DEClimate: A Dataset of Climate Change Discourse in German Politics

Ronja Memminger and Manfred Stede
Applied Computational Linguistics
University of Potsdam
(memminger|stede)@uni-potsdam.de

Abstract

Computational studies of climate change discourse in the political domain suffer from a scarcity of publicly available data. To help remedy this, we present a corpus of texts by the five German political parties currently represented in parliament, on the topic of climate change (CC). Spanning three million tokens and with texts sourced from a variety of communication channels, ranging from parliamentary debates to social media posts, our dataset allows for detailed analyses of the CC discourse in the German political sphere. We make the dataset publicly available.[1]

1 Introduction

Studies on climate change (CC) discourse in the political sphere have increased in recent years. Particularly the release of ClimateBERT (Webersinke et al., 2022) marked an improvement in climate-relevant text classification, easing access to English CC data by providing a filtering framework. However, computational studies on CC discourse rarely release their data publicly. This creates a barrier to applications of Natural Language Processing (NLP) analyses on climate related texts. Researchers must collect and curate specialized datasets for the topic, expending time and computation to filter relevant text. Furthermore, for languages other than English, tools for filtering text, or finalized corpora are rare.

We introduce DEClimate – a corpus of 50,091 text paragraphs by German political parties and politicians on the topic of climate change. It contains text from the last three German legislative periods from a variety of sources, capturing more than a decade of CC discourse over different mediums and contexts. The majority of the dataset is sourced from parliamentary speeches, but it also contains social media posts, manifestos and press releases. We make a the corpus publicly available in order to increase data accessibility for CC discourse research.

This dataset builds on the AfD Climate Change Corpus (AfD-CCC) (Stede and Memminger, 2025), a similar collection of CC paragraphs from the German right-wing populist party "Alternative für Deutschland". Using a similar framework, improved topic filtering, and a greater number of source domains, we extend this dataset to represent CC discourse in German politics at large.

2 Related Work

While there is a general scarcity of publicly available CC datasets, there exist some notable exceptions. A recent publication that has taken steps to close this gap is Policlim. With Policlim, Sanford et al. (2025) present a multilingual dataset of political party manifestos from 620 countries over three decades, as well as a text classifier for climate salience. Sanford et al. (2025) fine-tune an XLM-RoBERTa model on annotated party manifestos from the Manifesto Project (Version 2024a) (Lehmann et al., 2024) to detect climate-related sentences. They report an accuracy of 0.936 and F1-Score of 0.948, an improvement in 8 percentage points compared to ClimateBERT's performance on English translations of their annotated data (accuracy = 0.815, F1 = 0.867) (Sanford et al., 2025, p. 7). Their work represents, to our knowledge, the only multilingual political CC dataset and classifier publicly available to date.

The domain of climate claim detection and fact checking has also produced several public datasets, consisting of brief claims about CC, often sentence level, that are annotated for narratives or factuality. Two such datasets are ClimateFever (Diggelmann et al., 2020) and ClimateCheck (Abu Ahmad et al., 2025). ClimateFever (Diggelmann et al., 2020) is a collection of claims extracted from a "large Knowl-

[1] https://github.com/discourse-lab/DEClimate

edge Document Collection", such as encyclopediae, newspapers, and scientific publications (Diggelmann et al., 2020, p. 2). They are paired with five evidence sentences for each claim. ClimateCheck (Abu Ahmad et al., 2025) contains pairs of climate claims and abstracts from scientific publications that support or refute the claim. Claims were collected from Reddit and Twitter/X and some were created synthetically.

Other publicly available CC datasets include the data used in the development of ClimateBERT (Webersinke et al., 2022), as well as the NYTAC-CC (Grasso et al., 2024), a CC-focused subcorpus of the New York Times Annotated Corpus (licensing is required from LDC). ClimaConvo (Shiwakoti et al., 2024) is a collection of English posts on Twitter/X about CC, annotated for several discourse dimensions. Ustyianovych and Fedushko (2025) present a Ukranian dataset of Telegram messages on CC and environmental issues.

With the exception of Policlim (Sanford et al., 2025) and the AfD-CCC (Stede and Memminger, 2025), most public data on CC discourse is not explicitly drawn from political text. Furthermore, most datasets draw only on one text source. Our contribution of a comprehensive corpus of German political CC discourse seeks to address these gaps.

3 Corpus

The following sections outline our data collection process, preprocessing steps, and topic filtering method. Finally, we present an overview of the corpus and the size of each subset.

3.1 Data Collection

To construct the corpus, we collected texts from various sources in order to capture the discourse across parliamentary speech as well as different media. In order to represent the current political situation in Germany, we collect only texts authored by the parties currently represented in the German Parliament, as of the 2025 elections. These parties are the Christian Democratic Union (*Christlich Demokratische Union Deutschlands*, CDU) and Christian Social Union (*Christlich-Soziale Union in Bayern*, CSU), considered jointly as CDU/CSU, the Social Democrats (*Sozialdemokratische Partei Deutschlands*, SPD), the Green party (*BÜNDNIS 90/Die Grünen*), the Alternative for Germany (*Alternative für Deutschland*, AfD), and the Left party (*Die Linke*).

We take into account the last three German legislative periods (October 2013 to March 2025). The cut-off date for text collection was February 28, 2025. For the AfD, as they are a relatively young party and only entered into national and European parliaments later in time, parliamentary speech records could only be collected from 2014 onward for the European, and from 2017 onward for the German Parliament respectively.

Our focus lies on official party stances. We therefore only included texts which could be directly linked to official party resources (press releases and manifestos) or members of parliament (MPs). When collecting social media data, we only considered accounts directly representing the party (on a national or regional level) or an MP.

German Parliament Records of speeches held in the German Parliament were obtained from SpeakGer (Lange and Jentsch, 2023) for the time period 2013–2021, GermaParl (Blaette and Leonhardt, 2025) for the time period 2022–2023/2024, and recent speeches directly retrieved via the API of the DIP (*Dokumentations- und Informationssystem für Parlamentsmaterialien*)[2]. In total, we collected 451,210 paragraphs containing 34,341,657 tokens.

European Parliament We retrieved records of speeches held in the European Parliament primarily from ParlLawSpeech (Schwalbach et al., 2025) for the time period 2013–2024. Recent transcripts were collected via the European Parliament Open Data API[3]. This resulted in a dataset of 24,290 paragraphs over 2,497,417 tokens.

Press releases Press releases for each party were supplied by PARTYPRESS (Erfort et al., 2023) for the years 2013–2017. Texts from 2017–2025 were scraped from party websites and, in part, supplied to us directly by the parties. Before topic filtering, the press subset consisted of 64,216 paragraphs (6,355,703 tokens).

Twitter Tweets by official party and MP accounts were supplied by Lasser et al. (2022) for the time period 2016–2022. Unfortunately, more recent Tweets were not available, resulting in the Twitter subset not covering the years 2023–2025. The Twitter dataset contained 24,457 candidate paragraphs with 691,788 tokens. Due to Twitter/X privacy reg-

[2] https://dip.bundestag.api.bund.dev
[3] https://data.europarl.europa.eu/api/v2

ulations, we do not include the post text and only publish post IDs.

Telegram We collected messages from public party-led Telegram channels over the years 2019-2025 with the FROG tool (Primig and Fröschl, 2024). Table 1 lists the scraped channels. Channels were selected for scraping if they could be directly connected to the party, such as by being linked on a party or MP website, and contained recent content. We do not include channels that exclusively link to posts on other platforms with no additional content. Our sample of channels is by no means exhaustive, but covers a majority of the most active party channels.

In total, we collected 13,685 paragraphs containing 729,134 tokens. CDU and CSU are overrepresented in this dataset, with 360,660 tokens from their channels. The Greens and SPD supplied the smallest subsets, with roughly 45,000 tokens each.

Channel	Channel ID	# Mssg.
afdfraktionBB	13262765	1,589
afdfraktionimbundestag	1665003217	408
afdrlp	1498819263	756
CDU_de	1052803111	1,609
CDUCSU	1258610992	6,555
spd_ts	1757565797	383
spdaschaffenburg	1480585639	317
ruppert_stuewe	1150907951	68
gruenemahe	1412944746	687
gruenemuenchen	2099833854	99
telegruenth	2092050564	99
gruenemahlsdorf	1208683692	198
DIELINKE	1223301692	870
berlinerlinke	1139205190	47

Table 1: Overview of the Telegram channels scraped during data collection, with number of messages retrieved.

Manifestos Party Manifestos were retrieved from the Manifesto Project, Version 2025-1 (Lehmann et al., 2025). We collect manifestos of the selected parties for the years 2013–2025. Overall, the manifesto subset consisted of 7,717 paragraphs and 774,862 tokens.

3.2 Preprocessing

In order to reduce noise and ensure comparability between communication domains, we preprocess texts in three steps. Firstly, social media texts are cleaned of noise, which consists of removing emojis and extra spaces and line-breaks. Telegram messages are additionally trimmed of promotional segments repetitively found at the end of messages.

Secondly, we split texts into paragraphs. This is done primarily to ensure comparability in length of texts from different media, as parliamentary speeches and Tweets are, traditionally, of highly differing length. Furthermore, it serves to select only relevant passages from documents that may cover several topics alongside CC. Finally, it allows for texts to meet the maximum token limit of many BERT-encoders. To this end, we use semantic-text-splitter[4] with a maximum character length of 1000 to split texts into paragraphs.

Finally, we tokenize paragraphs using SpaCy's German pipeline `de_core_news_lg`[5]. Paragraphs shorter than 10 tokens in length are discarded to ensure sufficient analysis context[6].

3.3 Topic Filtering

In order to retrieve only climate-relevant paragraphs from the collected data, we pass the texts through the Klima-Klassifikator[7] (Memminger et al., 2026), a binary XGBoost classifier, which labels short texts on whether they contain the topic of climate change. This classifier is trained on short paragraphs by the AfD with binary labels for CC topic-relevance. It first counts the occurrences of climate change related keywords in each paragraph. The keywords were chosen after Schaefer et al. (2023):

> *klima* ('climate'), *erwärmung* ('warming'), *treihbaus* ('greenhouse'), *co2*, *kohle* ('coal'), *energiewende* ('energy transition'), *verkehrswende* ('transport transition'), *fridays for future*, *extinction rebellion*

Alongside each text, the keyword counts are passed into the classifier as features for prediction. The final output is a joint decision of the prediction and the keywords. If a text is classified as containing the topic of CC and contains at least one keyword, it is labeled 1, or CC-relevant. If neither or only one of these conditions are met, it is labeled 0 (or not CC-relevant).

[4] https://pypi.org/project/semantic-text-splitter/

[5] https://spacy.io/models/de#de_core_news_lg

[6] This minimum length gate may be removed in a future version.

[7] https://github.com/rmemminger/klima-klassifikator

Domain	CDU/CSU	SPD	AfD	Greens	Left	Total
German Parl.	504,393	452,821	190,236	436,020	156,032	1,739,502
EU Parl.	63,086	33,554	31,470	52,858	10,334	191,302
*Twitter**	88,010	106,943	53,672	343,233	80,050	671,908
Telegram	20,405	1,060	10,453	2,822	5,727	40,467
Press	58,170	71,825	124,931	168,766	83,548	507,240
Manifestos	9,253	9,326	2,726	38,091	19,660	79,056
Total	743,317	675,529	413,488	1,041,790	355,351	**3,229,475**

Table 2: Overview of the final dataset; given in number of tokens. * = subset can only be published in anonymized form, i.e. post ID, no text.

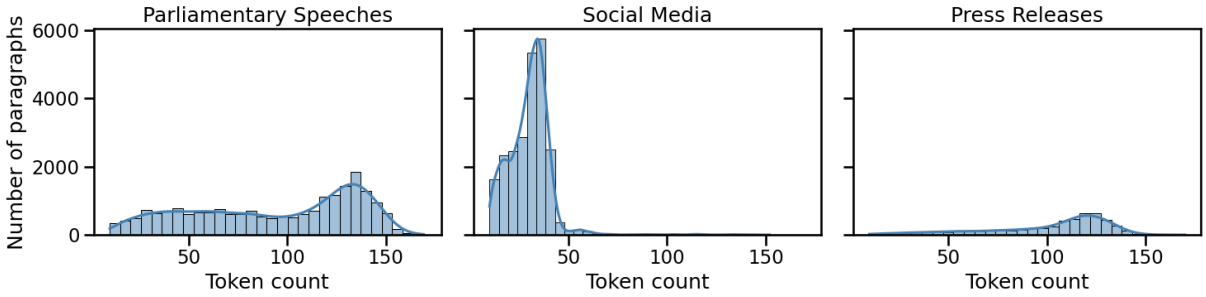

Figure 1: Distribution of tokens over the corpus, grouped by communication channel: Parliamentary speeches (German and European Parliament), Social Media (Twitter and Telegram), and press releases.

While it was trained only on AfD texts (parliamentary speeches, tweets, press releases, and Telegram messages), previous testing of the classifier on manually-annotated BT speeches from other parties showed that it can reliably generalize to other parties as well (Memminger et al., 2026). On this out-of-domain test with 140 paragraphs (14 CC-relevant, 126 non-relevant texts), the classifier achieved an F1-score of 0.93 and a macro-F1 score of 0.96. The test set was specifically chosen to have high class imbalance to measure precision, with the resulting score being 0.93 (Memminger et al., 2026).

3.4 DEClimate Corpus Statistics

After topic filtering, DEClimate contains 3,229,475 tokens, over a total of 50,091 paragraphs. Table 2 provides a detailed overview of the corpus and its subsets in number of tokens. Paragraphs from speeches held in the German Parliament (*Bundestag*) make up the largest subset of the corpus with roughly 1.7 million tokens. The smallest subsets are manifestos and Telegram messages. With over a million tokens, the Greens contribute the greatest amount of CC text.

The corpus is additionally enriched with various metadata. We include the following metadata for each text: party, ID[8], date of creation, author or speaker, tokenized text, text length in tokens, and counts for each keyword (listed in Section 3.3). We also include the source URL where applicable.

As mentioned in Section 3.2, paragraphs shorter than 10 tokens in length were removed during preprocessing. Paragraphs in the final corpus range from 10 to 170 tokens in length. Figure 1 gives an overview of the lengths of texts from different communication channels. (Manifestos are not included here.) Parliamentary speech paragraphs show the greatest variation in length, while most social media posts are less than 50 tokens long. The majority of press release paragraphs range from 100–150 tokens.

4 Conclusion & Outlook

We have presented DEClimate, a collection of paragraphs authored by German political parties or their MPs on the topic of climate change. By making this data available to the public, we hope to encourage further studies of CC discourse. With over three million tokens from different political parties

[8] In the case of social media posts, the ID represents the post ID on the respective service. Otherwise, it is a unique ID assigned by us.

and communication channels, our corpus allows for detailed analyses on paragraph and sentence levels. The paragraph length of the texts makes them compatible with state-of-the-art contextual embedding models and retains context for tasks such as claim detection and topic modeling.

Future work can expand the corpus with speeches from state parliaments (*Landtag*), where local climate policy effects may be discussed in more detail. Adding further social network data, such as posts from party and MP accounts on Bluesky, Threads, Facebook, as well as WhatsApp Channels, could add further detail to the corpus.

5 Acknowledgments

We would like to thank Dr. Cornelius Erfort for giving us access to the PARTYPRESS dataset, as well as the Manifesto Project for allowing us to include the manifestos in our corpus release. We further thank the students who helped with the data collection process: Annika Blietz, Maya Angelova, Jelle Psurek, Veronika Hentze, Zaher Alkaei, and Ian Clotworthy.

Finally, we want to thank the anonymous reviewers for their helpful feedback. Any remaining errors are our own.

References

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Andreas Blaette and Christoph Leonhardt. 2025. GermaParl Corpus of Plenary Protocols.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*.

Cornelius Erfort, Lukas F Stoetzer, and Heike Klüver. 2023. The PARTYPRESS Database: A new comparative database of parties' press releases. *Research & Politics*, 10(3).

Francesca Grasso, Ronny Patz, and Manfred Stede. 2024. NYTAC-CC: A Climate Change Subcorpus of New York Times articles. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 403–409, Pisa, Italy. CEUR Workshop Proceedings.

Kai-Robin Lange and Carsten Jentsch. 2023. SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 19–28, Ingolstadt, Germany. Association for Computational Lingustics.

Jana Lasser, Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, David Garcia, and Stephan Lewandowsky. 2022. Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, 1(4).

Pola Lehmann, Simon Franzmann, Denise Al-Gaddooa, Tobias Burst, Christoph Ivanusch, Jirka Lewandowski, Sven Regel, Felicia Riethmüller, and Lisa Zehnter. 2024. The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR) (Version 2024a). Berlin: WZB Berlin Social Science Center/Göttingen: Institute for Democracy Research (IfDem).

Pola Lehmann, Simon Franzmann, Denise Al-Gaddooa, Tobias Burst, Christoph Ivanusch, Jirka Lewandowski, Sven Regel, Felicia Riethmüller, and Lisa Zehnter. 2025. The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR) (Version 2025-1). Berlin: WZB Berlin Social Science Center/Göttingen: Institute for Democracy Research (IfDem).

Ronja Memminger, Dietmar Benndorf, and Manfred Stede. 2026. Politische Texte zum Thema Klimawandel automatisch identifizieren: Ein XGBoost Modell. In *Book of Abstracts - DHd 2026*, pages 608–609.

Florian Primig and Fabian Fröschl. 2024. Introducing the FROG tool for gathering Telegram data. *Mobile Media & Communication*, 12(2):449–453.

Mary Sanford, Silvia Pianta, Nicolas Schmid, and Giorgio Musto. 2025. Policlim: A dataset of climate change discourse in the political manifestos of forty-five countries from 1990 to 2022. *British Journal of Political Science*, 55:e131.

Robin Schaefer, Christoph Abels, Stephan Lewandowsky, and Manfred Stede. 2023. Communicating Climate Change: A Comparison Between Tweets and Speeches by German Members of Parliament. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 479–496, Toronto, Canada. Association for Computational Linguistics.

Jan Schwalbach, Lukas Hetzer, Sven-Oliver Proksch, Christian Rauh, and Miklós Sebők. 2025. Parllawspeech. (Version 1.0.0) [Data set]. GESIS, Köln.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on Twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and*

Evaluation (LREC-COLING 2024), pages 984–994, Torino, Italia. ELRA and ICCL.

Manfred Stede and Ronja Memminger. 2025. AfD-CCC: Analyzing the climate change discourse of a German right-wing political party. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 163–174, Vienna, Austria. Association for Computational Linguistics.

Taras Ustyianovych and Solomia Fedushko. 2025. Climate Change and Environmental Issues Dataset from Ukrainian Telegram Channels. Harvard Dataverse, V1.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.