The InsightsNet Climate Change Corpus (ICCC) -

Compiling a Multimodal Corpus of Discourses in a Multi-Disciplinary Domain

Elena Volkanovska[1] Sherry Tan[2] Changxu Duan[3] Sabine Bartsch[4] and Wolfgang Stille[5]

Abstract: The discourse on climate change has become a centerpiece of public debate, thereby creating a pressing need to analyze the multitude of communications created by the participants in this communication process. In addition to text, information on this topic is communicated multimodally, through images, videos, tables and other data objects that are embedded within documents and accompany the text. This paper presents the process of building a multimodal pilot corpus to the InsightsNet Climate Change Corpus (ICCC) using natural language processing (NLP) tools to enrich corpus metadata, thus building a dataset that lends itself to the exploration of the interplay between the various modalities that constitute the discourse on climate change.

Keywords: corpus; climate change; computational linguistics; annotation; metadata

1 Introduction

In recent years, the topic of climate change has taken center stage in discourses across different segments of society through different channels, media and publications. While climate scientists are in agreement that climate change is ongoing and real, debates on this topic as well as its influences on policy-makers remain highly controversial [SP21].

With the surge of published data on the topic of climate change, linguistics as well as other related disciplines have identified the study of data representing discourses on climate change as a research desiderate in order to gain a better understanding of this multidisciplinary field and the role played by a diverse set of participants with different scientific and political backgrounds who are assuming different roles and interests. In order to enable such studies, research is needed to collect and organize suitable corpora in a comprehensive and

[1] Technische Universität Darmstadt, Corpus and Computational Linguistics, Residenzschloss 1, 64283 Darmstadt, Deutschland elena.volkanovska@tu-darmstadt.de
[2] Technische Universität Darmstadt, Corpus and Computational Linguistics, Residenzschloss 1, 64283 Darmstadt, Deutschland sherry.tan@tu-darmstadt.de
[3] Technische Universität Darmstadt, Corpus and Computational Linguistics, Residenzschloss 1, 64283 Darmstadt, Deutschland changxu.duan@tu-darmstadt.de
[4] Technische Universität Darmstadt, Corpus and Computational Linguistics, Residenzschloss 1, 64283 Darmstadt, Deutschland sabine.bartsch@tu-darmstadt.de
[5] Technische Universität Darmstadt and Hessian Center for Artificial Intelligence (hessian.AI) wolfgang.stille@hessian.ai

doi:10.18420/BTW2023-59

meaningful way to inform the different communities engaging and interested in relevant discourses as well as processes concomitant with their roles as scientists, laypersons, politicians, managers and many others involved in the relevant debates and policy making processes. According to [LCJ20], the climate change related topic of global warming "has received little attention in natural language processing [NLP] despite its real world urgency". One plausible reason for this may be attributed to the lack of available corpora focusing on climate change. Additionally, as a topic - like many topics with a multidisciplinary coverage - climate change is represented in many publications not merely by means of natural language text, but also by means of a multitude of modalities such as images, maps, data tables and visualizations that are hardly captured, let alone systematically analysed for their contribution at all. So while there may be an abundant volume of digital text to be potentially included in corpora, the demonstration of textual and embedded multimodal data objects extracted and stored together in a corpus on the topic of climate change is still lacking. Therefore, the research reported in this paper aims to fill this gap by building multimodal corpora representing discourses from the domain of climate change across different genres. We furthermore set out to demonstrate some exemplary methods from corpus and computational linguistics to enrich the corpus data by metadata and annotations to allow for more in-depth analyses to further our understanding of discourses on the topic of climate change. We believe the analysis of such corpora and the study of the interlinking between the multimodal objects with its textual counterparts will create new insights into the topic of climate change and drive new discussions across various communities.

2 Overview of corpora for discourse analysis on climate change

Prior to embarking on corpus-building, we explored existing corpora and datasets that have been used in previous studies on the climate change discourse. A good overview of datasets used to investigate the debate on climate change by practitioners in the community of NLP and social sciences is provided in [SP21]; unfortunately, none of these studies takes multimodality into account. A further potentially relevant climate change dataset is the Science Daily Climate Change (SciDCC) dataset, presented in [MM21], which includes approximately 11,000 news articles scraped on the topics "Earth and Climate" and "Plant and Animals" of the Science Daily website. Yet, this is a text-only resource as well. There is a limited number of studies on the topic of climate change conducted on multimodal corpora, but these are largely combinations of texts and photographic illustrations (see [ADY11] and [We16]).

The exploration of existing corpora on the topic of climate change revealed that while they are well-suited for text-based discourse analysis, none of them can help us fully address the objective of our study, which is to analyse the climate change discourse as an interaction between various modalities. The corpora that we inspected do not store data objects of different formats in a single corpus in a manner that lends itself to the study of the interplay between a document's text and any multimedia content embedded in it. In addition, existing

multimodal corpora take into consideration a set number of media types, which does not allow for the exploration of the range of embedded media types. Rather than moulding our research to fit the data that was readily available at the time this study began, we decided to build a multimodal corpus from authentic data that would allow us to examine (1) the type of modalities embedded in a document, and (2) how different modalities contribute to the discourse on climate change.

3 Developing a pilot corpus

The pilot corpus described in this section is a precursor to ICCC. The objective is to explore the possibilities of developing a multimodal corpus on climate change and to systematically learn more about the challenges before expanding it. At the onset of the corpus-building process for the pilot corpus two main criteria were devised: the corpus had to contain content in both English and German, and any collected multimedia content had to be embedded in the document. We refrain from incorporating stand-alone collections of single-modality data such as collections of images or photos etc. We did not set a limit on the types of multimodal data to be collected with the expectation that we will encounter data objects beyond images and videos. Beyond this, we adhere to a fairly standard corpus-design procedure, which includes the following steps: (1) identify genres of interest and data sources that contain suitable content; (2) contact copyright holders to obtain their approval to collect and use the data; (3) define metadata properties to store relevant information; (4) collect the data from each data source, (5) parse it in a project-specific corpus structure.

3.1 Identifying genres, data sources, and obtaining copyright permissions

The objective in this step was to ensure that each genre included in the corpus represents various entities or members of society that actively take part in the public discourse on climate change. The pilot corpus entails content from three genres: academic papers on climate change, reports published by the International Panel on Climate Change (IPCC), and content published on the websites of Greenpeace International and Greenpeace Germany (Non-Governmental Organisations (NGOs)). Academic papers can be found either under a free open access (OA) policy, which does not require specific copyright permissions, or hidden behind a paywall, in which case the rules for content use are governed by the specific publisher. IPCC reports can be downloaded from the official website of IPCC[6] and used for personal, non-commercial purposes as long as the source is duly acknowledged. Translation of IPCC reports into German is managed by the German IPCC Coordination Office[7] and the translated content can be retrieved from their website. Content published on the two Greenpeace websites posed the most complex copyright case, mostly because of the

[6] https://www.ipcc.ch/
[7] https://www.de-ipcc.de/index.php

different copyright rules applicable to text on the one hand, and multimedia content on the other. Greenpeace has granted us approval to use images and videos that have been created by and are sole property of Greenpeace, as long as the content is used for research purposes [Gr22a, Gr22b] only.

3.2 Developing and implementing a metadata scheme

Metadata support corpus management and exploitation and constitute an integral part of linguistic research. They can be retrieved from the content description provided by the publisher, or obtained through data post-processing, including linguistic processing and information extraction. The metadata framework for the pilot corpus uses properties from the Dublin Core Metadata Initiative (DCMI Metadata Terms) as its backbone. We opted for the DCMI framework because it provides descriptive terms for data objects of different formats and constitutes a widely acknowledged standard that has been used in the description of both web and physical collections. This allows us to use the same schema for digitised collections which were not primarily designed to serve as web content.

At the time of the property selection, DCMI Metadata Terms entailed 55 properties [Du20], accompanied by a set of datatypes and vocabulary encoding schemes for the description of digital resources of various formats (including image, video, and audio). We selected 14 DCMI metadata terms: title, type, subject, publisher, contributor, identifier, rights, format, bibliographicCitation, rightsHolder, license, extent, created, accrualMethod. For a more detailed description of each term please see [Du20]. This information should be retrievable for each document in the corpus.

While the DCMI Metadata Terms provide a good selection of descriptive elements, they do not include fields for encoding all information of relevance to the project. Two containers of metadata properties were added to address this shortcoming: *linguisticInformation* and *mediaInformation*. The former is a container for project-relevant linguistic information gathered from both the given metadata, that is, metadata provided by the publisher, and for metadata derived by performing linguistic processing on the corpus. The latter gives information about the number and type of multimedia data objects embedded in a document. The two metadata containers are flexible and more properties can be added as necessary. At the moment, *linguisticInformation* stores information about genre, language, text type, status of content (archived or not, for more information see section 4.3), number of tokens, number of words, word types, content words, type-token ratio, lexical density, information about sentence, word, and token length, named entities and abbreviations. Each document was given a *filename* according to an agreed workable convention so that various media types can be linked to the document in which they are embedded. The intention is to apply this scheme to each document collected from the three data sources described in section 3.1. The collected metadata is added to each document and helps us build a profile of the whole corpus.

As already mentioned, some of the metadata properties are obtained by conducting linguistic processing and annotation of the content, which at the moment entails tokenization, part-of-speech (POS) tagging, dependency parsing, named entity recognition (NER), and abbreviation extraction.[8]

4 Data collection

This section elaborates on the data collection process from three sources: academic papers, the website of the International Panel on Climate Change (IPCC), and the content published on the websites of Greenpeace International and Greenpeace Germany.

4.1 Academic Papers

As a starting point for data collection for the pilot corpus, we used an article published by CarbonBrief titled "The most influential climate change papers of all time" [Pi15]. In this article, eight academic writings [AH97, Ca38, MW67, Ke76, No91, GZ00, HS06, HSR12] were highlighted as the most "cited" papers, which is a measure and an indication of how much impact the paper has in the scientific world. These seed papers ranged between the years 1896 to 2012, giving us a wide range of different climate change perspectives as the topic has evolved over time. We coined these eight papers as "seed papers" and these papers provided a way for us to extract information from them that would link us to other related academic works along the same topics across different years, providing a way for us to build a more comprehensible corpus.

Building a corpus with the seed papers We explored two methods for building a corpus using the eight seed papers that we have obtained: (1) checking the overlap of references between the seed papers, (2) extracted keywords and keyphrases from the academic papers were used as *seed terms* for search of more academic papers in the similar topic in Google Scholar.

With the first approach, we were not able to find any overlap between the references of the seed papers. Therefore, we did a search on Dimensions[9] for a list of the top citation references for each seed paper and from there we looked for overlapping citations. If a paper referenced to at least two seed papers, then that paper was taken to be included in the corpus. Based on this method, a total of 84 papers were initially collected.

[8] Any metadata obtained through content post-processing will be affected by the choice of tool used to perform this process. This paper demonstrates how such tools can be applied for metadata enrichment and does not focus on ways of improving their performance.

[9] https://www.dimensions.ai/

The second method was based on information extraction. The text content of the seed papers was extracted and analyzed with KeyBERT[Gr20]. KeyBERT provides integration of different pre-trained language models and since we only have academic papers in the English language, we opted for the model[10] developed initially by [RG19].

The top 10 keywords/keyphrases from each paper were extracted and these were grouped together according to semantic similarity. After the first iteration of extracting the keywords/keyphrases from the seed papers and grouping them together, a total of 9 clusters of keywords were formed. Each cluster of keywords was used as seed terms to search Google Scholar with *AND* operator between the terms and the top 20 results were taken and added to our collection. This iterative process was completed when we evaluated the corpus and found that we had obtained 1,812 academic papers using this method (see figure 1 for visualization of the process). The total number of academic papers downloaded was 1887, ranging from the years 1895 to 2022.

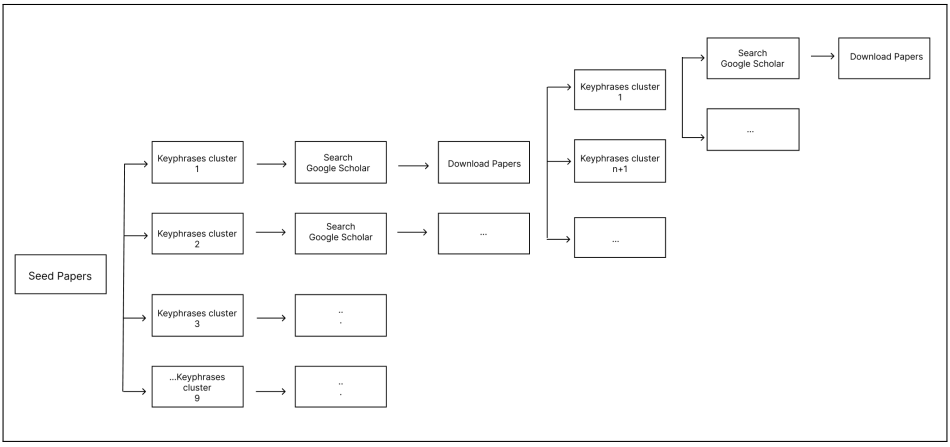

Fig. 1: An example of the iterative process for keywords extraction from seed papers to downloading the papers from Google Scholar.

4.2 Reports published by the International Panel on Climate Change (IPCC)

In the pilot corpus, we included the IPCC synthesis reports from each reporting period[11] [12]. These reports are originally published in English; translations into German were collected

[10] all-MiniLM-L6-v2

[11] "Climate Change: The IPCC 1990 and 1992 Assessments", "SAR Climate Change 1995: Synthesis Report", "TAR Climate Change 2001: Synthesis Report", "AR4 Climate Change 2007: Synthesis Report", "AR5 Synthesis Report: Climate Change 2014".

[12] At the time of writing this paper, the publication of the AR6 Synthesis Report is pending; only an outline of the report is available.

when available[13]. This means that the pilot corpus contains 5 synthesis reports in English and 3 synthesis reports or part of the synthesis report in German.

4.3 Greenpeace International and Greenpeace Germany

The web pages from Greenpeace International and Greenpeace Germany relevant to our project were retrieved by entering the prompt 'climate change' and 'Klimawandel' respectively in the search bar on each organisation's web site[14] [15]. The search, performed in March 2022, returned 4057 links to web pages from Greenpeace International, of which 698 were hosted on the domain of Greenpeace International, while 3359 were archived and hosted on the domain of the Wayback Machine - Internet Archive[16]. We only include the 698 web pages hosted on the Greenpeace International domain in our pilot corpus in order to have a balanced number of tokens in each language (see table 1 for corpus size). From Greenpeace Germany, the search returned 1281 links to web pages.

5 Data parsing

The process described in section 4 resulted in files in two formats: PDF and HTML. This section discusses the tools used to parse the documents and extract relevant information.

5.1 Academic papers and IPCC reports

All academic papers and IPCC reports are saved and parsed as PDF files. For the scanned versions of PDF files, we use Tesseract [Ka07] to do OCR on the text and append the results as transparent text layers to the original PDF pages.

We combine the VILA [Sh22] and Resnet101 [He15] that was trained on DocBank [Li20] to parse PDF Files. VILA is a model for token sequence prediction, which does not predict the images in the document. Resnet101 takes as input the rendered image of each page of the document and identifies only the location of the figures on each page. The output label set is *Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table* and *Title*.

The models for parsing PDF files are run in an Online Learning [Ho18] framework and are loaded in Label Studio [Tk22] as a machine learning backend service. We import the

[13] At the moment, there are full translations of the synthesis reports for the years 2007 and 2014, and a translation of the Summary for Policymakers from 2001.
[14] https://www.greenpeace.org/international/
[15] https://www.greenpeace.de
[16] https://archive.org/web/

documents as a rectangular label object detection task into the front end, use the original models to make predictions for a small subset of documents, correct the prediction manually, and then fine-tune the models. Finally, we parse all documents using the updated models. The goal is not only to extract the text and other data objects from the PDF files, but also to retain the layout information of the documents so that we can examine the interactions between the data objects.

5.2 Greenpeace International and Greenpeace Germany

Since many of the webpages that we needed to download and parse were dynamic, we used Selenium[17] to retrieve the HTML from the collected links (see section 4.3) and BeautifulSoup [Ri07] to parse the content. When extracting the relevant data objects, we made sure to preserve the order of appearance of HTML elements containing important information, to extract their position on the web page, and to extract all HTML elements that might contain links to data objects in a modality different than text. This resulted in documents that mirror the output of the PDF parsing process described in section 5.1, hence allowing us to examine the interaction between various modalities.

6 Data Annotation

This section will highlight some of the methods used to annotate the parsed data. We will mainly discuss the annotation process for: (1) linguistic annotations including sentence splitting and tokenization, part-of-speech tagging and dependency parsing, (2) Named-entity recognition and (3) keywords/keyphrases extraction.

6.1 Linguistic annotation and Named-Entity Recognition (NER)

The linguistic annotation was done in a bottom-up approach: the text of a document was split into sentences, which were then run through an annotation pipeline. Three libraries constitute the annotation pipeline for English texts: spacy-stanza[18], Stanford CoreNLP [Ma14][19], and SciSpacy[Ne19][20]. Stanford CoreNLP is used to complement the named entity extraction for categories of named entities not covered by spacy-stanza. The extracted named entities from the English texts belong to the following 24 categories: PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK-OF-ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL, TITLE, CITY, IDEOLOGY,

[17] https://github.com/SeleniumHQ/selenium, v.3.14.0
[18] https://spacy.io/universe/project/spacy-stanza, running on stanza language model 1.4.1
[19] Version 4.4.0
[20] https://github.com/allenai/scispacy

RELIGION, CRIMINAL-CHARGE, and CAUSE-OF-DEATH. With SciSpacy we extract abbreviations from each sentence. The German documents were processed with stanza only, since both stanza and Stanford CoreNLP have only four categories of named entities for German language texts (ORG, PERSON, LOC, MISC).

In addition to saving the extracted annotations in a JSON file, each document with linguistic annotations is saved as a pickle file (German content) and both pickle file and spaCy object (English content). This step allows for consistency should we decide to extract linguistic patterns or another type of linguistic information.

The linguistic annotation served as the backbone of the linguistic information extracted and calculated for each document. The result of this process feeds back into the metadata, where the information is saved in the metadata container *linguisticInformation* as mentioned in section 3.2.

6.2 Keywords/Keyphrases extraction

As previously mentioned in section 4.1, KeyBERT was implemented to capture keywords and keyphrases that are semantically similar to the document content. The same approach was used to annotate the documents in our corpus. Since our corpus contains documents both in English and German, using pretrained language models in English is not enough. Therefore, a multilingual model [21] developed by [RG20] was used for German texts.

Through our implementation and experimentation of using KeyBERT for identifying seed terms as seen in section 4.1, we found that KeyBERT has the tendency to create noisy results, which did not have much effect on our results when using them as seed terms to search on Google Scholar, but would potentially have a much larger effect on keywords/keyphrases annotation of the data. Therefore, we combined the KeyBERT approach with the textrank approach proposed by [MT04]. We implemented textrank using PyTextRank [Na16] in the spaCy pipeline. Both English[22] and German[23] models were used through spaCy.

The keywords and keyphrases that had a semantic similarity score of 0.7 or higher were extracted using KeyBERT and the list was compared to the set that were extracted using PyTextRank; overlapping results were discarded and the final list of keywords and keyphrases was added to the metadata term *subject*.

7 Results

We present the results obtained from the data curation process of the pilot corpus and the type of multimedia data objects retrieved from each of the three data sources.

[21] paraphrase-multilingual-MiniLM-L12-v2
[22] en_core_web_sm and en_core_web_trf
[23] de_core_news_sm and de_dep_news_trf

Academic papers A total of 1,887 academic papers were collected and 15,461 images and figures were extracted from the PDFs. 1,095 equations and 2,207 tables were extracted and these are listed under "Other" in table 1. A total of over 43 million tokens were extracted from these documents.

IPCC reports The 5 English IPCC reports contained 315 images and figures and 104 tables/equations. A total of 496,477 tokens were extracted. For the 3 reports in German, 135 images and figures were extracted with 31 tables/equations and a total of 170,617 tokens were extracted.

Greenpeace International and Greenpeace Germany The 698 documents of Greenpeace International have 2066 embedded images, 123 embedded videos, 67 videos added to the content as hyperlinks, and 458 other types of multimedia objects. In 1, "Other" entails iframes, which are web pages embedded within another web page. In the context of the Greenpeace International corpus, iframes store videos, text, images, animations, dynamic charts, tweets, Facebook posts, Instagram posts, and files in a PDF format. The 1281 documents of Greenpeace Germany contained 2463 images and 14 videos. We retrieved 188 YouTube videos from Greenpeace International, whose total duration was 25.55 hours. The 14 videos of Greenpeace Germany amounted to 3.35 hours. Total number of tokens in the Greenpeace International transcripts were 157,115 and 27,586 tokens for Greenpeace Germany.

It is evident that of the total number of collected documents (3,874), the majority have embedded multimedia content (3,317), compared to text-only documents (557). This finding underpins the need for awareness of the various media types that support textual content, and for incorporating processing techniques that would enable researchers to analyse media content in the context of a document as a whole.

Data Source	Docs without MC*	Docs with MC*	Multimedia Content			# of Tokens
			Imgs/Figs	Videos	Other	
Academic Paper	100	1 787	15 461	-	3 302	43 152 714
IPCC reports EN	0	5	315	-	104	496 477
IPCC reports DE	0	3	135	-	31	170 617
Greenpeace International	228	470	2 066	188	458	676 879
Greenpeace Germany	229	1 052	2 463	14	463	645 962

*MC: Multimedia Content

Tab. 1: Summary of the pilot corpus with number of extracted contents.

8 Discussion

This paper describes the process of building a multimodal pilot corpus comprising both a substantial number and a wide range of data types in documents. The pilot corpus was the starting point for developing methodologies that would allow us to better design and curate the ICCC. We believe such a multimodal dataset is needed as a starting point in order to gain further insights to the climate change topic by analyzing multimodal data and exploring the additional information that can be obtained.

Possible use-case One example use-case of such analysis can be the study of political views on a specific policy. With textual information, one can analyze the textual data with sentiment analysis to extract the sentiments regarding the policy. With the addition of multimodal data, we can also extract the sentiment in those data objects and determine if they play a role in strengthening the sentiment found in the textual data or not. Such insight can lead to a deeper understanding of the views and opinions concerning the specific policy.

Lessons learned We present some of the lessons learned in terms of data modelling of multimodal corpora, application of data annotation and information retrieval techniques, and challenges of working with a bilingual corpus.

It is evident that discarding multimedia data objects, as is common practice in corpus development, results in the possible loss of relevant information and eliminates the opportunity to investigate the interaction between data objects of different modalities. As seen in this paper, creating a data model that lends itself to modelling and analyzing interactions between different types of data objects presents another layer of complexity in the process of collecting, parsing, and analysing multimodal data.

Another important task was to enrich the corpus metadata by incorporating NLP tools for corpus annotation and information retrieval. While some of these techniques evidently enriched the metadata of the corpus, others performed well on one type of data, but not on another, highlighting the necessity of employing tools or models built for specific task in the specific target domain. More work will need to be done in this respect to seek out the appropriate tools and models and fine-tuning them to our domain-specific documents.

Lastly, we found that it is difficult to achieve entirely matching annotations for English and for German corpora, mostly because existing tools and models for processing German texts do not offer the same level of granularity and linguistic detail. Improving this situation is identified as a research desiderate in our future work.

As stated previously, the goal of this paper on the creation of ICCC pilot corpus is to explore the design and curation process of a multimodal corpus. The impact for future research is to provide a corpus-building methodology that will be applied in the next step of the research,

which is to expand the ICCC by looking further into collecting data from other sources in the climate change domain.

Acknowledgments We would like to thank Rasmus Beckmann and Jasper Korte of the Institute for Software Technology, German Aerospace Centre (DLR) for their assistance in identifying sources of data relevant to our research.

The research reported is this paper was conducted within the research project InsightsNet (`https://insightsnet.org/`) which is funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01UG2130A.

Bibliography

[ADY11] Anne DiFrancesco, Darryn; Young, Nathan: Seeing climate change: The visual construction of global warming in Canadian national print media. cultural geographies, 18(4):517–536, 2011.

[AH97] Arrhenius, S.; Holden, Edward S.: ON THE INFLUENCE OF CARBONIC ACID IN THE AIR UPON THE TEMPERATURE OF THE EARTH. Publications of the Astronomical Society of the Pacific, 9(54):14–24, 1897.

[Ca38] Callendar, Guy Stewart: The artificial production of carbon dioxide and its influence on temperature. Quarterly Journal of the Royal Meteorological Society, 64(275):223–240, 1938.

[Du20] DublinCore: DCMI Metadata Terms. 2020.

[Gr20] Grootendorst, Maarten: , KeyBERT: Minimal keyword extraction with BERT., 2020.

[Gr22a] Greenpeace, International: Email to Elena Volkanovska, 13 April. 2022.

[Gr22b] Greenpeace, International: Email to Elena Volkanovska, 15 March. 2022.

[GZ00] Guisan, Antoine; Zimmermann, Niklaus E: Predictive habitat distribution models in ecology. Ecological modelling, 135(2-3):147–186, 2000.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: , Deep Residual Learning for Image Recognition, 2015.

[Ho18] Hoi, Steven C. H.; Sahoo, Doyen; Lu, Jing; Zhao, Peilin: , Online Learning: A Comprehensive Survey, 2018.

[HS06] Held, Isaac M; Soden, Brian J: Robust responses of the hydrological cycle to global warming. Journal of climate, 19(21):5686–5699, 2006.

[HSR12] Hansen, James; Sato, Makiko; Ruedy, Reto: Perception of climate change. Proceedings of the National Academy of Sciences, 109(37):E2415–E2423, 2012.

[Ka07] Kay, Anthony: Tesseract: An Open-Source Optical Character Recognition Engine. Linux J., 2007(159):2, jul 2007.

[Ke76] Keeling, Charles D; Bacastow, Robert B; Bainbridge, Arnold E; Ekdahl Jr, Carl A; Guenther, Peter R; Waterman, Lee S; Chin, John FS: Atmospheric carbon dioxide variations at Mauna Loa observatory, Hawaii. Tellus, 28(6):538–551, 1976.

[LCJ20] Luo, Yiwei; Card, Dallas; Jurafsky, Dan: Detecting stance in media on global warming. arXiv preprint arXiv:2010.15149, 2020.

[Li20] Li, Minghao; Xu, Yiheng; Cui, Lei; Huang, Shaohan; Wei, Furu; Li, Zhoujun; Zhou, Ming: , DocBank: A Benchmark Dataset for Document Layout Analysis, 2020.

[Ma14] Manning, Christopher D; Surdeanu, Mihai; Bauer, John; Finkel, Jenny Rose; Bethard, Steven; McClosky, David: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60, 2014.

[MM21] Mishra, Prakamya; Mittal, Rohan: NeuralNERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction. In: Tackling Climate Change with Machine Learning Workshop at ICML. 2021.

[MT04] Mihalcea, Rada; Tarau, Paul: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 404–411, 2004.

[MW67] Manabe, SvUkURO; Wetherald, Richard T: Thermal equilibrium of the atmosphere with a given distribution of relative humidity. 1967.

[Na16] Nathan, Paco: , PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents, 2016.

[Ne19] Neumann, Mark; King, Daniel; Beltagy, Iz; Ammar, Waleed: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, Florence, Italy, pp. 319–327, August 2019.

[No91] Nordhaus, William D: To slow or not to slow: the economics of the greenhouse effect. The economic journal, 101(407):920–937, 1991.

[Pi15] Pidcock, Roz: The most influential climate change papers of all time. 2015.

[RG19] Reimers, Nils; Gurevych, Iryna: 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, 11 2019.

[RG20] Reimers, Nils; Gurevych, Iryna: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2020.

[Ri07] Richardson, Leonard: Beautiful soup documentation. Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018], 2007.

[Sh22] Shen, Zejiang; Lo, Kyle; Wang, Lucy Lu; Kuehl, Bailey; Weld, Daniel S; Downey, Doug: VILA: Improving structured content extraction from scientific PDFs using visual layout groups. Transactions of the Association for Computational Linguistics, 10:376–392, 2022.

[SP21] Stede, Manfred; Patz, Ronny: The climate change debate and natural language processing. In: Proceedings of the 1st Workshop on NLP for Positive Impact. pp. 8–18, 2021.

[Tk22] Tkachenko, Maxim; Malyuk, Mikhail; Holmanyuk, Andrey; Liubimov, Nikolai: , Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio.

[We16] Wessler, Hartmut; Wozniak, Antal; Hofer, Lutz; Lück, Julia: Global multimodal news frames on climate change: A comparison of five democracies around the world. The International Journal of Press/Politics, 21(4):423–445, 2016.