	Opposition	10,072	4741 (2203)	17.1	512,803	50.91
Die Grünen	Left, ecological	Opposition	Government	9576	3382 (1706)	16.3	537,408	56.12
FDP	Liberalism	Opposition	Government	6610	1802 (912)	3.1	356,789	53.98
AfD	Far right	Opposition	Opposition	11,625	4543 (2887)	7.7	592,828	51.00
Die Linke	Far left	Opposition	Opposition	9628	3781 (2229)	16.4	522,322	54.25
Total	–	–	–	58,864	21,959 (11,817)	100	3,145,722	53.44

Fig. 1 Examples of images posted by political party accounts (see Table 8 for details on example images)

Table 2 Statistics on the entire corpus collected by Hellwig et al. (2023): Tweets that mentioning accounts of the respective parties

Mentioned party	Political orientation	Pre-election	Post election	# Tweets	# Images (with text)	%	# Tokens	Avg. Tweets length
SPD	Center left	Government	Government	228,415	10,387 (8233)	32.3	7,153,549	31.32
CDU/CSU	Center right	Government	Opposition	227,683	11,678 (8916)	32.2	7,097,145	31.17
Die Grünen	Left, ecological	Opposition	Government	73,261	3932 (2946)	10.4	2,408,946	32.88
FDP	Liberalism	Opposition	Government	79,815	3818 (3058)	11.3	2,607,610	32.67
AfD	Far right	Opposition	Opposition	57,572	2548 (1923)	8.1	1,636,144	28.42
Die Linke	Far left	Opposition	Opposition	40,495	1778 (1295)	5.7	1,340,331	33.10
Total	–	–	–	707,241	34,141 (26,371)	100	22,243,725	31.45

semantic relationships (Devlin et al., 2019). One way to use such language models for topic modeling is BERTopic which was introduced by Grootendorst (2022) and gained attention recently. A document is first converted into a dense vector representation using Sentence-BERT (Grootendorst, 2022). The dimensionality of the document embeddings is then reduced with the help of Uniform Manifold Approximation and Projection (UMAP) (Grootendorst, 2022). Finally, clustering algorithms such as k-Means or Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) can be used to identify clusters or topics, respectively (Grootendorst, 2022). Three corpora were considered for evaluation by Grootendorst (2022), including a corpus consisting of 44,253 tweets posted by Donald Trump. When comparing BERTopic with LDA in terms of the common evaluation metrics topic diversity and topic coherence, BERTopic always outperformed LDA in the case of all three corpora, including the Twitter corpus.

2.2 Topic modeling on Twitter for political research

Topic modeling has proven to be an effective approach for extracting information within various political contexts on the Twitter. Previous research has applied topic modeling techniques to identify different topics within a broader political scope. For example, Miller (2019) examined a publicly

Fig. 2 Examples of images posted by users mentioning political party accounts (see Table 9 for details on example images)

available corpus of IRA/Russian Federation associated tweets from July 2014 to September 2017 related to the 2016 US presidential election to identify key topics being discussed in such tweets. Among the 56 topics identified, seven were found to be associated with the campaign of presidential candidate Hillary Clinton, and these were all identified as being mostly negative in terms of their sentiment. In contrast, there were ten topics that referenced presidential candidate Donald Trump, with four of them expressing support. Additionally, topics emerged that involved President Putin and the Russian Federation in general, but no consistently positive or negative sentiment of these topics was discernible. The author's conclusion emphasizes that topic modeling uncovered the heterogeneous and contradictory nature of the corpus, as tweets addressed a wide range of political positions and issues.

Dahal et al. (2019) collected 309,016 geotagged tweets related to climate change from July 1, 2016, to February 28, 2018. They identified topics such as transportation, energy, fossil fuel industry, and international agreement within this context. They showed that certain topics are discussed more frequently at times when they are also of importance in the political debate. For example, the topic of international agreement was extensively discussed at the time when the US declared the withdrawal from the Paris Climate Accords. When comparing different countries in which tweets were posted, it was observed that for Australia, nearly 50% of the collected tweets were assigned to the topic of energy, whereas in other countries like Canada, the United Kingdom, and the United States, this proportion was significantly lower (below 25%).

Achmann and Wolff (2023) applied BERTopic to explore topics in posts and stories (temporary available photo or video content) posted by political parties and politicians on Instagram in the context of the 2021 German federal election campaign. Achmann and Wolff (2023) employed BERTopic to derive 25 topics each from the posts and stories. The majority of posts dealt with policy issues, while the majority of stories did not deal with policy issues. Instead, they were concerned with the documentation of the rallies and campaign trail or call-to-action content, i.e. content that call on people to vote for a certain party or politician.

Furthermore, there have been studies investigating elections and the topics discussed by users on Twitter. Karami et al. (2018) collected 24 million tweets on the two candidates Barack Obama and Mitt Romney in the context of the 2012 US presidential election. Topics with words in the context of economy, jobs, budget deficit, healthcare and tax were identified. Taking the sentiment of tweets into account, it was shown that the topic of budget deficit was significantly more negative in tweets collected for Mitt Romney compared to those collected for Barack Obama.

Overall, topic modeling is widely utilized in the political context of Twitter to identify the wide range of topics discussed in tweets. Once topics are identified, it is common to examine, variations in topic prevalence over time and in different geographic regions as well as differences in sentiment of tweets on specific topics.

3 Methodology

3.1 Data acquisition

3.1.1 Tweets by political party accounts

In order to examine the topics addressed in tweets posted by political party accounts during the election year 2021, we used a Twitter corpus by Schmidt et al. (2022). The corpus comprises 58,864 tweets (see Table 1), whereby these were posted by 89 politicians and political party accounts of the seven largest parties represented in the Bundestag before and after the election (see Table 6 for the full list of accounts). The CDU and its sister party, the Bavarian regional party CSU, were considered as one party. For every party, the ten politician accounts and the four official party accounts with the highest number of followers as of January 2022 were considered. Furthermore, for each tweet in the corpus, there is already a sentiment classification for the text of the tweet, which can be either positive, negative or neutral. The sentiment was determined using a BERT model, which was fine-tuned by Schmidt et al. (2022).

We will include the text of the tweets into our upcoming topic modeling analysis. The tweets of this corpus by Schmidt et al. (2022) are referenced in the following as

Table 3 Identified topics in tweets posted by political party accounts

Topic	# Tweets	Top 3 tokens	Topic	# Tweets	Top 3 tokens
(*a*) *Text of the tweets*					
1	16,757	afd, berlin, bundestag	14	805	landwirtschaft, wald, landwirt
2	9143	corona, impfstoff, impfung	15	1078	twitter, tweet, trump
3	3383	klimaschutz, klima, sozial	16	606	fußball, spiel, hsv
4	3173	digital, digitalisierung, uhr	17	485	rassismus, hanau, rassistisch
5	2767	euro, rente, steuer	18	613	nazi, gedenken, opfer
6	3024	csu, cdu, abgeordneter	19	629	bahn, auto, mobilität
7	2000	eu, russland, belarus	20	926	arbeit, beschäftigter, lohn
8	1861	kind, schule, familie	21	504	pflege, krankenhaus, gesundheit
9	1906	youtube, live, orbit	22	452	maske, ffp, spahn
10	1854	polizei, bundeswehr, werden	23	301	atomkraft, fukushima, atomwaffe
11	1182	afghanistan, taliban, kabul	24	330	türkei, erdogan, türkisch
12	1182	israel, antisemitismus, jude	25	470	preis, wasserstoff, energiepreis
13	1164	frau, gender, sprache			
(*b*) *Text extracted from images of tweets*					
1	2601	deutschland, corona, afd	14	225	cdu, cduzcsu, ralph
2	1275	spd, cdu, fdp	15	145	scholz, olaf, spd
3	459	uhr, digital, facebook	16	105	afghanistan, taliban, bundeswehr
4	404	klimaschutz, energie, klima	17	108	russland, datenträger, unterlage
5	428	euro, rente, einkommen	18	113	virus, coronavirus, tz
6	310	cases, covid, data	19	115	arbeit, cduo, arbeiten
7	242	grün, grüne, top	20	104	landwirtschaft, künast, renate
8	215	sonntagsfrage, befragung, befragter	21	128	covid, patient, inzidenz
9	224	rassismus, opfer, antisemitismus	22	152	berlin, hbf, dielinke
10	204	pmk, unternehmen, quellcode	23	116	pandemie, kramp, karrenbauer
11	189	impfstoff, impfpflicht, biontech	24	93	steuer, geld, haushalt
12	178	kind, schule, familie	25	78	bartsch, dietmar, fraktionsvorsitzender
13	139	twitter, tweet, iphone			

tweets by political party accounts. In addition to analyzing the tweet's text, we also examine the text in the images posted in tweets. In order to do so, we subsequently extracted the text of the images via OCR. Version 5.3.1 of Tesseract[4] was used, and we were able to extract text from around half of the images (21,959) posted by political party accounts. Some images are mainly photographs of the politicians of the respective party and thus offer no text content. In case text is discernible within an image, the text often comprises quotations or political statements (see Fig. 1 for some examples).

3.1.2 Tweets mentioning politicians

In order to analyze which topics were addressed by accounts that mentioned political party accounts on Twitter in the election year, topic modeling was carried out on the corpus already acquired by Hellwig et al. (2023) in a previous study. The corpus comprises 707,241 tweets (see Table 2) that mentioned (using @-sign) the 89 party and politician accounts considered by Schmidt et al. (2022). Similar to the tweets in the corpus by Schmidt et al. (2022), there is a sentiment classification for the text of each tweet, which can be either positive, negative or neutral. Again, the sentiment was identified with the help of a fine-tuned BERT model (Hellwig et al., 2023).

Again, by using *Tesseract*, text was extracted from the images, although the number of images that could be extracted from the tweets was lower in proportion to the total number of tweets comparing it with the tweets posted by the political party accounts. When looking at some of the images, it can be noticed that the type of images is very different (see Fig. 2) to the ones by parties and politicians. Commonly, the visual content includes screenshots, such as those from news articles, as well as photographs depicting politicians affiliated with a respective political party.

[4] GitHub repository of Tesseract: https://github.com/tesseract-ocr/tesseract.

Table 4 Identified topics in tweets mentioning political party accounts

Topic	# Tweets	Top 3 tokens	Topic	# Tweets	Top 3 tokens
(*a*) *Text of the tweets*					
1	236,107	cdu, gut, kind	14	5721	digital, digitalisierung, internet
2	67,815	impfung, impfstoff, impfen	15	5236	angst, panik, panikmach
3	54,590	deutschland, deutsch, berlin	16	4553	korruption, korrupt, cdukorruption
4	24,157	geld, steuer, zahlen	17	3461	cannabis, legalisierung, cannabislegalisierung
5	11,354	tweet, twitter, lesen	18	4467	maske, tragen, ffp
6	13,716	klimaschutz, klima, klimawandel	19	4062	christlich, kirche, wert
7	11,994	union, arbeit, job	20	4535	flüchtling, migrant, migration
8	12,127	energie, strom, wind	21	3484	alkohol, trinken, rauchen
9	12,461	bild, journalist, tv	22	3661	droge, pharma, medikament
10	5744	israel, antisemitismus, islam	23	4051	lachen, satire, lustig
11	11,893	frau, mann, rassismus	24	3664	lockdown, hart, lockdowns
12	7758	auto, fahren, bahn	25	5016	fdp, liberal, konservativ
13	4832	nazi, afdbeobachtungjetzt, fckafd			
(*b*) *Text extracted from images of tweets*					
1	6700	deutschland, afd, cdu	14	378	frau, mann, arbeitsförderung
2	2960	euro, jahr, deutschland	15	287	flüchtling, deutschland, migration
3	1338	tweet, twitter, antwort	16	279	idiot, journalist, ideologie
4	1351	covid, cases, data	17	410	virus, variante, cov
5	1302	impfstoff, impfung, biontech	18	205	maske, ffp, belegung
6	1204	corona, jahr, coronavirus	19	213	israel, jude, antisemitismus
7	1147	covid, fall, patient	20	604	cdu, csu, maskendeal
8	783	politik, partei, politisch	21	274	scholz, olaf, spd
9	660	kind, schule, kitas	22	172	prime, deputy, party
10	685	union, spd, bundestagswahl	23	291	grüne, antrag, grün
11	321	cannabis, legalisierung, droge	24	188	digital, digitalisierung, neu
12	476	freiheit, recht, grundgesetz	25	246	lauterbach, karl, spd
13	351	afghanistan, taliban, bundeswehr			

3.2 Analyzing political topics in tweets using BERTopic

3.2.1 Topic modeling

Topic modeling was carried out for both the corpus with tweets from the political party accounts and the corpus with tweets that mentioned these accounts. A model was fitted for each of the following sub-corpora separately:

- Text extracted from the tweets posted by political party accounts.
- Text extracted from the images posted by political party accounts.
- Text extracted from tweets mentioning political party accounts.
- Text extracted from images of tweets mentioning political party accounts.

BERTopic was chosen because it incorporates the semantic relations between words by using embedding representation of documents, which helps to generate more meaningful topics (as already outlined in Sect. 2). Furthermore, it enabled us to analyze the frequency of documents on specific topics over time and the comparison of the topics identified in each sub-corpora, which is relevant to answer the research questions. Apart from utilizing BERTopic, we also employed Latent Dirichlet Allocation (LDA) to identify topics in the sub-corpora. However, the LDA models produced a relatively fewer number of meaningful topics. Often, there was no distinct thematic pattern that could be discerned based on the most prevalent tokens.[5]

For the concrete application of BERTopic, we used the official implementation for Python[6], which uses HDBSCAN

[5] GitHub repository with supplements: https://github.com/NilsHellwig/Topic_Modeling_Twitter_German_Federal_Election_2021.

[6] BERTopic—Python Package Index: https://pypi.org/project/bertopic/.

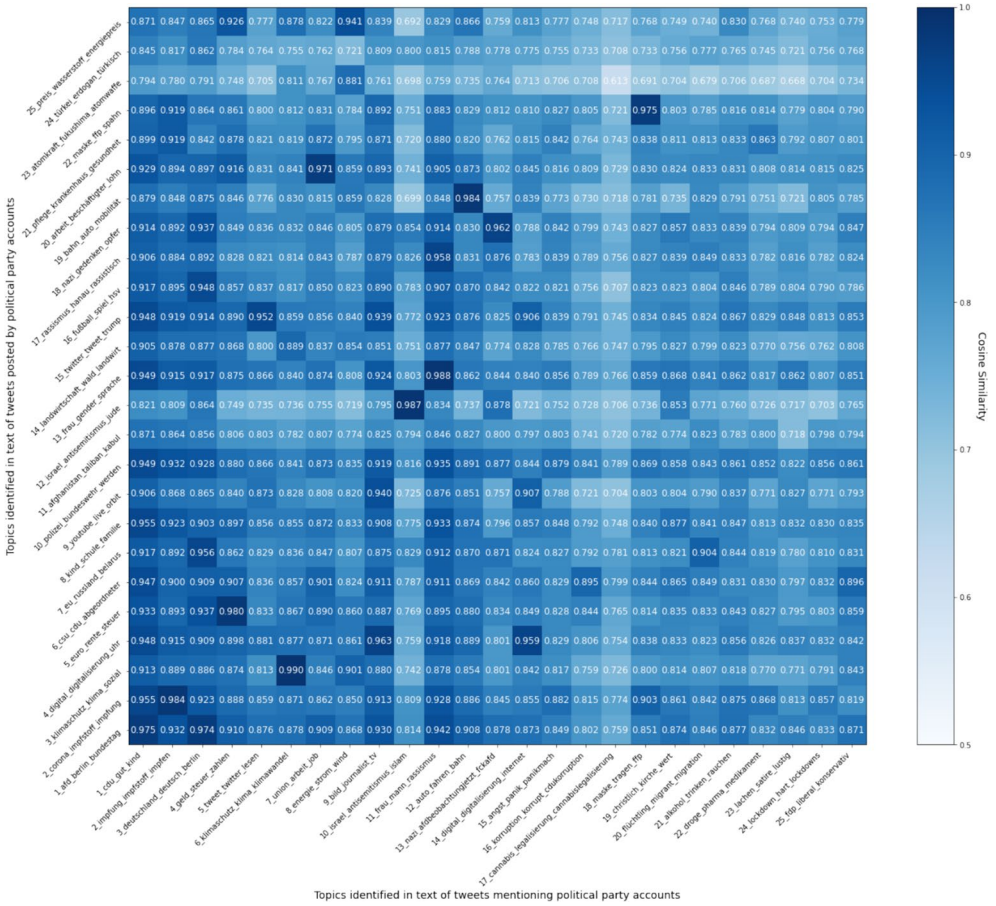

Fig. 3 Tweet text: Cosine similarity matrix for topic similarity comparison of topics identified in the corpus of tweets posted by political party accounts and tweets mentioning political party accounts

as a clustering algorithm. Without limiting the minimum number of documents per cluster, well over 50 clusters which are regarded as topics were identified for each of the four sub-corpora, however with some topics having a very small number of documents associated with them. In order to obtain more meaningful and important topics, a minimum number of documents per topic was defined based on the size of the sub-corpora. We determined that a topic must encompass a minimum of 1/250 of the total documents in the sub-corpus. Taking this limit into account, we were able to generate 25 topics for each sub-corpus.

3.2.2 Preprocessing and representation

The text of both tweets and the extracted text from images underwent preprocessing, which included removing punctuation and stop words, converting the text to lowercase as well as lemmatization via the Python package *Spacy*[7] which proved to be beneficial preprocessing steps for German language text (Fehle et al., 2021). Additionally, documents with fewer than five words were excluded. To represent the preprocessed tweets, a Sentence-BERT model, specifically "*paraphrase-multilingual-MiniLM-L12-v2*"[8], which is suggested for multilingual documents or any other language than English.[9] All tweets were transformed into embeddings using an NVIDIA GeForce GTX 1080Ti GPU with 11 GB VRAM. As suggested by Grootendorst (2022), UMAP is used to reduce the dimensionality of the embeddings, which can improve the performance of clustering algorithms such as k-Means or HDBSCAN in terms of clustering accuracy and time (Allaoui et al., 2020).

[7] Spacy, Python Package Index: https://pypi.org/project/spacy/.

[8] "*Paraphrase-multilingual-MiniLM-L12-v2*", HuggingFace: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

[9] BERTopic FAQ: https://maartengr.github.io/BERTopic/faq.html.

4 Results

4.1 Identified topics

In this section, we present the topics derived from tweets originating from the political party accounts, as well as those mentioning these accounts. The appendix D encompasses diagrams illustrating the prevalence of the topics identified for each of the four sub-corpora over the course of the election year. For the topics identified within the sub-corpora of tweet texts, supplementary diagrams are provided, offering insights into the sentiment trends associated with each topic throughout the election year.

4.1.1 Tweets of politicians

Topics identified in tweets posted by political party accounts are presented in Table 3. An English translation for all tokens can be found in the Table 10. For both sub-corpora, it can be noticed that there is one topic to which a comparatively large number of tweets were assigned (topic 1). These were mainly tweets that could not be assigned to a policy-focused topic, but referred to politicians, parties and their programs in general. In this context, some identified topics focused specifically on individual parties (e.g. tweet text: topic 6; tweet image: topic 7, topic 14, topic 15).

Looking at the policy-focused topics, some were debated throughout the election year. An extensively discussed topic is the COVID-19 pandemic and the challenges it posed to the society (tweet text: topic 2; tweet image: topic 6). However, topics related to COVID-19, such as masks (tweet text: topic 22) or vaccinations (German: "Impfung") (tweet image: topic 11) were also identified. In the majority of tweets assigned to this topic, a negative sentiment was expressed. Furthermore, the topic was less frequently discussed between May and October compared to the other months. Further policy-related topics that endured throughout the election year include climate policy (tweet text: topic 3; tweet image: topic 4), digitization (tweet text: topic 4; tweet image: topic 3) and finance policy (tweet text: topic 5; tweet image: topic 5).

Furthermore, some topics were primarily discussed at certain points in time. As an example, there's a topic surrounding Afghanistan and the return to power of the Taliban in August 2021, which was identified in both sub-corpora (tweet text: topic 12; tweet image: topic 16) and was discussed the most in August 2021. Similarly, a topic focused on Israel and antisemitism garnered significant attention in May 2021 (tweet text: topic 12).

4.1.2 Tweets mentioning accounts by politicians

All topics that were identified in the two sub-corpora of tweets mentioning the political party accounts are presented in Table 4). An English translation for all tokens can be found in the Table 11. As for the tweets posted by the political party accounts, a comparatively large topic could be identified for both sub-corpora to which tweets were assigned that were not policy-focused (tweet text: topic 1, tweet images: topic 1).

Policy-focused topics that were debated throughout the election year were identified. As for the tweets of the political party accounts, topics focusing on the COVID-19 pandemic were identified. There are several topics related to different aspects of the COVID-19 pandemic, such as vaccinations (tweet text: topic 2, tweet images: topic 5), masks (tweet text: topic 18, tweet image: topic 18) or lockdowns (tweet text: topic 24).

Other policy-focused topics include finance policy (tweet text: topic 4) or social issues like gender equality and racism (tweet text: topic 11, tweet image: topic 14). In both sub-corpora, we identified a prevalent topic focusing on cannabis (tweet text: topic 17, tweet image: topic 11), which remained a prominent subject of discussion throughout the entire election year. It is worth mentioning that as of April 1st, 2024, the possession of small quantities has been legalized in Germany. However, there was an outlier in October for both sub-corpora, meaning that the topic was debated comparatively much in that month (see Fig. 6).

In addition, it can be noticed that within the sub-corpus containing tweet texts, two distinct topics emerged, encompassing tweets that not only garnered significant attention throughout the election year but almost exclusively express a negative sentiment. This is on the one hand a topic focusing on the party AfD (tweet text: topic 13) and on the other hand a topic focusing on corruption scandals of the parties CDU and CSU and corruption in general (tweet text: topic 16).

As for the tweets posted by political party accounts, we identified topics that underwent extensive debates on Twitter at specific points in time. As for the political party accounts, a topic with regard to Afghanistan and the return to power of the Taliban in August 2021 (tweet image: topic 13) was identified but only for the sub-corpus of images. Equally, we identified a topic focusing on Israel and antisemitism (tweet text: topic 10, tweet image: topic 19), which was extensively discussed in May 2021.

4.2 Topic similarity between corpora

To further illustrate the presence of topics addressed in both tweets of political party accounts and those mentioning them, we identified similar topics between the sub-corpora using their corresponding topic embeddings. Subsequently, we compared these topics addressed by these two perspectives using cosine similarity, and for visualization purposes, we utilized a cosine similarity matrix.

We considered on the one hand topics that could be identified in the tweet text (see Fig. 3), but on the other hand also topics that were identified in the text extracted from images (see Fig. 8). Upon examining the two resulting cosine similarity matrices, several topics are being discussed by both perspectives, such as COVID-19, climate policy, financial policy, antisemitism, and social issues like gender equality and racism. Topics like cannabis (text of mentions: topic 17, text extracted from images: topic 11) or humor and satire (text of mentions: topic 23) could not be identified on the side of political party tweets. On the other hand, there are topics like Turkey (text of political party tweets: topic 24) for which no equivalent topic could be identified on the side of tweets that mentioned political party accounts.

5 Discussion

In the following section, we discuss and interpret the overall results, highlighting notable findings that emerged from our analysis. First, we collected the images posted in tweets from the corpora curated by Schmidt et al. (2022) and Hellwig et al. (2023) and used OCR to extract their texts. The average number of images per tweet was higher in tweets posted by political party accounts compared to tweets that mentioned political party accounts.

We employed BERTopic to identify topics in four sub-corpora: the two newly curated sub-corpora and the existing sub-corpora containing the tweet texts curated by Schmidt et al. (2022) and Hellwig et al. (2023). Similar to the findings by Achmann and Wolff (2023), we observed the presence of a comparatively large topic within all sub-corpora we examined, which does not focus on specific policies, but concentrates on politicians, parties and their programs in general. Both on the side of tweets that were posted by political party accounts and on the side of tweets that mentioned political party accounts, the COVID-19 pandemic was an intensively discussed topic. Additionally, various other topics also addressed different aspects partly related to COVID-19. Schmidt et al. (2022) speculated that the COVID-19 pandemic might be a reason for the overall negative sentiment in tweets from political party accounts. Our findings support this hypothesis, as we observed that tweets related to COVID-19 were predominantly associated with negative sentiment.

Furthermore, a topic surrounding Afghanistan and the return to power of the Taliban was identified in the tweets of political party accounts. Since the sentiment is primarily negative, this debated topic could indeed be a cause for the overall negative sentiment of tweets from some parties in August, as suggested by Schmidt et al. (2022). Other debated topics include finance policy, social issues and energy policy, which have been subjects of debate on Twitter in previous elections as well (Karami et al., 2018; Miller, 2019).

We proceeded with a comprehensive comparison of topic embeddings to ascertain the presence of topics addressed in both tweets by political party accounts and tweets that mentioned those. Topics such as the COVID-19 pandemic, climate policy, financial policy, antisemitism and social issues were discussed by both sides. Nevertheless, certain topics, such as cannabis, were exclusively addressed in tweets mentioning political party accounts. Notably, it is likely that these topics can also be identified in tweets of the other perspective by increasing the number of topics to be identified beyond 25.

Overall, BERTopic proved to be effective in identifying meaningful topics, although, especially considering the image sub-corpora, there were topics for which the thematic context can not be clearly discerned based on the most frequent tokens and tweets assigned to them (e.g. Table 3b, topic 22).

6 Limitations

Our work provides valuable insights into the topics debated in both tweets by political party accounts and tweets mentioning them on Twitter in the election year 2021. However, there are certain limitations of our work that we intend to address: While we attempted to capture the topics addressed through images by extracting text from them, we need to acknowledge that this approach may only partially capture the range of topics conveyed via this medium. In future work, we intend to enhance the analysis by incorporating automated image captioning methods or human annotations to gain a more comprehensive understanding of the topics addressed through images and how they are portrayed in terms of sentiment and presentation. In addition, it is important to note that, occasionally, the applied OCR might misinterpret text in the images, particularly in the case of proper nouns or unique words. Another limitation is the restriction of identifying a maximum of 25 topics for each sub-corpus. As a result, there are inevitably other topics that were discussed in tweets of a certain sub-corpus, but do not represent an independent topic because not enough tweets were assigned to them. Finally, it is worth noting that Twitter's popularity and usage in Germany is not as widespread as in other countries. Only 10% of Germans regularly use Twitter,[10] in contrast to 23% of U.S. adults.[11] As a consequence,

[10] Germans' use of online services—Statista: https://de.statista.com/statistik/daten/studie/171006/umfrage/in-anspruch-genommene-angebote-aus-dem-internet/.

[11] Twitter users in the US, Statista: https://www.statista.com/statistics/232818/active-us-twitter-user-growth/.

the corpora curated only represent a limited subsection of public (social media) sentiment.

7 Conclusion and future work

In conclusion, this study provides insights into the Twitter discourse surrounding the 2021 German federal election by utilizing BERTopic for topic modeling. Examining both the tweet text and the text extracted from images allowed to analyze a wide range of topics discussed on Twitter throughout the election year. Our analysis of tweets posted by accounts of the major German political parties and those mentioning them revealed that the most discussed topics and sentiment trends throughout the election year. In all sub-corpora examined, we noted the prevalence of a larger topic not focussing on policy-oriented topics. Instead, that topic tends to focus on politicians, parties, and their general campaigns. Looking at policy-focused topics, we found that the COVID-19 pandemic, climate policy, finance policy and social issues like racism and gender equality were among the most prominent topics discussed on Twitter during the election.

In future work, we plan to conduct a more in-depth analysis of visual content such as images and videos shared in the context of the election. In addition to extracting text from images, we intend to explore the emotion tone and sentiment conveyed through these using computer vision techniques (Schmidt et al., 2021b; El-Keilany et al., 2022). We also see potential in switching from the basic sentiment concept to more fine-grained emotion analysis (Schmidt et al., 2021a; Dennerlein et al., 2023) and also include methods of aspect based sentiment analysis to analyse the cause-effect relation of sentiments and emotions (Fehle et al., 2023). Finally, one could investigate the alignment between topics discussed on Twitter and the debates in the German Bundestag. By exploring potential overlaps between political discourse on social media platforms and formal legislative discussions, we seek to explore the extent to which public concerns find resonance in the legislative process.[12]

Appendix A: Results of German federal election 2021

See Appendix Table 5.

Table 5 Election results of the 2021 federal election and changes compared to the previous election in 2017

Party	Full name	2021 (%)	2017 (%)	Change (%)
SPD	Social Democratic Party of Germany	25.7	20.5	+ 5.2
CDU/CSU	Christian Democratic Union/Christian Social Union (Bavaria)	24.1	32.9	− 8.8
Die Grünen	The Greens	14.8	8.9	+ 5.9
FDP	Free Democratic Party	11.5	10.7	+ 0.8
AfD	Alternative for Germay	10.3	12.6	− 2.3
Die Linke	The Left	4.9	9.2	− 4.3

[12] GitHub Repository with supplements: https://github.com/NilsHellwig/Topic_Modeling_Twitter_German_Federal_Election_2021.

Appendix B: Twitter accounts for data acquisition

Appendix B.1: Parties

See Appendix Table 6.

Table 6 The three main accounts with the most followers of each party

SPD	CDU	CSU	Die Grünen	FDP	AfD	Die Linke
@spdde Follower: 417k Tweets: 22,138	@CDU Follower: 378k Tweets: 37,100	@CSU Follower: 229k Tweets: 9072	@Die_Gruenen Follower: 649k Tweets: 30,560	@fdp Follower: 414k Tweets: 27,981	@AfD Follower: 173k Tweets: 8330	@dieLinke Follower: 350k Tweets: 14,135
@spdbt Follower: 217k Tweets: 9809	@cducsubt Follower: 166k Tweets: 13,250		@GrueneBundestag Follower: 186k Tweets: 6399	@fdpbt Follower: 39k Tweets: 8194	@AfDimBundestag Follower: 68k Tweets: 4713	@Linksfraktion Follower: 108k Tweets: 2994
@jusos Follower: 77k Tweets: 1847	@Junge_Union Follower: 79k Tweets: 931		@gruene_jugend Follower: 76k Tweets: 1290	@fdp_nrw Follower: 28k Tweets: 884	@AfDBerlin Follower: 19k Tweets: 364	@dielinkeberlin Follower: 19k Tweets: 1228

Appendix B.2: Politicians

See Appendix Table 7.

Table 7 The 10 accounts with the most followers of each party

SPD	CDU	CSU	Die Grünen	FDP	AfD	Die Linke
@Karl_Lauterbach Follower: 770k Tweets: 132,526	@jensspahn Follower: 279k Tweets: 35,571	@Markus_Soeder Follower: 341k Tweets: 30,495	@cem_oezdemir Follower: 290k Tweets: 9942	@c_lindner Follower: 552k Tweets: 19,942	@Alice_Weidel Follower: 138k Tweets: 9367	@SWagenknecht Follower: 518k Tweets: 7177
@HeikoMaas Follower: 660k Tweets: 6431	@ArminLaschet Follower: 188k Tweets: 36,161	@DoroBaer Follower: 103k Tweets: 2560	@GoeringEckardt Follower: 202k Tweets: 5227	@MaStrackZi Follower: 46k Tweets: 2453	@Joerg_Meuthen Follower: 76k Tweets: 4813	@GregorGysi Follower: 439k Tweets: 1722
@OlafScholz Follower: 324k Tweets: 27,414	@_FriedrichMerz Follower: 179k Tweets: 23,651	@andreasscheuer Follower: 63k Tweets: 2431	@JTrittin Follower: 115k Tweets: 1782	@MarcoBuschmann Follower: 46k Tweets: 10,062	@Beatrix_vStorch Follower: 68k Tweets: 3962	@katjakipping Follower: 130k Tweets: 1072
@KuehniKev Follower: 323k Tweets: 5192	@JuliaKloeckner Follower: 74k Tweets: 3357	@ManfredWeber Follower: 54k Tweets: 527	@KonstantinNotz Follower: 85k Tweets: 2144	@KonstantinKuhle Follower: 44k Tweets: 2710	@gottfriedcurio Follower: 37k Tweets: 275	@DietmarBartsch Follower: 82k Tweets: 3409
@larsklingbeil Follower: 116k Tweets: 5669	@n_roettgen Follower: 68k Tweets: 4645	@DerLenzMdB Follower: 10k Tweets: 236	@RenateKuenast Follower: 77k Tweets: 2026	@johannesvogel Follower: 38k Tweets: 2121	@MalteKaufmann Follower: 36k Tweets: 5149	@anked Follower: 43k Tweets: 935
@hubertus_heil Follower: 108k Tweets: 2406	@PaulZiemiak Follower: 58k Tweets: 12,723	@hahnflo Follower: 9k Tweets: 2900	@Ricarda_Lang Follower: 65k Tweets: 3546	@Wissing Follower: 32k Tweets: 2805	@JoanaCotar Follower: 30k Tweets: 4330	@b_riexinger Follower: 41k Tweets: 1399
@EskenSaskia Follower: 101k Tweets: 7180	@groehe Follower: 49k Tweets: 79	@smuellermdb Follower: 9k Tweets: 239	@KathaSchulze Follower: 37k Tweets: 4609	@Lambsdorff Follower: 27k Tweets: 884	@Tino_Chrupalla Follower: 21k Tweets: 2875	@jankortemdb Follower: 34k Tweets: 743
@Ralf_Stegner Follower: 64.9k Tweets: 7061	@HBraun Follower: 39k Tweets: 3212	@DaniLudwigMdB Follower: 8k Tweets: 3821	@BriHasselmann Follower: 37k Tweets: 1795	@ria_schroeder Follower: 23k Tweets: 359	@StBrandner Follower: 23k Tweets: 11,914	@Janine_Wissler Follower: 37k Tweets: 1046
@KarambaDiaby Follower: 55.6k Tweets: 392	@rbrinkhaus Follower: 30k Tweets: 4280	@ANiebler Follower: 6k Tweets: 25	@nouripour Follower: 29k Tweets: 505	@LindaTeuteberg Follower: 23k Tweets: 328	@GtzFrmming Follower: 17k Tweets: 984	@SevimDagdelen Follower: 35k Tweets: 172
@MiRo_SPD Follower: 39k Tweets: 350	@tj_tweets Follower: 17k Tweets: 396	@MarkusFerber Follower: 5k Tweets: 21	@MiKellner Follower: 28k Tweets: 3436	@f_schaeffler Follower: 20k Tweets: 1092	@PetrBystronAFD Follower: 17k Tweets: 496	@SusanneHennig Follower: 29k Tweets: 4463

Appendix C: Information regarding example images

Appendix C.1: Tweets by political party accounts

See Appendix Table 8.

Table 8 URL and author of the tweet from the examples of images posted by political party accounts

Image URL	Account	Associated party
https://twitter.com/dielinkeberlin/status/1447545102235222019	@dielinkeberlin	Die Linke
https://twitter.com/CSU/status/1406945392181252100	@CSU	CSU
https://x.com/BriHasselmann/status/1367109678350667776	@BriHasselmann	Die Grünen

Appendix C.2: Tweets mentioning political party accounts

See Appendix Table 9.

Table 9 URL and author of the tweets from the examples of images posted by accounts mentioning political party accounts

Image URL	Mentioned account	Associated party
https://twitter.com/ChaosMono/status/1423962932833103883	@DerLenzMdB	CSU
https://twitter.com/DC08817836/status/1408871840215453700	@Alice_Weidel	AfD
https://twitter.com/ChristianLange_/status/1374082373579595784	@ria_schroeder	FDP

Appendix D: Number of tweets for each identified topic over the course of the election year

Legend:
green—positive sentiment, yellow—neutral sentiment, red—negative sentiment, grey/black—total number of tweets in month

Appendix D.1: Tweets posted by political party accounts

See Appendix Figs. 4 and 5.

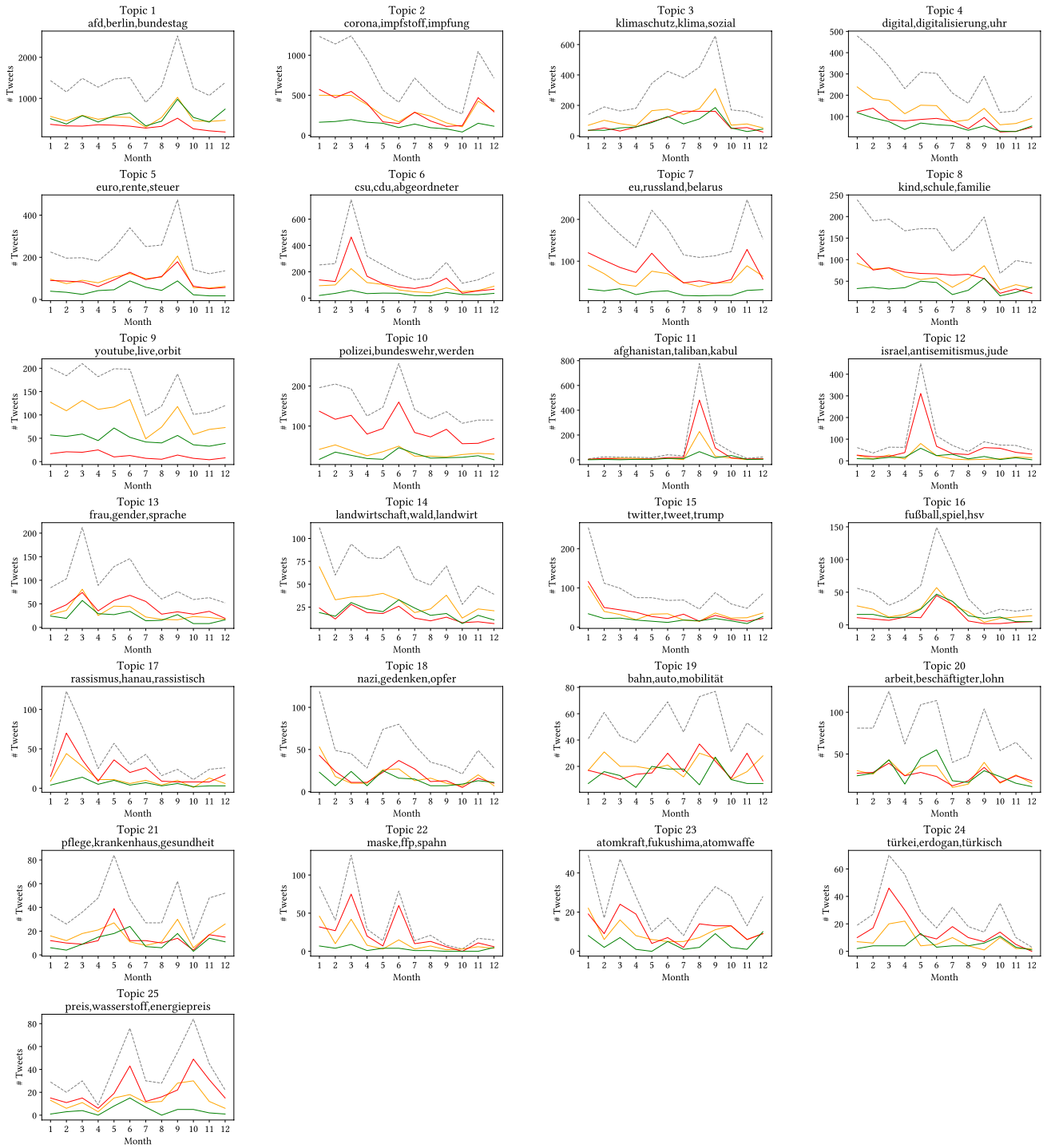

Fig. 4 Text of tweets posted by political party accounts: Number of tweets for each identified topic over the course of the election years

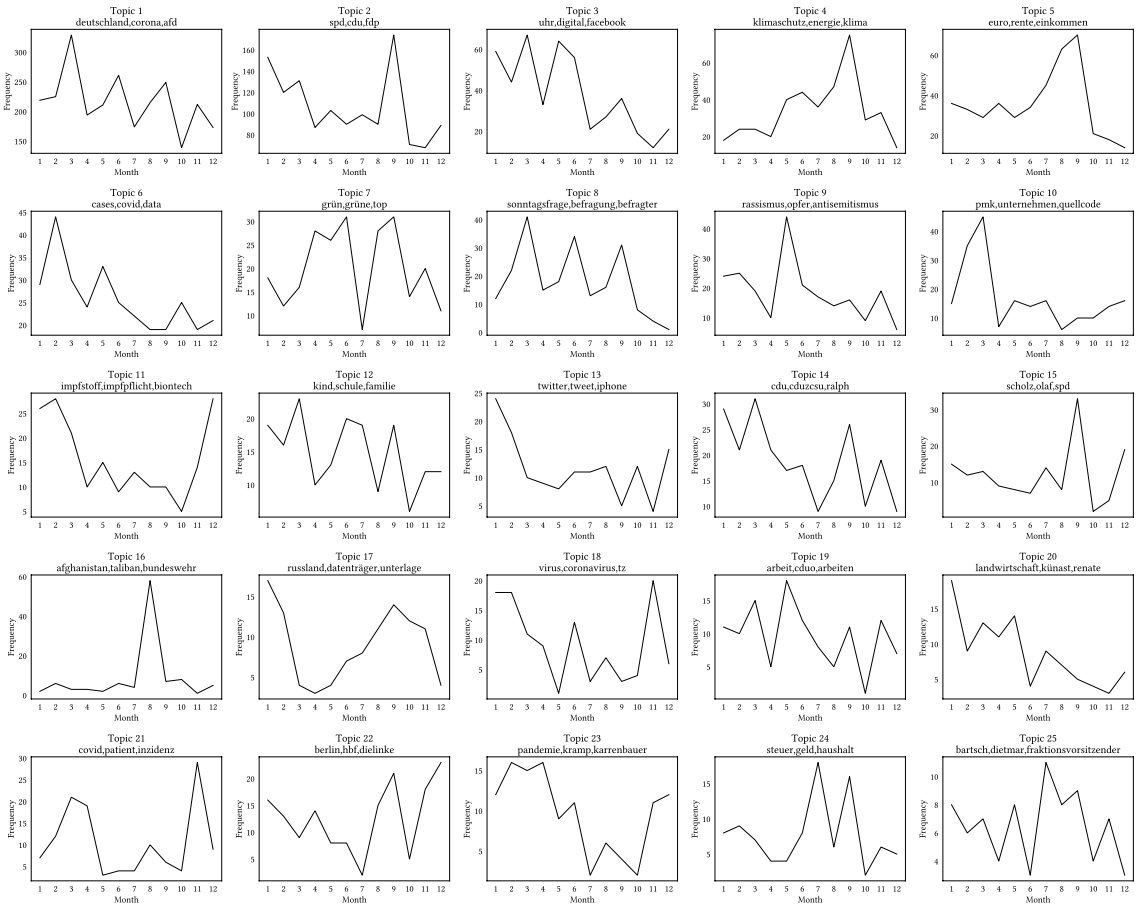

Fig. 5 Text extracted from images of tweets posted by political party accounts: Number of tweets for each identified topic over the course of the election years

Appendix D.1.2: Tweets mentioning political party accounts

See Appendix Figs. 6 and 7.

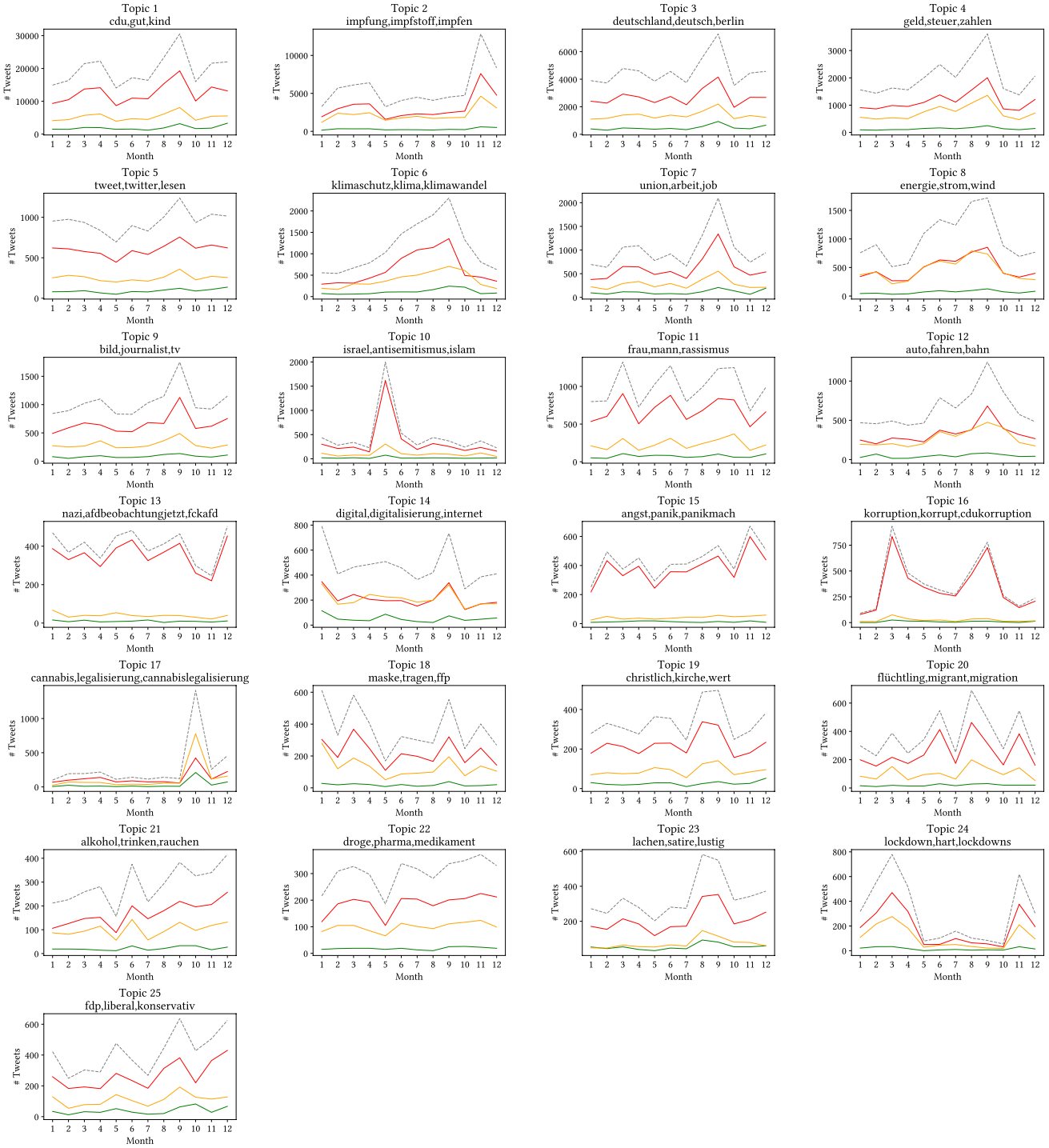

Fig. 6 Text of tweets mentioning political party accounts: Number of tweets for each identified topic over the course of the election years

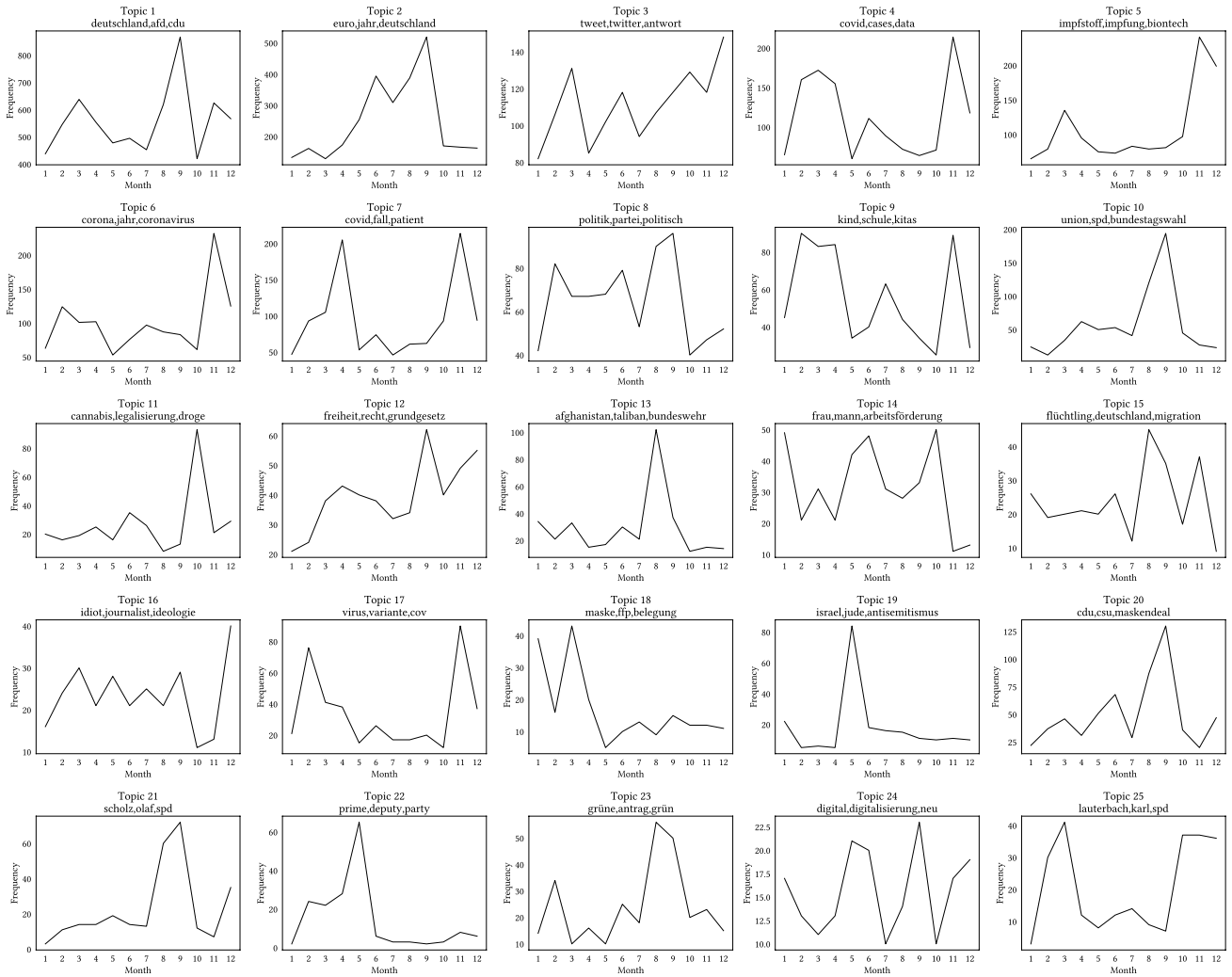

Fig. 7 Text extracted from images of tweets mentioning political party accounts: Number of tweets for each identified topic over the course of the election years

Appendix E: English translation of the most frequent tokens for the identified topics

See Appendix Tables 10 and 11.

Table 10 Identified topics in tweets posted by political party accounts

Thema	# Tweets	Top 3 Tokens	Thema	# Tweets	Top 3 Tokens
(*a*) *Text of the tweets*					
1	16,757	afd, berlin, bundestag	14	805	agriculture, forest, farmer
2	9143	corona, vaccine, vaccination	15	1078	twitter, tweet, trump
3	3383	climate protection, climate, social	16	606	football, game, hsv
4	3173	digital, digitization, clock	17	485	racism, hanau, racist
5	2767	euro, pension, tax	18	613	nazi, remembrance, victim
6	3024	csu, cdu, member of parliament	19	629	train, car, mobility
7	2000	eu, russia, belarus	20	926	work, employee, wage
8	1861	child, school, family	21	504	care, hospital, health
9	1906	youtube, live, orbit	22	452	mask, ffp, spahn
10	1854	police, military, become	23	301	nuclear power, fukushima, nuclear weapon
11	1182	afghanistan, taliban, kabul	24	330	turkey, erdogan, turkish
12	1182	israel, antisemitism, jew	25	470	price, hydrogen, energy price
13	1164	woman, gender, language			
(*b*) *Text extracted from images of tweets*					
1	2601	germany, corona, afd	14	225	cdu, cduzcsu, ralph
2	1275	spd, cdu, fdp	15	145	scholz, olaf, spd
3	459	clock, digital, facebook	16	105	afghanistan, taliban, bundeswehr
4	404	climate protection, energy, climate	17	108	russia, storage medium, document
5	428	euro, pension, income	18	113	virus, coronavirus, tz
6	310	cases, covid, data	19	115	work, cduo, working
7	242	green, greens, top	20	104	agriculture, künast, renate
8	215	opinion poll, survey, respondent	21	128	covid, patient, incidence
9	224	racism, victim, antisemitism	22	152	berlin, hbf, die Linke
10	204	pmk, company, source code	23	116	pandemic, kramp, karrenbauer
11	189	vaccine, vaccination obligation, biontech	24	93	tax, money, budget
12	178	child, school, family	25	78	bartsch, dietmar, parliamentary group leader
13	139	twitter, tweet, iphone			

Table 11 Identified topics in tweets mentioning political party accounts

Thema	# Tweets	Top 3 Tokens	Thema	# Tweets	Top 3 Tokens
(*a*) *Text of the tweets*					
1	236,107	cdu, good, child	14	5721	digital, digitization, internet
2	67,815	vaccination, vaccine, vaccinating	15	5236	fear, panic, scaremongering
3	54,590	germany, german, berlin	16	4553	corruption, corrupt, cdubribery
4	24,157	money, tax, paying	17	3461	cannabis, legalization, cannabislegalization
5	11,354	tweet, twitter, reading	18	4467	mask, wearing, ffp
6	13,716	climate protection, climate, climatechange	19	4062	christian, church, value
7	11,994	union, work, job	20	4535	refugee, migrant, migration
8	12,127	energy, electricity, wind	21	3484	alcohol, drinking, smoking
9	12,461	picture, journalist, tv	22	3661	drug, pharmaceutical, medication
10	5744	israel, antisemitism, islam	23	4051	laughter, satire, funny
11	11,893	woman, man, racism	24	3664	lockdown, hard, lockdowns
12	7758	car, driving, train	25	5016	fdp, liberal, conservative
13	4832	nazi, fckafd, fckafd			
(*b*) *Text extracted from images of tweets*					
1	6700	germany, afd, cdu	14	378	woman, man, jobpromotion
2	2960	euro, year, germany	15	287	refugee, germany, migration
3	1338	tweet, twitter, reply	16	279	idiot, journalist, ideology
4	1351	covid, cases, data	17	410	virus, variant, cov
5	1302	vaccine, vaccination, biontech	18	205	mask, ffp, occupancy
6	1204	corona, year, coronavirus	19	213	israel, jew, antisemitism
7	1147	covid, case, patient	20	604	cdu, csu, maskdeal
8	783	politics, party, political	21	274	scholz, olaf, spd
9	660	child, school, daycare	22	172	prime, deputy, party
10	685	union, spd, federal election	23	291	green, motion, green
11	321	cannabis, legalization, drug	24	188	digital, digitization, new
12	476	freedom, law, constitution	25	246	lauterbach, karl, spd
13	351	afghanistan, taliban, bundeswehr			

Appendix F: Cosine similarity matrices for comparison of topic similarity

See Appendix Fig. 8.

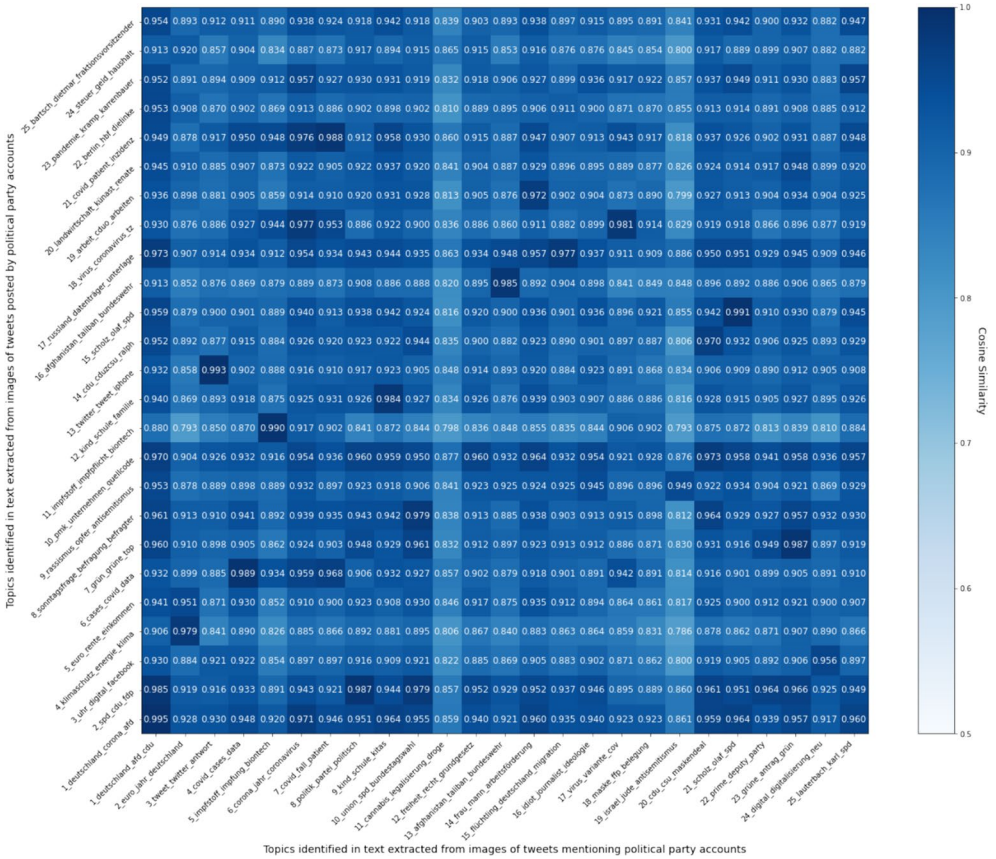

Fig. 8 Image text: cosine similarity matrix for topic similarity comparison of topics identified in the corpus of tweets posted by political party accounts and tweets mentioning political party accounts

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The tweets, including their text and any potentially included images, are available upon request for scientific purposes. Programming code, visualizations and corpus information are publicly available on GitHub.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Achmann, M., & Wolff, C. (2023). Policy issues vs. documentation: Using bertopic to gain insight in the political communication in Instagram stories and posts during the 2021 German federal election campaign. *Digital Humanities in the Nordic and Baltic Countries Publications, 5*(1), 11–28. https://doi.org/10.5617/dhnbpub.10647

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In *Image and signal processing: 9th international conference (ICISP 2020),* proceedings 9, (pp. 317–325), June 4–6, 2020, Marrakesh, Morocco.

Asuncion, H. U., Asuncion, A. U., & Taylor, R. N. (2010). Software traceability with topic modeling. In *Proceedings of the 32nd ACM/IEEE international conference on software engineering-volume 1* (pp. 95–104).

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet allocation. *Advances in Neural Information Processing Systems, 14*. https://doi.org/10.7551/mitpress/1120.003.0082

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of mixed membership models and their applications* (Vol. 225255).

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia presidential election from twitter using sentiment

analysis. *Journal of Big data, 5*(1), 1–10. https://doi.org/10.1186/s40537-018-0164-1

Costa, C., Aparicio, M., & Aparicio, J. (2021). Sentiment analysis of Portuguese political parties communication. In *Proceedings of the 39th ACM international conference on design of communication* (pp. 63–69).

Dahal, B., Kumar, S. A., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, 9*, 1–20.

Dennerlein, K., Schmidt, T., & Wolff, C. (2023). Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century. *Digital Scholarship in the Humanities, 38*(4), 1466–1481. https://doi.org/10.1093/llc/fqad046

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics.

El-Keilany, A., Schmidt, T., & Wolff, C. (2022). Distant viewing of the Harry Potter movies via computer vision. In K. Berglund, M. La Mela, & I. Zwart (Eds.), *Proceedings of the 6th digital humanities in the Nordic and Baltic countries Conference (DHNB 2022)* (pp. 33–49). Uppsala, Sweden. Retrieved from https://ceur-ws.org/Vol-3232/paper03.pdf

Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023). Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews. In M. Georges, A. Herygers, A. Friedrich, & B. Roth (Eds.), *Proceedings of the 19th conference on natural language processing (konvens 2023)* (pp. 202–218). Association for Computational Lingustics. Retrieved from https://aclanthology.org/2023.konvens-main.21

Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques. In K. Evang, L. Kallmeyer, R. Osswald, J. Waszczuk, & T. Zesch (Eds.), *Proceedings of the 17th conference on natural language processing (konvens 2021)* (pp. 86–103). KONVENS 2021 Organizers. Retrieved from https://aclanthology.org/2021.konvens-1.8

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794

Hellwig, N. C., Bink, M., Schmidt, T., Fehle, J., & Wolff, C. (2023). Transformer-based analysis of sentiment towards German political parties on Twitter during the 2021 election year. In M. Abbas & A. A. Freihat (Eds.), *Proceedings of the 6th international conference on natural language and speech processing (ICNLSP 2023)* (pp. 84–98). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.icnlsp-1.9

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88).

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*, 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Karami, A., Bennett, L. S., & He, X. (2018). Mining public opinion about economic issues: Twitter and the U.S. presidential election. *International Journal of Strategic Decision Sciences, 9*(1), 18–28. https://doi.org/10.4018/IJSDS.2018010102

Miller, D. T. (2019). Topics and emotions in Russian twitter propaganda. *First Monday*. https://doi.org/10.5210/fm.v24i5.9638

Moßburger, L., Wende, F., Brinkmann, K., & Schmidt, T. (2020). Exploring online depression forums via text mining: A comparison of Reddit and a curated online forum. In G. Gonzalez-Hernandez et al. (Eds.), *Proceedings of the fifth social media mining for health applications workshop & shared task* (pp. 70–81). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.smm4h-1.11

Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, pp. 265–272).

Schmidt, T., Bauer, M., Habler, F., Heuberger, H., Pilsl, F., & Wolff, C. (2020a). Der einsatz von distant reading auf einem korpus deutschsprachiger songtexte. In C. Schöch (Ed.), *Dhd 2020: Spielräume; digital humanities zwischen Modellierung und Interpretation. Konferenzabstracts; (pp. 296–300),* Universität Paderborn, 2. bis 6. März 2020, Paderborn, Germany. Retrieved from https://epub.uni-regensburg.de/43704/

Schmidt, T., Dennerlein, K., & Wolff, C. (2021a). Emotion Classification in German plays with transformer-based language models pretrained on historical and contemporary language. In *Proceedings of the 5th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (pp. 67–79). Association for Computational Linguistics. Retrieved 16 December 2021, from https://aclanthology.org/2021.latechclfl-1.8

Schmidt, T., El-Keilany, A., Eger, J., & Kurek, S. (2021b). Exploring computer vision for film analysis: A case study for five canonical movies. In *2nd international conference of the European association for digital humanities (EADH 2021)*. Krasnoyarsk, Russia. Retrieved 21 April 2022 from https://epub.uni-regensburg.de/50867/

Schmidt, T., Fehle, J., Weissenbacher, M., Richter, J., Gottschalk, P., & Wolff, C. (2022). Sentiment analysis on Twitter for the major German parties during the 2021 German federal election. In R. Schaefer, X. Bai, M. Stede, & T. Zesch (Eds.), *Proceedings of the 18th conference on natural language processing (konvens 2022)* (pp. 74–87). KONVENS 2022 Organizers. Retrieved from https://aclanthology.org/2022.konvens-1.9

Schmidt, T., Hartl, P., Ramsauer, D., Fischer, T., Hilzenthaler, A., & Wolff, C. (2020b). Acquisition and analysis of a meme corpus to investigate web culture. In L. Estill & J. Guiliano (Eds.), *15th annual international conference of the alliance of digital humanities organizations (DH 2020), conference abstracts.* Ottawa, Canada. Retrieved from https://epub.uni-regensburg.de/49294/

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using latent Dirichlet allocation for topic modeling on twitter. *PloS ONE, 15*(9), e0239441. https://doi.org/10.1371/journal.pone.0239441