NYTAC-CC: A Climate Change Subcorpus based on New York Times Articles

Francesca Grasso[1,*,†], Ronny Patz[2,†] and Manfred Stede[2,†]

[1]University of Turin, Corso Svizzera 185, 10149, Turin, Italy
[2]University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476, Potsdam, Germany

Abstract

Over the past decade, the analysis of discourses on climate change (CC) has gained increased interest within the social sciences and the NLP community. Textual resources are crucial for understanding how narratives about this phenomenon are crafted and delivered. However, there still is a scarcity of datasets that cover CC in *news media* in a representative way. This paper presents a CC-specific subcorpus of 3,630 articles extracted from the 1.8 million New York Times Annotated Corpus, marking the first CC analysis on this data. The subcorpus was created by combining different methods for text selection to ensure representativeness and reliability, which is validated using ClimateBERT. To provide initial insights into the CC subcorpus, we discuss the results of a topic modeling experiment (LDA). These show the diversity of contexts in which CC is discussed in news media over time.

Keywords

Climate Change, Corpora, Topic Modeling

1. Introduction

We present NYTAC-CC, a topic-specific subcorpus with 3,630 articles addressing climate change (CC), derived from the *New York Times Annotated Corpus*. This subcorpus covers a 20-year period, drawing from NYTAC's collection of 1.8 million articles published between 1987 and 2007, which is available through the *Linguistic Data Consortium*. The original corpus, and thus also the subcorpus, includes a variety of metadata such as the 'desk' (the newspaper branch) and both manually- and automatically-labeled content categories, with many articles also featuring hand-written summaries. The extensive use of NYTAC in NLP research over the last 15 years (e.g., [1, 2]) benefits CC researchers, allowing for detailed historical analysis of CC discussions in news media. This includes exploring how CC debates were interwoven with topics like domestic and foreign policy, science reporting, and arts and culture coverage. Unlike other CC-focused resources that often contain shorter documents, the NYTAC-CC subcorpus offers a diverse array of articles with varying lengths and complex content, making it a unique resource for investigating the evolution of CC narratives over time.

The contribution of this paper is threefold:

(*i*) We present the NYTAC-CC subcorpus and its construction using blending of dictionary-based and supervised methods in order to ensure *representativeness* as well as *validity* and *reliability*, which are key in social science research [3]. This hybrid approach addresses the challenges of refining a topic-specific subcorpus from a larger corpus, aiming to mitigate the limitations of traditional keyword-based sampling that often results in false positives.

(*ii*) To demonstrate the validity of the subcorpus, and thus its reliability for further downstream tasks, we illustrate the results of a classification experiment using ClimateBERT [4]. While this experiment further validates that the articles in our NYTAC-CC subcorpus are, indeed, true positives, it also shows limitations of ClimateBERT. As ClimateBERT falsely classifies a number of true positives from our subcorpus as (false) negatives, we demonstrate that our approach achieves better results in ensuring recall of relevant CC articles from the NYTAC corpus.

(*iii*) To gain initial insights into the CC subcorpus coverage, we use keyword analysis and topic modeling (specifically LDA) to track specifics of CC reporting over the 1987-2007 time span. The results show important trends over time, including key periods of reporting and a large variety of contexts in which CC is discussed.

Thus, our goal is to provide a substantively new and relevant subcorpus, developed and validated in multiple iterations, and to then provide a first overview of the NYT's coverage of climate change during the time period covered in our corpus. Although several studies have explored U.S. print media's reporting on anthropogenic CC, we cover an important 20-year period in which much of today's climate change discourse evolved.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

†These authors contributed equally.

fr.grasso@unito.it (F. Grasso); ronny.patz@uni-potsdam.de (R. Patz); stede@uni-potsdam.de (M. Stede)
0000-0001-8473-9491 (F. Grasso); 0000-0002-0761-086X (R. Patz); 0000-0001-6819-2043 (M. Stede)

2. Related Work: CC in News

Despite the growing interest in addressing climate change among various academic communities, as pointed out by Luo et al. [5], the topic has so far received limited attention within the 'core' NLP community. This is largely due to the NLP field's focus on standardized datasets and shared tasks, where the topic of CC has been scarcely addressed.

Efforts can be observed within the context of social media, with datasets made available for CC-related tasks [6, 7]. However, there remains a scarcity of work addressing CC at the news article level, which is essential for the NLP community investigating CC narratives in media or performing downstream tasks involving longer texts. In contrast, the analysis of CC discourse on both social media and traditional media has been extensively studied in various social science disciplines [8, 9]. In the following, we will focus on prominent work targeting traditional news media.

A widely-cited early study by Trumbo [10] examined the framing techniques used by various "claim makers" in the online editions of five U.S. newspapers. After querying with different terms and manually filtering the results, the remaining articles were thoroughly investigated. Boykoff [11] later studied the "claims and frames" issue in a similar manner. Legagneux et al. [12] conducted a comparative study of scientific literature and press articles to investigate coverage differences between CC and biodiversity. They analyzed materials from the USA, Canada, and the United Kingdom spanning 1991 to 2016, using representative keywords to query and retrieve relevant content. Similarly, [13] examined how journalistic norms affected CC reporting in U.S. TV and newspapers. Other studies examined the frequency of CC mentions, or the 'attention cycle'. Brossard et al. [14] compared CC reporting between the NYT and the French *Le Monde*. Grundmann and Krishnamurthy [15] analyzed newspapers from four countries, enhancing article counts with word frequency and collocation analyses using corpus-linguistic tools, where the outcomes are manually interpreted. The work of [16] highlights one of the few instances where NLP technology is used to analyze CC in newspapers, where authors applied supervised classification to construct a corpus and identify frame categories within four U.S. papers. Continuing in the NLP domain, [4] utilized a specialized corpus that includes CC-related news articles, though details on data retrieval are not available. [17] compiled a dataset of 11k news articles from *Science Daily* through web scraping.

In conclusion, there remains a scarcity of available corpora containing larger text units like entire articles, which are essential for the NLP community investigating CC narratives in traditional media or performing various downstream tasks involving news articles.

3. Building the NYTAC-CC

3.1. Challenges in CC Text Selection

The New York Times Annotated Corpus (LDC release)[1] contains 1,855,658 articles (1987-2007), each formatted as a single XML file. Metadata include date, author, and newsroom desk. Articles are manually annotated with locations, people, organizations, and key topics. However, topic labels are generally not sufficient for our purpose, that is, finding all CC-related articles, because (i) not all articles are labeled; (ii) some labels of potentially CC-relevant text are overly broad, e.g., 'weather,' which also encompasses many non-CC topics; and (iii) some articles we consider CC-relevant are tagged with labels that do not relate to CC.

Our goal is to design a retrieval method that not only ensures *validity* and *reliability* but also emphasizes *representativeness*, ensuring that the corpus adequately covers content related to the specific subject it aims to represent. Traditional approaches, such as the use of keywords or n-grams, can be inadequate if used alone and can lead to misclassifications due to both false positives and false negatives. Crucially, this holds even with advanced models, particularly when tasked with processing large linguistic units such as entire articles [18]. The changing use of language in time-spanning corpora can further challenge single-method approaches, since they must handle texts that, although consistent in topic, may cover the phenomenon in varied ways over time.

Moreover, we aim for an approach that is reproducible, i.e., that can also be applied to other corpora that do not come with this type of metadata. We have therefore opted for a hybrid approach that combines the advantages of both keyword-based methods and automatic classification, while also aiming to overcome the weaknesses of both.

3.2. Our Hybrid Approach

Our subcorpus construction is built on text retrieval methods previously used in studies on CC discourse (see, e.g., Section 2), but merges them into a hybrid approach to address their strengths and weaknesses. In the literature, we identified the following approaches:

1. Search with bigrams: typically, this involves terms like "climate change," sometimes accompanied by one or two others, notably "global warming" and "greenhouse effect"; e.g., [10, 12];

2. Search with a longer list of keywords, followed by manual filtering; e.g., [19, 18];

[1] https://www.ldc.upenn.edu

3. Complex Boolean queries with keywords and operators (AND, OR, NOT); e.g., [20];

4. Manual annotation of training data followed by supervised classification; e.g., [16].

As a first exploratory step, we experimented with method (1), obtaining the expected unsatisfactory results. We subsequently refined our retrieval process from the NYTAC by extending methods (2) and (4). Texts that we consider relevant for the CC topic must not only merely mention CC in passing, but should discuss aspects of anthropogenic CC, relate substantial information, or convey a stance on its existence or urgency.

Bigram search. Initially, we experimented with a list of bigrams (see Appendix A) sourced from the BBC Climate Change Glossary[2]. This was done to cover terminologies used over the two decades spanned by the corpus. This method led to the retrieval of 10,707 articles. Upon manual inspection, we found that many were false positives, addressing general environmental issues but not specifically related to CC. Conversely, many articles we regarded as relevant did not contain the bigram "climate change" (searching for this bigram yielded only 2,080 texts). Consequently, this led us to seek a more elaborate approach.

Keyword search. In response to the limited performance of the bigram search, we proceeded to extract CC-related articles using keywords that were employed by [19] to identify topic-relevant articles in *Nature* and *Science* (see Appendix B). To these, we added the keyword "Kyoto", given the specific time period of our corpus where the Kyoto conference had a similar importance as later the "Paris agreement". However, the resulting subcorpus still contained many false positives, primarily from long list-like articles combining various news items. To ensure homogeneity, we excluded these articles, resulting in an intermediate corpus of 12,883 articles.

Text ranking and supervised classification. To overcome the presence of false positives, we implemented an additional, more elaborate filtering step on the intermediate corpus. Initially, we heuristically ranked the articles for topic relevance, using a score based on accumulated keyword weights. This score reflects both the frequency of the keywords and their position within the article, as content in the beginning is generally considered most important. Specifically, we multiply the number of keyword occurrences per sentence by a score representing sentence prominence (1 for the first sentence, 0.9 for the second, 0.8 for the third, and so on). After automatically ranking the articles, we selected 450 articles for manual tagging: the top 150, the last 150, and 150 from the middle. We manually assessed them to determine if they were at least partially about CC, using the labels '1' (CC-related) or '0' (not CC-related).

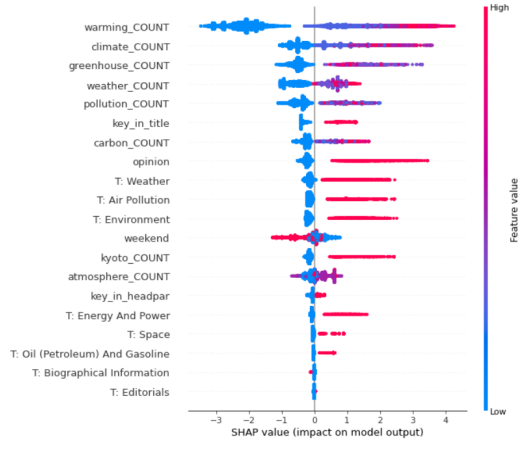

Figure 1: Key features in classifying "climate change" articles

We used the manually-annotated data to train and test an XGBoost classifier, configured to differentiate between CC-related and non-CC articles. The features used included keyword counts, (those from [21], plus 'Kyoto'), the 50 most frequent 'topic' labels from the article metadata, and several binary features: whether an article was published by (i) the 'Dining' or 'Style' desks or by (ii) other desks; whether it was published on the weekend; whether a keyword appeared in the title or the first paragraph; and whether the article was (i) an opinion piece or a letter versus (ii) another type of article. The classifier achieved a precision score of 1.0 and a recall score of 0.94 on our held-out evaluation set of 100 texts. Subsequently, we used the classifier to label the entire intermediate corpus, labeling 9,253 articles as not CC-related and 3,630 CC-related, thus forming what we now refer to as our final 'NYTAC climate change subcorpus' and make available as the list of document IDs.[3] Figure 1 illustrates the features that had the greatest impact on the classification decisions.

3.3. Evaluation with ClimateBERT

We aim to demonstrate (i) the relevance of our 3,630-article subcorpus in genuinely consisting of climate change (CC)-related articles and, thereby, (ii) the validity of our combined method for retrieving topic-consistent texts from a larger, heterogeneous collection while minimizing false positives. To perform that validation, we employed ClimateBERT, specifically *ClimateBert$_F$* [4], a BERT-based model trained on CC-related texts. In particular, we used *distilroberta-base-climate-detector* from the

[2]https://www.bbc.com/news/science-environment-11833685

[3]https://github.com/discourse-lab/NYTAC-CC

Hugging Face platform[22], a fine-tuned version with a classification head for detecting climate-related paragraphs. Given its specialization in CC-related texts, we deemed ClimateBERT a very suitable tool to confirm the accuracy of our dataset. In doing so, we are also indirectly assessing the model's capability in detecting CC-related content within larger portions of texts. As the model's context length is limited to 512 tokens, we addressed this limitation by adopting two different approaches described below.

In the first approach, longer texts were truncated due to the model's limited context length. Of the 3,630 instances, the model recognized 3,468 articles as +climate. We manually inspected the remaining 162 texts classified as -climate, i.e., as false negatives. We found that the model clearly misclassified 75 texts, which included relevant CC content appearing beyond the initial 512 tokens. More qualitative insights on these 162 texts are provided in the subsection below.

In addition, we attempted a second approach to overcome the context length constraint by using a sliding window technique. This involved creating chunks of longer texts (> 512 tokens), classifying each chunk, and labeling the entire text as +climate if any of the chunks were labeled as such. This second approach led to significantly different results, as only 3 out of 3,630 instances were labeled -climate.

These results demonstrate both the representativeness of our corpus and the validity of our hybrid subcorpus selection method. In addition, we show how automatic classification models can be limiting when dealing with long text units, therefore reinforcing the need for a combined approach to build topic-relevant (sub)corpora.

3.4. Analysis of the ClimateBERT misclassifications

As discussed in Section 3.3, we manually inspected 162 articles that ClimateBERT initially classified as false negatives within our subcorpus. Of these, 75 were clearly related to CC. Specifically, 48 articles featured significant discussions on CC-related issues beyond the model's 512-token limit. Additionally, 27 articles contained detailed CC narratives within the first 512 tokens, often intersecting with other topics like politics (e.g., conferences on CC) and population (e.g., CC impacts on specific regions). This misclassification highlights the models' limitation extending beyond the mere input token limitation, underscoring the challenges in handling topic intersections.

Although not the primary focus, CC was still mentioned in the remaining articles. In particular, 51 articles included CC in contexts marginally related to their main narratives, integrating CC with other discussions. In another 36 articles, CC was a secondary topic, occasionally mentioned only in passing, such as references to the Kyoto Protocol or metaphorical uses of global warming.

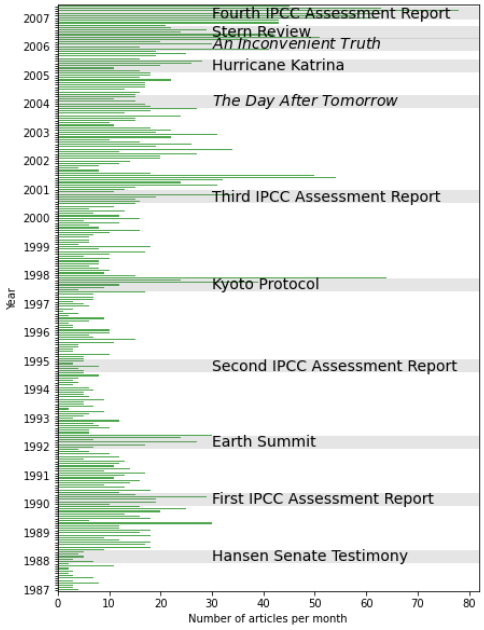

Figure 2: Monthly article count in CC subcorpus

4. Overview of NYTAC-CC

In this section, we provide an initial overview of the NYTAC-CC coverage, including the article distribution over time and a preliminary subtopics exploration.

4.1. Temporal and Keyword highlights

We examine the temporal distribution of articles and key lexical features in our corpus to illuminate trends and shifts in CC coverage over time (see Figure 2).

The analysis reveals a peak in articles during 1990, with up to 50 mentions per month, followed by a decline to 20 articles per month in the mid-90s. After the Kyoto Protocol in December 1997, the curve shows a steady rise with intermittent bursts in coverage. In the figure, we have marked important 'climate events' corresponding to the years they occurred.

The frequency ratios of the top eight lexical features determined by the classifier (cf. Figure 1) over time in Figure 3 illustrate the dominance of 'greenhouse' in the late 1980s. 'Warming' remains the most frequent term throughout, but in the final years, 'climate' gains prominence, suggesting a shift of term preference from 'global warming' to 'climate change'—a transition noted in various other studies as well. Also, the two 'Kyoto' events

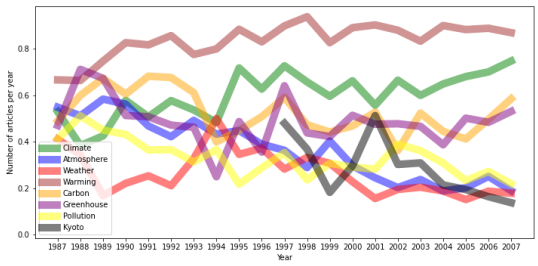

Figure 3: Keyword distributions over time

are clearly visible: the international accord was reached in 1997, and the Bush administration's decision not to ratify it occurred in 2001.

At the same time, we also find that many articles focused on weather or pollution primarily addressed these issues directly, mentioning climate change only tangentially. This reduces the co-occurence of other prominent CC terms in these articles.

4.2. Document Structuring with LDA

Building on the basic statistics discussed in the previous subsection, we delved deeper into the range of subtopics within the CC corpus using topic modeling, specifically Latent Dirichlet Allocation (LDA). This approach helps to uncover underlying thematic structures in the data, which are not immediately apparent from simple keyword analysis.

Preprocessing Steps To prepare the texts for LDA, we performed several preprocessing steps on article titles and bodies, including removing punctuation, lemmatizing words, and converting all text to lowercase to ensure consistency. We also joined frequently co-occurring bigrams into single terms to preserve important phrases. For our topic modeling, we focused on nouns and proper nouns that ranked among the top 10,000 by frequency and had more than two letters. This refinement allowed us to emphasize key entities and their relationships, central to the content of the articles, and avoid the dilution of thematic significance by less informative parts of speech, enhancing consistency through the use of pseudowords.

Model Selection The best LDA model was chosen based on the coherence score, calculated using the Python *Gensim* library. This ensures an objective selection process, minimizing subjective interpretation. We prioritized coherence to ensure that the topics generated by the model are interpretable and meaningful. The optimal model identified 18 topics, with a coherence score of .56, indicating a reasonable level of interpretability. We chose the highest-ranked term as the 'name' of each topic and listed five additional representative terms as follows:

1. **emission:** country, world, greenhouse_gas, carbon_dioxide, global_warming
2. **administration:** president, policy, white_house, bill, congress
3. **people:** time, life, book, world, earth
4. **scientist:** temperature, climate, study, research, university
5. **energy:** oil, fuel, gas, production, power
6. **city:** new_york, people, park, town, mayor, manhattan
7. **company:** business, project, program, group, director
8. **global_warming:** report, climate_change, scientist, panel, editor
9. **plant:** coal, company, emission, power, utility
10. **water:** area, land, river, population, fish
11. **state:** pollution, air, ozone, epa, smog
12. **china:** government, people, war, security, country
13. **car:** vehicle, fuel, gasoline, hydrogen, auto
14. **ice:** sea, arctic, ocean, glacier, bear
15. **forest:** tree, plant, species, fire, crop
16. **weather:** winter, temperature, snow, degree, heat
17. **storm:** el_nino, drought, hurricane, wind, flood
18. **island:** bird, beach, garden, long_island, sand

As is common with topic models, some overlap between topics can occasionally be observed when examining the complete top-30 term lists, for example, between topics *company* and *plant*. Additionally, we find some apparent 'outlier' terms in all the topics.

As a preliminary approximation, we tagged each text in the subcorpus with the predominant topic identified by the model, allowing us to track the evolution of topic coverage over time (see Figure 4). This LDA-based analysis highlights how the context of CC-related coverage in the NYTAC corpus shifts over time, for example from a framing within science and pollution debates to a discourse context in which greenhouse gas emissions were central. Further, our findings complement the manual inspection discussed in Section 3.3, illustrating how climate change discussions, while sometimes secondary in broader articles on government policy (topic 'administration'), are integral to discussions on foreign policy ('China') and cultural topics ('people').

5. Conclusion and Future Work

In this paper, we introduced the NYTAC-CC, a specialized subcorpus of 3,630 climate change articles from the New York Times Annotated Corpus spanning 1987 to 2007,

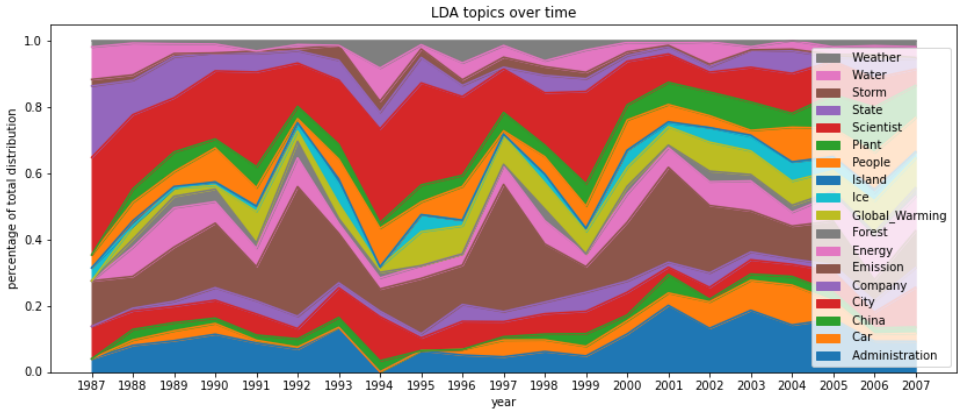

Figure 4: Topic coverage over the 20-year period

marking the first CC analysis with this dataset. Addressing the lack of available news-based textual resources for NLP tasks, we employed a hybrid method combining keyword-based prefiltering and automatic classification to optimize the corpus construction. The representativeness of the subcorpus was confirmed using ClimateBERT, but additional manual inspection of ClimateBERT's classification of a relevant amount of true positives as (false) negatives also showed the model's limitations and the benefits of the hybrid approach chosen.

Initial analyses of the subcorpus, including statistics, keyword searches, and topic modeling, highlight the corpus's potential for detailed diachronic and subtopic exploration.

Thus, the NYTAC-CC subcorpus can be a useful resource for examining the historical narrative of climate change in news media. As it builds on the NYTAC corpus, it adds to previous work on this data, providing valuable insights for social science research. It also serves as a beneficial dataset for developing NLP applications that require a deep understanding of climate-related discourse. While the size of the subcorpus may restrict certain quantitative analyses, its rich, concentrated content is ideal for qualitative studies. Furthermore, it offers the potential for expansion and further integration with additional sources to enhance its utility and relevance for ongoing climate change research. Future work will expand on these findings with advanced topic modeling techniques and integrate more recent articles to enrich the diachronic analysis.

References

[1] Y. Zhang, A. Jatowt, S. S. Bhowmick, K. Tanaka, Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time, in: Annual Meeting of the Association for Computational Linguistics, 2015. URL: https://api.semanticscholar.org/CorpusID:1121386.

[2] O. Alonso, K. Berberich, S. J. Bedathur, G. Weikum, Time-based exploration of news archives, 2010. URL: https://api.semanticscholar.org/CorpusID:2353972.

[3] C. Kantner, M. Overbeck, Exploring soft concepts with hard corpus-analytic methods, in: N. Reiter, A. Pichler, J. Kuhn (Eds.), Reflektierte algorithmische Textanalyse, De Gruyter, Berlin, 2020.

[4] N. Webersinke, M. Kraus, J. Bingler, M. Leippold, 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, 2022. doi:https://doi.org/10.48550/arXiv.2212.13631.

[5] Y. Luo, D. Card, D. Jurafsky, Detecting stance in media on global warming, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020, pp. 3296–3315.

[6] D. Effrosynidis, A. Karasakalidis, G. Sylaios, A. Arampatzis, The climate change twitter dataset, Expert Syst. Appl. 204 (2022) 117541. URL: https://api.semanticscholar.org/CorpusID:248807383.

[7] A. Samantray, P. Pin, Data and code for: Credibility of climate change denial in social media (2019). URL: https://doi.org/10.7910/DVN/LNNPVD. doi:10.7910/DVN/LNNPVD.

[8] T. Diehl, B. Huber, H. G. de Zúñiga, J. H. Liu, Social media and beliefs about climate change: A cross-national analysis of news use, political ideology, and trust in science, International Journal of Public Opinion Research (2019). URL: https://api.semanticscholar.org/CorpusID:214067785.

[9] A. Shehata, J. Johansson, B. Johansson, K. Andersen, Climate change frame acceptance and resistance: Extreme weather, consonant news, and personal media orientations, Mass Communication and Society 25 (2021) 51 – 76. URL: https://api.semanticscholar.org/CorpusID:238720934.

[10] C. Trumbo, Constructing climate change: claims and frames in US news coverage of an environmental issue, Publ. Underst. Science 5 (1996) 269–283.

[11] M. Boykoff, The cultural politics of climate change discourse in UK tabloids, Political Geography 27 (2008) 549–569.

[12] P. Legagneux, N. Casajus, K. Cazelles, C. Chevallier, M. Chevrinais, L. Guéry, C. Jacquet, M. Jaffré, M.-J. Naud, F. Noisette, P. Ropars, S. Vissault, P. Archambault, J. Bêty, D. Berteaux, D. Gravel, Our house is burning: Discrepancy in climate change vs. biodiversity coverage in the media as compared to scientific literature, Frontiers in Ecology and Evolution 5 (2018). URL: https://api.semanticscholar.org/CorpusID:39805874.

[13] M. Boykoff, J. Boykoff, Climate Change and Journalistic Norms: A Case-Study of US Mass-Media Coverage, Geoforum 38 (2007) 1190–2004.

[14] D. Brossard, J. Shanahan, K. McComas, Are issue-cycles culturally constructed? A comparison of French and American coverage of global climate change, Mass Communication and Society 7 (2004) 359–377.

[15] R. Grundmann, R. Krishnamurthy, The Discourse of Climate Change: A Corpus-based Approach, Critical Approaches to Discourse Analysis across Disciplines 4 (2010) 113–133.

[16] D. A. Stecula, E. Merkley, Framing Climate Change: Economics, Ideology, and Uncertainty in American News Media Content From 1988 to 2014, Frontiers in Communication 4 (2019).

[17] P. Mishra, R. Mittal, Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction, in: ICML 2021 Workshop on Tackling Climate Change with Machine Learning, 2021. URL: https://www.climatechange.ai/papers/icml2021/76.

[18] M. Leippold, F. S. Varini, Climatext: A dataset for climate change topic detection, in: NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning, 2020. URL: https://www.climatechange.ai/papers/neurips2020/69.

[19] M. Hulme, N. Obermeister, S. Randalls, M. Borie, Framing the challenge of climate change in Nature and Science editorials, nature climate change 8 (2018) 515–521.

[20] A. Schmidt, A. Ivanova, M. S. Schäfer, Media Attention for Climate Change around the World: A Comparative Analysis of Newspaper Coverage in 27 Countries, Global Environmental Change 23 (2013) 1233–1248.

[21] M. Hulme, Why we disagree about climate change: Understanding controversy, inaction and opportunity, Cambridge UP, Cambridge, 2009.

[22] J. Bingler, M. Kraus, M. Leippold, N. Webersinke, How Cheap Talk in Climate Disclosures Relates to Climate Initiatives, Corporate Emissions, and Reputation Risk, Working paper, Available at SSRN 3998435, 2023.

A. List of Bigrams

climate change, global warming, greenhouse effect, acid rain, ozone layer, greenhouse gases, fossil fuels, greenhouse emissions, ice shelves, ice sheets, rising sea, sea levels, Kyoto Protocol, Montreal Protocol, carbon footprint, carbon dioxide, carbon neutral, emission trading, feedback loop, global dimming, renewable energy, Stern Review.

B. List of Keywords

climate, atmosphere, weather, warming, carbon, greenhouse, pollution.