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

Anonymous ACL-IJCNLP submission

Abstract

Over the past decade, the analysis of discourses on climate change (CC) has gained increased interest within both the social sciences and the NLP community. Textual resources are crucial for understanding how narratives about this phenomenon are crafted and delivered. However, while there is growing attention on social media resources, there still is a scarcity of datasets that cover CC in *news media* in a representative way. This paper presents a CC-specific subcorpus 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 of the subcorpus, 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, which is relevant for various downstream tasks.

1 Introduction

We present NYTAC-CC, a climate change (CC) topic-specific subcorpus of news media articles covering a 20-year period, based on the *New York Times Annotated Corpus* (NYTAC). NYTAC is a collection of 1.8 million articles published by the paper between 1987 and 2007 and made available by the *Linguistic Data Consortium*[1]. The original corpus, and thus also the subcorpus, includes a variety of metadata, including the 'desk' (the newspaper branch) and both manually- and automatically-labeled tags that categorize the content. Furthermore, a sizable subset of the articles includes hand-written summaries.

NYTAC has been used for various research purposes in the Natural Language Processing (NLP) community since its release in 2008, which allows the CC research community to profit from 15 years of related NLP work (e.g. Zhang et al. (2015); Alonso et al. (2010)). In addition, given the extensive temporal coverage, the NYTAC-CC subcorpus serves as a valuable resource for investigating how CC has been discussed and portrayed in the news media over time, including how early CC debates were embedded within other subtopics such as domestic and foreign policy, science reporting, or articles from the domains of arts and culture. Compared to other CC-related resources that focus on shorter documents, the NYTAC-CC subcorpus includes documents of variable length.

The contribution of this paper is threefold:

- *First*, we present the NYTAC-CC subcorpus (by publishing the filenames) 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 (cf. Kantner and Overbeck (2020)). This hybrid approach is designed to address the challenges associated with refining a topic-specific subcorpus or extracting relevant information from a larger, existing corpus. It aims to overcome the limitations of traditional sampling techniques, which often involve retrieving articles via a small set of keywords or bigrams and can lead to the inclusion of false positives in the datasets.

- *Second*, to demonstrate the representativeness of the subcorpus and its reliability for further downstream tasks, we illustrate the results of a classification experiment using ClimateBERT (Webersinke et al., 2022), a BERT-based model specifically trained on CC-related texts, to validate that the articles in our NYTAC-CC subcorpus are true positives.

[1] http://ldc.upenn.edu

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

Thus, our goal is to utilize the NYTAC corpus to gain a comprehensive picture of the NYT's coverage of CC during the specified time period through our subcorpus. While several studies have examined U.S. print media's reporting on anthropogenic CC (see Section 2), to our knowledge, this is the first work that specifically addresses the 20-year period covered by the NYTAC.

The rest of the paper is structured as follows: Section 2 summarizes relevant related work. Section 3 describes the process of creating the topic-specific subcorpus from the NYTAC through the aforementioned combined method, as well as results from classification task using ClimateBERT to validate the NYTAC-CC subcorpus. Section 4 offers key observations on the content and distribution of NYTAC-CC articles, including the outcome of an LDA experiment for unsupervised sub-topic exploration within the subcorpus. We conclude in Section 5 with suggestions for future work based on the subcorpus.

2 Related Work: Climate Change in News

Despite the growing interest in addressing climate change among various academic communities, as pointed out by Luo et al. (2020), 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 (Effrosynidis et al., 2022; Samantray and Pin, 2019). Other existing datasets mostly remain limited to the sentence or paragraph levels (Leippold and Varini, 2020; Diggelmann et al., 2020; Laud et al., 2023). However, there remains a scarcity of NLP works (focusing here on English text) that address the CC discourse at the news article level, where the majority of studies on CC within traditional media have been conducted in various social science disciplines (Diehl et al., 2019; Shehata et al., 2021). This section will focus on prominent work targeting traditional news media.

A widely-cited early study by Trumbo (1996) 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 (2008) later studied the "claims and frames" issue in a similar manner.

Legagneux et al. (2018) 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. Boykoff and Boykoff (2007) analyzed CC coverage in U.S. TV and newspapers (1988-2004) to see if journalistic norms like personalization, drama, and balance hindered reporting on anthropogenic CC. Their study, using manual coding, revealed that an emphasis on balance and drama often gave undue prominence to fringe scientists. Other studies examined the frequency of CC mentions, or the 'attention cycle'. Brossard et al. (2004) compared CC reporting between the NYT and the French *Le Monde*. Grundmann and Krishnamurthy (2010) analyzed newspapers from four countries, enhancing article counts with word frequency and collocation analyses using corpus-linguistic tools, with outcomes manually interpreted. The work of Stecula and Merkley (2019) highlights one of the few instances where NLP technology is extensively used to analyze CC in newspapers. They applied supervised classification to construct a corpus and identify frame categories within four U.S. papers. Continuing within the NLP field, Webersinke et al. (2022) use a corpus that includes, among its subsets, the NEWS dataset containing CC-related news articles. However, the dataset is not publicly available, nor are the specifics on how these data were retrieved detailed. Mishra and Mittal (2021) curated a dataset of 11k news articles by web scraping from the Science Daily website.

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

3 Building the NYTAC-CC Subcorpus

3.1 Challenges in CC Text Selection

In the LDC release, the New York Times Annotated Corpus comprises 1,855,658 articles published between 1987 and 2007, each provided as a single XML file. The corpus contains detailed metadata, including information about the date, author, and newsroom desk that published the article. Additionally, the documents are manually annotated with information about locations, people, organizations, and topics that are prominent in an article. Annotators could choose as many tags as they wished from a set of entities that appears to have expanded over the years. The 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 climate change.

Our goal is to design a retrieval method that not only meets the requirements of *validity* and *reliability* but also emphasizes *representativeness*, ensuring the corpus adequately covers the range of content related to the specific subject matter 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. This holds even with advanced models, particularly when tasked with processing large linguistic units such as entire articles (Leippold and Varini, 2020). 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

To refine our method, we first reviewed the methods for text retrieval that have been used in previous studies on (CC) discourse, such as those mentioned in Section 2, as well as in other work targeting blogs and Twitter. 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., (Trumbo, 1996; Legagneux et al., 2018)
2. Search with a longer list of keywords, followed by manual filtering; e.g., (Hulme et al., 2018; Leippold and Varini, 2020)
3. Complex Boolean queries with keywords and operators such as AND, OR, NOT; e.g., (Schmidt et al., 2013)
4. Manual annotation of training data followed by supervised classification; e.g., (Stecula and Merkley, 2019)

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 should deal with some aspect of anthropogenic climate change, relate information about it, or convey a stance on the existence or urgency of the problem. It is not sufficient to merely mention CC in passing, for example, as one among a series of other ongoing crises. The most challenging judgments arise with texts that deal with an environmental problem possibly related to climate change (such as CO_2 emissions, ozone depletion, deforestation, etc.), even though CC is not explicitly mentioned. Our criterion is that the connection to CC must be clearly inferable from the text. For instance, an article that provides merely statistics on different types of air pollution would not qualify as CC-related unless the link to climate change is made explicit by the author.

Bigram search. Initially, we experimented with a list of bigrams[2] sourced from the BBC Climate Change Glossary[3]. This was done to cover terminologies that were used over the two decades

[2] 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

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

spanned by the corpus. We applied the criterion that an article must contain at least one instance of one of the bigrams, which led to the retrieval of 10,707 articles. Upon manual inspection, we found that many articles were false positives, addressing general environmental problems but not specifically related to climate change. 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 employed by Hulme et al. (2018) for identifying all topic-relevant articles in the journals *Nature* and *Science* between 1966 and 2016.[4] To these, we added the keyword ”Kyoto.” An article had to contain at least two different keyword types to be selected. However, the resulting subcorpus still contained many false positives. One source of these was very long list-like articles from the corpus, categorized as ”Listing,” ”News Summary,” ”Business Digest,” ”Inside,” or ”Observatory,” which combined a variety of different news in a single document. In the interest of homogeneity, we removed all articles from these categories, which led to an intermediate corpus consisting of 12,883 articles.

Text ranking and supervised classification. To overcome the presence of false positives within the intermediate corpus, we decided to implement an additional, more elaborate filtering step on the intermediate corpus. Initially, we 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. For instance, 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). For example, if the word ”climate” appears once in the second sentence, 0.9 * 1 is added to the overall score of the text.

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 climate change. An article received the label ’1’ if it referenced the problem of global climate change or an aspect thereof—consistent with our previous characterization. It was labeled ’0’ if it did not meet this criterion. Similarly, articles where the words ’climate’ and ’weather’ were used figuratively, as in ’They weathered the climate,’ were also labeled ’0’.

We used the manually-annotated data to train and test an XGBoost classifier, which we configured to differentiate between CC-related and non-CC articles. The features used included keyword counts (those from (Hulme, 2009), 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 achieves 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. A total of 9,067 articles were labeled as -climate (not CC-related), while 3,630 were designated as +climate (CC-related), forming what we now refer to as our final ’NYTAC climate change subcorpus.’[5] The graph in Figure 1 illustrates the features that had the greatest impact on the classification decisions.

3.3 Evaluating the Subcorpus 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 in retrieving topic-consistent texts from a larger and heterogeneous collection while minimizing the inclusion of false positives. To perform that validation, we conducted experiments with ClimateBERT, specifically $ClimateBert_F$ (Webersinke et al., 2022), a BERT-based model trained on CC-related texts. In particular, we used *distilroberta-base-climate-detector* from the Hugging Face platform[6] (Bingler et al., 2023), a fine-tuned version of $ClimateBert_F$ with a classification head for

[4]climate, atmosphere, weather, warming, carbon, greenhouse, pollution

[5]To facilitate future research, we will make the IDs of the texts available upon the publication of this paper.

[6]https://huggingface.co/

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, which we describe below. For our experiments, we used only the text of the articles as input data, without employing any advanced pre-processing steps or including additional metadata.

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 conducted a manual inspection of the remaining 162 texts that the model classified as -climate, i.e. as false positives for our corpus. We found that the model clearly misclassified 75 texts, which after manual inspection turned out to include relevant sections on CC. However, in part, this was due to the model's input limitations, which led to the misclassification of longer texts containing relevant climate-related parts later in the text. More qualitative insights on these 162 texts initially identified by ClimateBERT as false positives are provided in Section 4.1. 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.

4 Overview of the NYTAC-CC Corpus

In this section, we aim to provide an initial overview of the coverage within the NYTAC-CC, including specifics on the distribution of articles over time and an initial examination of the subtopics it portrays. We begin by illustrating a qualitative content analysis of the articles classified as false positives by ClimateBERT. Following this, we conduct a straightforward yet informative keyword-based analysis and a topic modeling experiment that offers a preliminary insight into some of the key subtopics covered by the subcorpus.

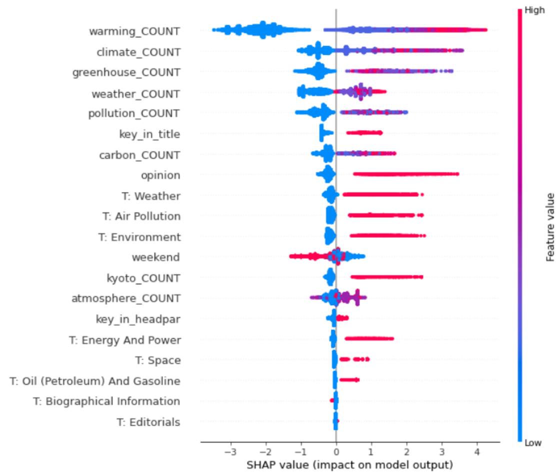

Figure 1: High-impact features in classifying "climate change" articles

4.1 Qualitative Analysis of Misclassified Articles

As discussed in Section 3.3, we performed a manual inspection of the 162 articles that ClimateBERT initially classified as false positive within our subcorpus. We found that 75 out of these 162 articles were, in fact, clearly related to CC. Specifically, 48 of these articles included substantial discussions on CC and related issues occurring after token 512, which indicates that the model's limited context length significantly impacted its classification accuracy. Furthermore, 27 articles were either entirely about CC or contained several paragraphs explicitly detailing CC stories before token 512. These discussions often intersected with other themes such as politics (e.g., conferences on CC) and population effects (e.g., impacts of CC on specific regions).

Despite not being the main focus, the remaining articles still mentioned CC-related information. Specifically, in 51 articles, CC was mentioned within sections that were marginally related to the main narrative, showing that CC could be interwoven with discussions on other topics. Additionally, in 36 articles, CC played a secondary role, being mentioned as part of a longer list of issues or events or merely in passing—for instance, some articles mentioned the Kyoto Protocol as an example, while others used global warming metaphorically.

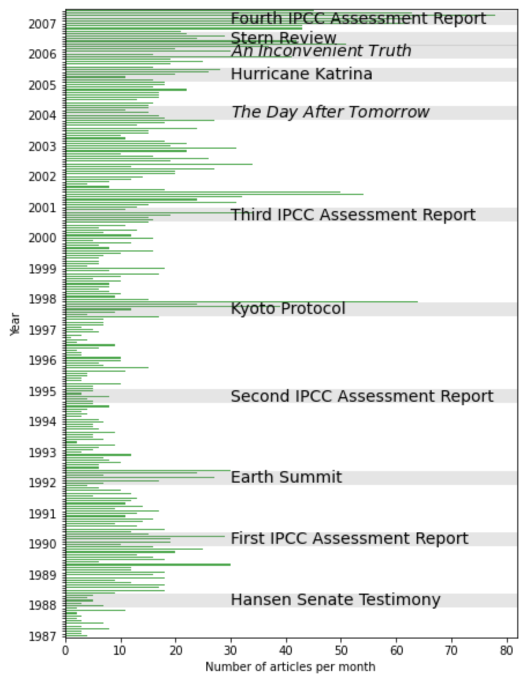

Figure 2: Monthly article count in CC subcorpus

4.2 Temporal and Keyword highlights

We examine the temporal distribution of articles and the usage of key lexical features in our corpus. This analysis helps illuminate trends and shifts in the coverage of climate change over time. When looking at the distributions of articles over time (Figure 2), we observe a peak around the year 1990, with up to 50 articles on climate change per month. This is followed by a drop in coverage, with only 20 articles per month during the mid-90s. From the beginning of 1998 – and following the adoption of the Kyoto Protocol in December 1997 – the curve shows a steady rise with intermittent bursts in coverage. In the figure, we have inserted several important 'climate events' (taken from various online sources) 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 are clearly visible: the international accord was reached in 1997, and the Bush administration's decision not to ratify it occurred in 2001.

Upon examining the co-occurrence of the top keywords, we noticed that 'atmosphere', 'weather', 'pollution', and 'Kyoto' are outliers, generally co-occurring less frequently with other terms. This observation supports our earlier description of varying degrees of CC-topicality: many articles discussing weather or pollution primarily address these issues directly, mentioning climate change only tangentially, which results in a low frequency of other prominent CC terms in these articles.

4.3 Structuring the Document Set with LDA

Building on the basic statistics previously discussed, 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 implemented several preprocessing steps on both the titles and bodies of the articles. These included: removing punctuation symbols; lemmatizing words to reduce them to their base or dictionary form; applying POS-tagging to identify parts of speech, as we focused on nouns; lower-casing all words to ensure uniformity; joining commonly co-occurring bigrams into single terms to preserve significant phrases. Additionally, to refine the focus of our topic modeling, we retained only words that met all the following criteria: (i) Classified as nouns or proper nouns, (ii) Ranked among the top 10,000 nouns and proper nouns by frequency, (iii) Comprised of more than two letters. We restricted our analysis to nominal phrases, concentrating on entities and their relationships within the topic model to emphasize concepts central to the content of the articles. This simplification helps to avoid the dilution of thematic significance by less informative parts of speech and is supported by transforming common bigrams into single pseudowords for clarity and consistency.

Model Selection The best LDA model was chosen based on the coherence score, calculated using the Python *Gensim* library[7]. This method ensures that our model selection is objective, minimizing

[7] https://pypi.org/project/gensim/

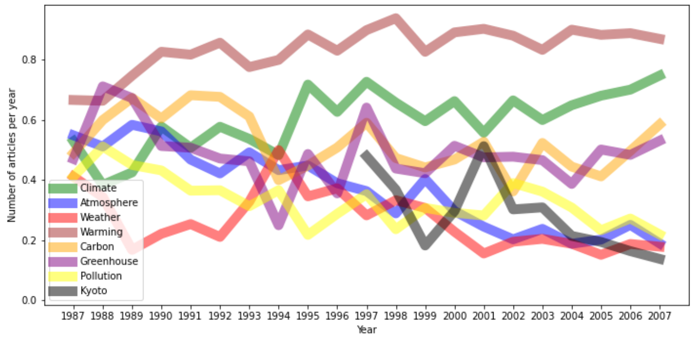

Figure 3: Keyword distributions over time

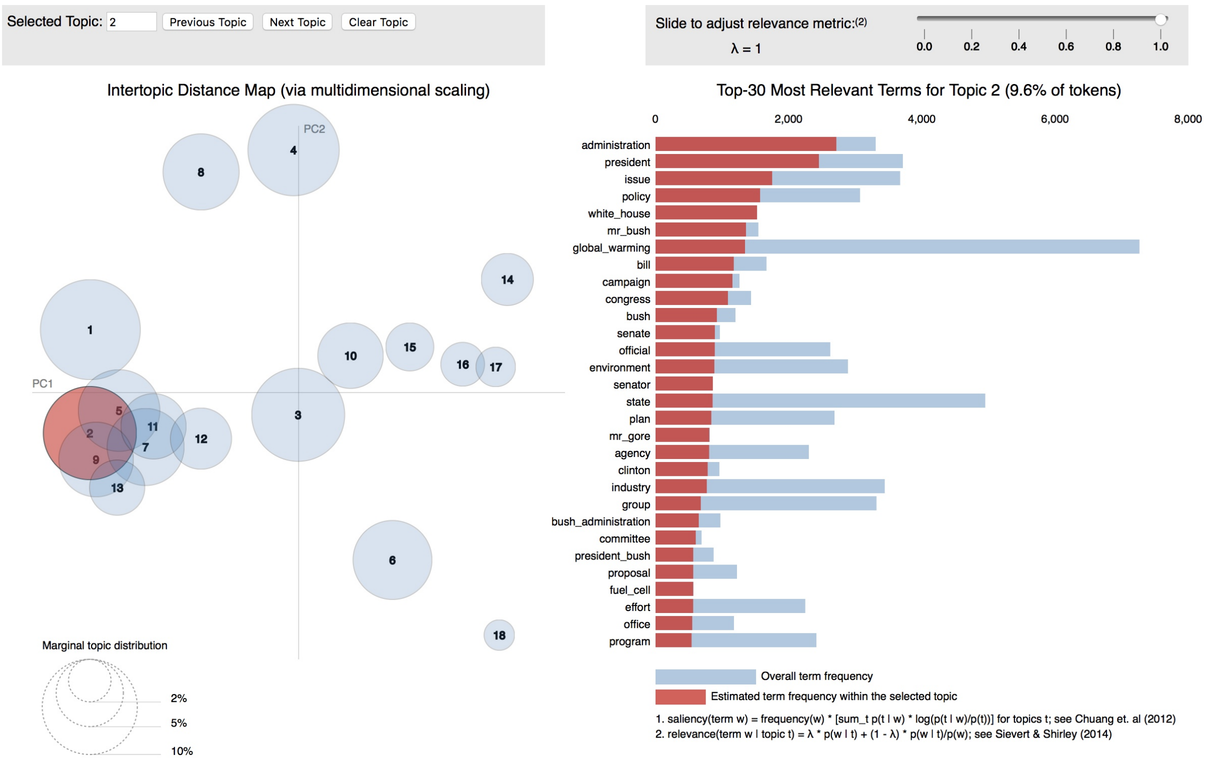

Figure 4: Two-dimensional plot of the LDA topics, with terms of topic 2 ("administration") on the right

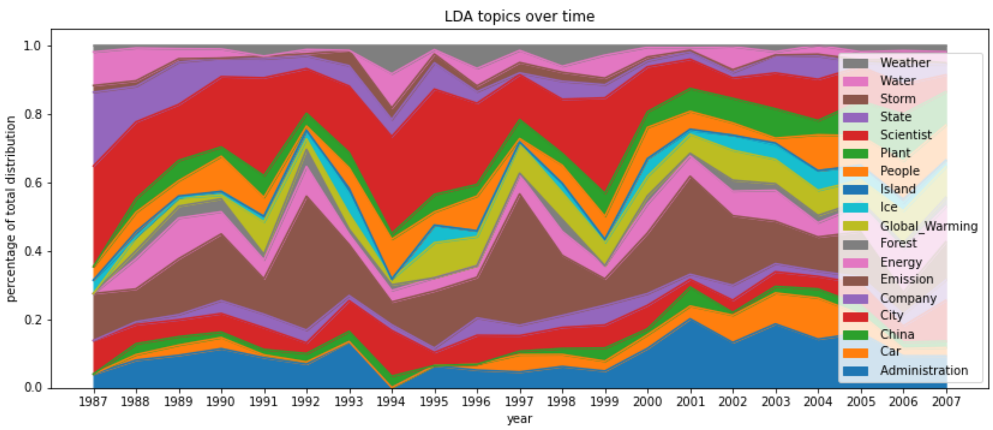

Figure 5: Topic coverage over the 20-year period

subjective interpretation in the analysis. We prioritized coherence to enhance the likelihood that the topics generated by the model are interpretable and meaningful. The optimal model identified consists of 18 topics, with a coherence score of .56, indicating a reasonable level of interpretability. For each topic, we chose the highest-ranked term as the 'name' of the topic and list 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. However, on the whole, we believe that the reduction to nominal terms has led to a rather clear delineation of subtopics relevant to the problem of climate change. Figure 4 displays a plot where (on the left-hand side) the neighborhood of topics can be studied. Notably, observe the proximity between topics 16 (*weather*) and 17 (*storm*), or the isolated position of topic 18 (*island*). The close clustering of topics 2 (*administration*), 5 (*energy*), 7 (*company*), 9 (*plant*), and 11 (*state*) appears to be significant, indicating thematic interrelations.

As a preliminary approximation, we tagged each text in the subcorpus with the predominant topic identified by the model. This enables us to visualize the development of topic coverage over time; see Figure 5. 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. Adding to our manual inspection in Section 3.3 that showed how climate change can be one of many issues in longer articles about general government policy (topic "administration"), the topic modeling also indicates how CC debates may be embedded in broader discussions about foreign policy ("China") or in articles about culture and arts ("people").

5 Conclusion and Future Work

In this paper, we introduced the NYTAC-CC, a topic-specific subcorpus of 3,630 articles on climate change (CC) drawn from the New York Times Annotated Corpus covering the span from 1987 to 2007. This marks the first CC analysis using this data set. Our work addresses the scarcity of available news-based textual resources for climate change, crucial for many NLP downstream tasks. We constructed the corpus using a hybrid approach that combines keyword-based prefiltering with automatic classification, effectively optimizing the extraction process. The representativeness of the subcorpus is validated through the application of ClimateBERT, with classification results revealing both the model's intrinsic limitations and the topic-consistency of the subcorpus. Initial explorations, including basic statistics, keyword analysis, and topic modeling, have already highlighted the potential for nuanced diachronic analysis and more fine-grained subtopic exploration. As future work, we plan to extend these preliminary findings by employing advanced topic modeling techniques, such as structured topic modeling that systematically incorporates time and dynamic topic modeling.

References

Alonso, O., Berberich, K., Bedathur, S. J., and Weikum, G. (2010). Time-based exploration of news archives.

Bingler, J., Kraus, M., Leippold, M., and Webersinke, N. (2023). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Working paper, Available at SSRN 3998435.

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

Boykoff, M. and Boykoff, J. (2007). Climate Change and Journalistic Norms: A Case-Study of US Mass-Media Coverage. *Geoforum*, 38(6):1190–2004.

Brossard, D., Shanahan, J., and McComas, K. (2004). Are issue-cycles culturally constructed? A comparison of French and American coverage of global climate change. *Mass Communication and Society*, 7(3):359–377.

Diehl, T., Huber, B., de Zúñiga, H. G., and Liu, J. H. (2019). 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*.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. (2020). Climate-fever: A dataset for verification of real-world climate claims. abs/2012.00614.

Effrosynidis, D., Karasakalidis, A., Sylaios, G., and Arampatzis, A. (2022). The climate change twitter dataset. *Expert Syst. Appl.*, 204:117541.

Grundmann, R. and Krishnamurthy, R. (2010). The Discourse of Climate Change: A Corpus-based Approach. *Critical Approaches to Discourse Analysis across Disciplines*, 4(2):113–133.

Hulme, M. (2009). *Why we disagree about climate change: Understanding controversy, inaction and opportunity*. Cambridge UP, Cambridge.

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

Kantner, C. and Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. In Reiter, N., Pichler, A., and Kuhn, J., editors, *Reflektierte algorithmische Textanalyse*. De Gruyter, Berlin.

Laud, T., Spokoyny, D. M., Corringham, T. W., and Berg-Kirkpatrick, T. (2023). Climabench: A benchmark dataset for climate change text understanding in english. *ArXiv*, abs/2301.04253.

Legagneux, P., Casajus, N., Cazelles, K., Chevallier, C., Chevrinais, M., Guéry, L., Jacquet, C., Jaffré, M., Naud, M.-J., Noisette, F., Ropars, P., Vissault, S., Archambault, P., Bêty, J., Berteaux, D., and Gravel, D. (2018). 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.

Leippold, M. and Varini, F. S. (2020). Climatext: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Luo, Y., Card, D., and Jurafsky, D. (2020). Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online.

Mishra, P. and Mittal, R. (2021). 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*.

Samantray, A. and Pin, P. (2019). Data and code for: Credibility of climate change denial in social media.

Schmidt, A., Ivanova, A., and Schäfer, M. S. (2013). Media Attention for Climate Change around the World: A Comparative Analysis of Newspaper Coverage in 27 Countries. *Global Environmental Change*, 23(5):1233–1248.

Shehata, A., Johansson, J., Johansson, B., and Andersen, K. (2021). Climate change frame acceptance and resistance: Extreme weather, consonant news, and personal media orientations. *Mass Communication and Society*, 25:51 – 76.

Stecula, D. A. and Merkley, E. (2019). Framing Climate Change: Economics, Ideology, and Uncertainty in American News Media Content From 1988 to 2014. *Frontiers in Communication*, 4(6).

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

Webersinke, N., Kraus, M., Bingler, J., and Leippold, M. (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*.

Zhang, Y., Jatowt, A., Bhowmick, S. S., and Tanaka, K. (2015). Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Annual Meeting of the Association for Computational Linguistics*.