Climate Change Indices from Online News: A Global Approach

F.A. Vavrik Felisberto (2677146)

June 30, 2022

Abstract

This paper concerns the textual analysis of climate change news obtained from GDELT GKG 1.0, creating indices with aim of quantifying the relative share and sentiment of Climate Change-related articles in global news. These indices capture global trends and climate change-related events, ranging from September 2013 until April 2022. It was found that the general Climate Change Index captures the effects of large environmental catastrophes, international conferences and agreements, US elections, and political decisions, in sentiment and relative frequency of publication.

Contents

1 Introduction

Climate Change is an unavoidable topic. News outlets are the main responsible vehicle for the distribution of this information, as they report on developments of climate disasters, policy, science, and investment on daily basis. The Internet has facilitated the access and flow of such information, making it even more available with just a few touches of a screen. This study builds on top of my research project, delivered in February 2022.

The aim of this paper is to analyze global English language news about climate change, using novel machine learning methods, to solve Natural Language Processing (NLP) tasks, such as text classification and sentiment analysis. These two tasks enable the creation of global climate change indices, capable of quantitatively measuring the effect of global climate change events, both terms of news publication frequency and sentiment.

The use of NLP (or textual analysis) has gained popularity in economic and financial literature, as a means to obtain high-frequency data, and uncover previously unobserved behaviors or relationships (Ghanem and Smith, 2021). Gentzkow et al., 2017 is seminal work in the area of textual analysis, as it provides a good overview of methods and applications of NLP techniques, in economic and financial academic research. A more applied example of such research is presented in Bybee et al., 2021, where a topic model is used to extract information from business news, using it to estimate the dynamics of macroeconomic variables and equity markets Barbaglia et al., 2021, uses news sentiment to improve the performance of regression models to forecast yield spreads of Italian Government Bonds. Bodas-Sagi and Labeaga, 2016, finds a weak correlation between the sentiment of news related to Spanish energy policy, and the average daily energy price, in Spain.

Research on the textual analysis of climate-related text is still in its infancy. Some examples of such research are Dahal et al., 2019, where public opinion on climate change is analyzed from social media posts. A more complex example of analysis of climate-related text is presented in Kölbel et al., 2020, which uses ClimateBERT (Webersinke et al., 2021), a neural language model trained on climate-related text, to analyze and detect climate risk from corporate financial statements, and relate it to the disclosing firm's corporate derivative spreads. ClimateBERT has also been used to analyze the presence of cherry-picking in voluntary climate risk disclosures, in Bingler et al., 2021.

Regarding the measurement of the effect of climate change news through indices, the only directly related literature is Engle et al., 2019 and Noailly et al., 2021.

Engle et al., 2019 presented a simple way of indexing the impact of climate change news. This is achieved by measuring the correlation of the content articles to a dictionary of climate-related terms. The index proposed is solely based on text from the Wall Street Journal, and was used to quantify, and hedge, the climate news exposure of a stock portfolio.

Noailly et al., 2021 provides a more sophisticated method of indexing climate change news, as it uses a broader set of US-based online newspaper articles. In the paper, a Support Vector Machine is trained to recognize and collect articles related to Environmental Policy, and the sentiment of the collected articles is returned using a fine-grained aspect-based Sentiment analysis method, allowing to compute the sentiment of a document, using a domain-specific lexicon. The indices developed in the paper, show that the relative frequency of publication and sentiment of environmental policy news is related to the likelihood of a clean energy startup receiving venture capital funding. It also shows a relation between the assets under management of renewable energy exchange-traded funds.

Questions regarding media bias and slant, due to political affiliations of newspapers, have already been addressed by Gentzkow and Shapiro, 2006 and Gentzkow and Shapiro, 2010. Specifically pertaining to bias in climate change news, Bohr, 2020 finds that the effects of media bias do not appear to significantly impact the volume of coverage of climate change. It also finds that the tone of the discourse and position towards topics such as climate mitigation, or denial, will differ depending on the political affiliations of the publisher.

This paper is set out as follows: Section 2, describes how I leverage Global Dataset on Language and Tone's Global Knowledge Graph 1.0 dataset (GDELT GKG 1.0) to obtain access to a set of articles that are climate change-related, English language, news articles from around the world. It shows, also, how these articles are divided into smaller, Climate Change related categories. Section 3, gives a brief explanation of how Transformer Neural Networks perform inference on text, and how these are adapted to perform text classification and sentiment analysis, in the context of this paper. Section 4, shows and discusses the resulting indices, and how they react to climate change-related events. Finally, in section 5, conclusions are drawn and possible future applications of the developed indices are discussed.

2 Data:

At the core of this analysis lies the Global Knowledge Graph 1.0 dataset, by GDELT (Leetaru and Schrodt, 2013). This is one of the largest news datasets available, providing access to links to daily online news articles and other categories, their perceived tone, persons and entities involved, location, and some labels which indicate the themes approached in the article. In total it contains over 300 million entries, spanning from April 1st, 2013 until April 1st, 2022.

Apart from the URL to the original source, the other data fields are powered by Google Cloud's Natural Language Processing engine and GDELT's own Global Content Analysis Measure algorithm. Since it is not publicly known how these algorithms operate, problems regarding the transparency of how the tone is measured arise, and therefore will not be included for the purpose of this analysis.

Labeling of data is made according to CrisisLex, World Bank Group taxonomies, as well as GDELT's own labeling system. [1]

2.1 Label-based filtering

The analysis of the data begins by manually selecting all the `theme` labels related to Climate Change from a set of over 2500 theme labels, in order to create a simple data filter, which allows to efficiently filter the complete database.

After label selection, the relevant labels are grouped into the main category, Climate Change, and subcategories like Climate Finance, Environmental Policy and management, Green Energy, Environmental Crimes and Disasters, and Green Economics. The list of labels that make up each category is specified in appendix **??**. A Python script is then used to check for the existence of relevant labels for each entry of the dataset. In case the labels given to the article match one of the relevant (sub)categories, the article is added to that (sub)category. These subcategories are not mutually exclusive as news articles can address several of these topics within their text, and therefore, these categories have some overlap.

The articles categorized as "Climate Change" include articles that discuss climate change as a general topic, but also ones that focus more on how it impacts communities, and how these adapt to its effects. This category includes all the 6,219,542 articles retrieved and serves as a means to ensure that all the articles present in other categories retrieved are climate change-related.

The remaining subcategories present more specific subtopics. Articles belonging to "Environmental Crimes and Disasters", discuss catastrophic events where there is harm to the environment. Both natural disasters, (such as storms, floods, or tornadoes) and caused by human action (such as oil spills, arson, and mass deforestation) are included in this subcategory.

"Green Energy" includes all articles which discuss wind, solar, geothermal, hydroelectric power generation, or bio-fuels. "Green Economics" includes topics such as green and sustainable growth, green employment, and the environmental impact of trade and transportation. Environmental Policy includes articles that discuss developments in environmental policy or the effects of the implementation of such policies. Finally, Climate Finance articles, discuss the issues related to the financing of climate-related projects climate and climate risk management.

The total amount of articles per category is specified in table 1.

2.2 Text Acquisition for Sentiment Analysis

In order to perform coherent sentiment analysis, it is necessary to access and download the text from the articles as some articles are mislabelled or irrelevant to the analysis. This was achieved by using the Python package `goose3`, allowing to acquire the text from news articles, already cleared

[1] The list of all labels is available in http://data.gdeltproject.org/api/v2/guides/LOOKUP-GKGTHEMES.TXT.

Category	Counts
Climate Change	6,219,542
Climate Finance	284,362
Environmental Policy and Management	205,627
Environmental Crimes and Disasters	5,714,374
Green Economy	317,760
Green Energy	1,102,262

Table 1: Article Categories and Counts

from HTML tags. Because it was necessary to consider the time constraints of this project, a decision was made to only access websites which can be scraped within 1 second of request. This, coupled with a significant amount of dead links and "pay-walled" articles, limited the access to the text of approximately 200,000 articles.

3 Neural Language Models and Sentiment Analysis

Quantifying the sentiment from an article is a complex task, that is central to analyzing the effect of climate change news. As mentioned before, GDELT itself provides sentiment analysis. The reasons why it was decided to abstain from its use are the following: Firstly, blindly considering the results from GDELT's sentiment analysis without looking further into the text can lead to errors, as it was noticed that some articles are mislabelled, compromising the quality of the sentiment analysis.

A second problem is the lack of transparency of the method used to analyze the articles, which can result in some incoherence between sentiment values. This is further exacerbated by the fact that since 2013, the field of Natural Language Processing has evolved to such an extent that the methods used to compute sentiment from the text are not easily comparable.

With this in mind, this paper proposes that the analysis of a large sample of articles is used to approximate the average daily sentiment of Climate Change news.

Here, a 2-staged approach to text classification is implemented, with aim of tackling the issues raised above. Starting with a sample drawn from the cleaned and categorized GDELT dataset. The first stage deals with removing articles that are not related to the main topic of climate change. In case an article is classified as "related", it then follows into the second stage, where the sentiment of the article is extracted.

The models used to perform this belong to the same type of Neural Language Model, namely, Transformer Neural Networks (Vaswani et al., 2017). These models follow a similar structure: Tokenization, Encoding (and Decoding), and the downstream task (classification, topic modeling, sentiment analysis, etc.).

The process of inference using these models happens in the following manner:

Firstly, tokeninzer transforms a sequence of words into vector representations (tokens). These are known as embeddings. It is necessary to note, that the embedding includes an encoding of the position of each token.

Afterward, the output of the tokenizer is passed onto the encoding stack. Here the input vectors are transformed, resulting in different feature representations. These representations are key to extracting meaning and context from a section. From these feature representations, a classifier head, which is generally a densely connected neural network, can be trained to perform sentiment analysis, which is the case in the used implementation.

A decoder, used in the case of the filtering network, has the task of reconstructing the original text from the encoded outputs of the encoder. This allows the model to reconstruct sentences from the encoded outputs, returning them as vectors of word probabilities. These outputs can also be used to construct a text classifier, which, in this paper is used to filter for climate-related articles. How these models were implemented is described in the sections below.

[2].

3.1 Zero-Shot Classification for Advanced Filtering

The typical approach taken towards classifying relevant text is to sample a sizable subset of the dataset, manually label it, and train on the specific classification task. This is the ideal approach, which was used in Bingler et al., 2021, and Kölbel et al., 2020. Unfortunately, the classifier head, which is capable of discerning climate change-related text from non-climate change-related text, is not publicly available, thus making the use of ClimateBERT impossible due to time constraints.

In light of this, it was decided that the only approach for this task was to base it on a Low Resource Natural Language Processing technique (Hedderich et al., 2021), namely, Zero-Shot classification.

This unsupervised transfer-learning approach is fully described in Yin et al., 2019. The approach solves the task of labeling a sequence of text, which belongs to a set of several classes not observed when training the set of the neural network.

The model to achieve this used in this paper uses the BART architecture (Lewis et al., 2019), from a model that was pre-trained on the MNLI dataset ("valhalla/distilbart-mnli-12-9 · Hugging Face", n.d.;Williams et al., 2018).

In this natural language understanding problem, the model is given 2 text sequences, a premise, and a hypothesis, and is trained to classify whether the hypothesis entails, contradicts, or is neutral towards the premise.

To convert a model from an entailment task into a zero-shot text classifier, the neutrality output of the model is dropped and the entailment/output is kept. Instead of passing a premise as the second input, a set of candidate labels serves as input. This output (which previously displayed the inferred probability of entailment), will now encode the probability of a candidate label being related to the article to be classified.

In order to formulate a filter for climate change-related articles, the following candidate labels were used: "climate", "sustainable", "renewable" and "environment".

These are not mutually exclusive as an article can often be classified to belong to several of these several classes. Given this, an article is considered related when, at least, one of these labels returns a probability of over 50%.

At face value, these labels seem too general, and this would indeed be the case when considering the unfiltered dataset, which could result in worse classifier performance. Since initial filtering has already been performed, the chosen set of candidate labels allows to filter out articles in which climate change-related themes are briefly mentioned, or not the central topic.

The precise accuracy of the classifier is hard to gauge, due to the unlabelled nature of the data, however from a limited sample of 1 day (89 articles - 1/4/2013), it was achieved a 70.0% accuracy and 70.1% recall. Although accuracy and recall are not optimal, due to the non-specific training set, the model still holds an acceptable performance.

The addition of more candidate labels could improve the performance of the classifier, however the inference time necessary scales linearly with the number of classes used, so a trade-off between

[2]All of the models used are compressed versions of the original models, referred to as distilled. This allows for quicker inference time while keeping up to 95% of the performance of the original model (Yuan et al., 2021).

decreasing the false-negative rate of the classifier and the speed at which class inference was made.

3.2 Sample Classification Result

The makeup of the sampled categories are visible on the table 2. It resulted in not enough data being gathered for the categories of "Climate Finance", "Environmental Policy and Management", "Green Economics and Sustainable Growth" and "Green Energy" to produce a representative sentiment index for these categories, and the global scale of the problem and the heterogeneous temporal distribution of the data. This is partly due to the lower article counts in these categories and the limited sample size. For these reasons only the sentiment of the general category, "Climate Change", and the topic-specific category, "Environmental Crimes and Disasters", will be plotted.

Table 2: Number of articles retrieved per category.

Category	Articles Retrieved
Climate Change	123446
Climate Finance	4672
Environmental Policy and Management	4248
Environmental Crimes and Disasters	110671
Green Economics and Sustainable Growth	7547
Green Energy	23842

3.3 Sentiment Analysis

After finding related articles, another model (BERT from Devlin et al., 2018) is used to quantify the sentiment behind each article. The model used was previously fine-tuned on the SST-2 database, and is able to distinguish sentiment polarity from text, outputting a label of "Positive" or "Negative" and the inferred label probability (Socher et al., 2013;"distilbert-base-uncased-finetuned-sst-2-english · Hugging Face", n.d.).

The chosen model has a limit input length of 512 tokens, which causes issues when analyzing longer sequences of text.

In order to avoid this issue, the input text is separated into paragraphs, and the sentiment is calculated for each paragraph.

In order to obtain a continuous scale ranging from -100 (Very Negative) to +100 (Very Positive), with the neutral point at 0.

This process is summarize by the equation:

$$S = 100 \cdot \frac{\sum_{n=1}^{N} pol_n \cdot p(pol_n)}{N} \tag{1}$$

where S is the average sentiment score from a given article with N paragraphs, and, pol_n represents the polarity of the n-th paragraph. $p(pol_n)$ represents the model's estimated probability of belonging to the classified polarity.

In the case where a paragraph is classified as having a "Negative" sentiment, pol_n is taken as -1. When classified as "Positive", the pol_n is taken to be $+1$.

4 Results - Visualizing the Indices:

4.1 Relative Frequency Indices

The basic label filtering of the dataset, allows us to observe the general changes in the relative media attention to Climate Change (and related subcategories) by plotting the relative share of articles published over time. This is done similarly to the frequency plots presented in Noailly et al., 2021, where the ratio between Climate Change related articles and the total amount of articles published is presented. The data is presented as the moving average, with a centered window of 31 days, and, is scaled by a factor of 10000.

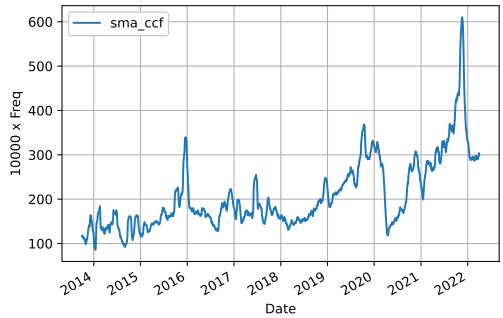

Figure 1: Relative Frequency of publication of Climate Change related articles from September 2013 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), Trumps Election and withdrawal from Paris Agreement (November 2016 and June 2017), COP 24 (December 2018), Australian Bushfires (September 2019 - Jan 2020), COP 2019, California Wildfires (February 2020) and COP 26 (November 2021).

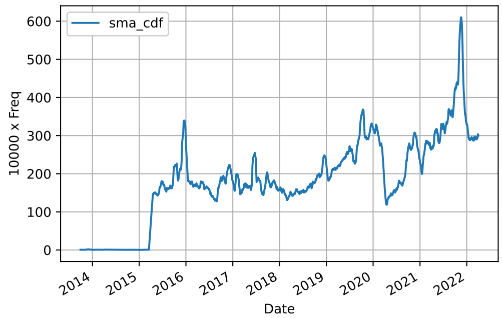

Figure 2: Relative Frequency of publication of Enviromental Crimes and Disasters, related articles from March 2015 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), Trumps Election and withdrawal from Paris Agreement (November 2016 and June 2017), COP 24 (December 2018), Australian Bushfires (September 2019 - Jan 2020), COP 2019, California Wildfires (February 2020) and COP 26 (November 2021).

The plots of the relative share indices can be observed in figures 1 - 4.

From the trends presented we observe a general increase in global news attention towards Climate Change, over the period analyzed.

A more specific trend, also present on all the indices, is the tendency for there to be a higher relative frequency of publication towards the last quarter of the year, with most peaks happening in the months of November and December.

These peaks are generally related to international climate conferences, namely the United Nations Framework Framework Convention on Climate Change Conference of Parties (COP) meetings, an event of major international importance as a sizeable portion of world leaders, academics, celebrities, and other prominent figures engage in climate change debates and speeches. This causes news outlets to steer more focus on news related to climate change and international climate agreements, for the duration of these conferences. These talks could also be a catalyst for publications to address local climate and environmental issues and policies, further raising the attention to the topic during the adjacent times.

All the time series plotted are influenced by these events to some extent.

One of the most prominent peaks is the one related to the COP21 and the signing of the Paris

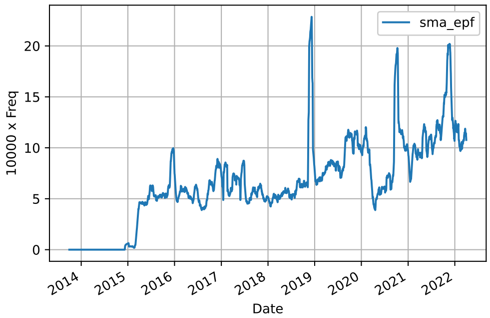

Figure 3: Relative Frequency of publication of Environmental Policy related articles, from April 2015, until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and debates (June-November 2020) and COP 26 (November 2021).

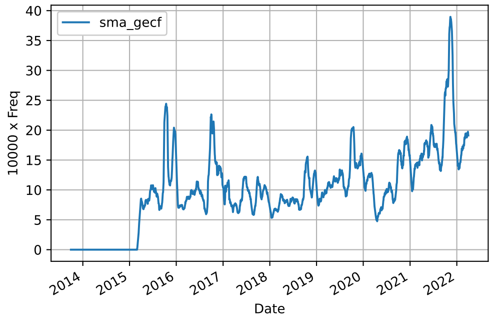

Figure 4: Relative Frequency of publication of Green Economics related articles from March 2015 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and Debates (June-November 2016) and COP 26 (November 2021).

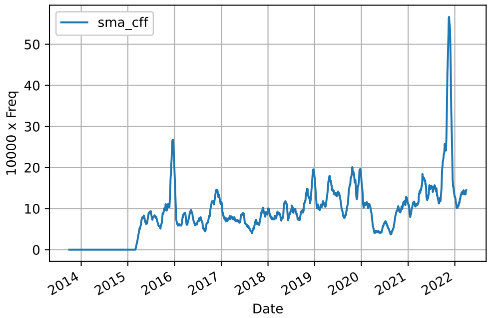

Figure 5: Relative Frequency of publication of Climate Finance related articles from September 2013 until April 2022.

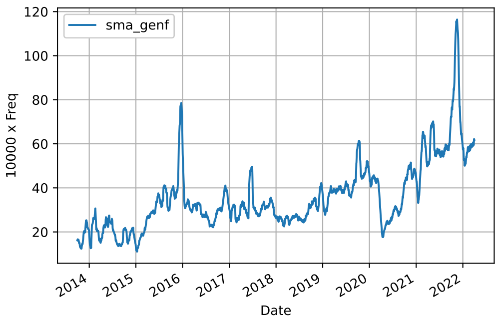

Figure 6: Relative Frequency of publication of Green Energy related articles from September 2013 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and debates (June-November 2020) and COP 26 (November 2021).

Agreement (30th November - 12th December 2015). At the time of occurrence, this was a very important international climate conference, as it established a multilateral deal between countries with the aim of limiting global warming to less than 2 $°C$ degrees by 2100. This peak is present in all the indices plotted.

The election of Donald Trump, in November 2016, caused a significant increase in the publication on the topics of Climate Finance (figure 5) and Green Economics (figure 4), due to his stance on climate change, a policy proposal, and views about the economy.

In the periods between the "COP-Peaks", there is no clear dominating topic, with exception of 2017, where on July 1st, the president of the United States announced the withdrawal from the Paris Agreement. This is particularly visible in the Renewable Energy presented in figure 6.

In the year 2020, there was a temporary decrease in the relative share of Climate Change as the COVID-19 epidemic took up a vast majority of media attention during the middle part of the year. At the end of that, year there was a resurgence to the end of the year. This peak is associated with the outcome of the US elections, and the 5th anniversary of the Paris Agreement, where heads of state reiterated their commitments to the goals of the previous agreement.

The peaks relative to 2020 are visibly noticeable on the indices related to Green Economy and Environmental Policy (figures 4 and 3), this being majorly related to the process and outcome of the 2020 United States elections, as both final candidates had opposite views with regards to environmental and climate policy.

The year 2021, shows the biggest peak in news attention towards climate change, this was once again driven by the COP 26 meeting. This year, the news picked up on the COP-related discourse earlier than in previous years, but also a major climate-related policy package passed congress in the US. Furthermore, an agreement was signed between US and China, encouraging both countries to cooperate in the development of regulatory frameworks and standards to limit greenhouse emissions and foster greener growth for both economies.

4.2 Sentiment Analysis Plots

Figures 7 and 8 show the sentiment measured on the 2 time-series, referent to the only two categories with a large enough samples of articles retrieved ("Climate Change" and "Environmental Crimes and Disasters").

Both the indices follow a similar trend in their plotted domain. This is expected as 89% of the articles in the total corpus belong to the category of "Environmental crimes and disasters".

The plotted indices show a general trend that the average sentiment of climate change news is neutral to slightly positive. This behavior is comparable to the one observed in the Environmental Policy Sentiment index, presented in Noailly et al., 2021. The developed sentiment indices also present a decrease in sentiment, starting with the 2016 US Election and remaining lower until the end of 2020. This coincides with Donald Trump's term as US president, which could be indicative of the weight of the political leadership in America, in the global Climate Change discourse. This is confirmed by the monthly mean sentiment values of the events identified in the relative frequency plots, displayed in table 3.

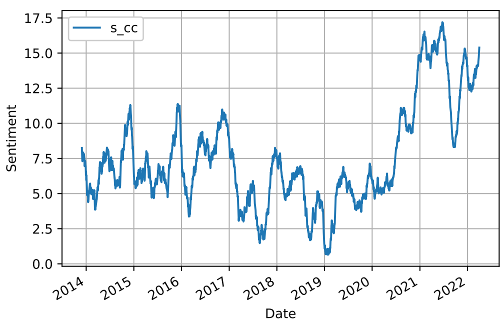

Figure 7: Measured daily average Sentiment of 'Climate Change' news.

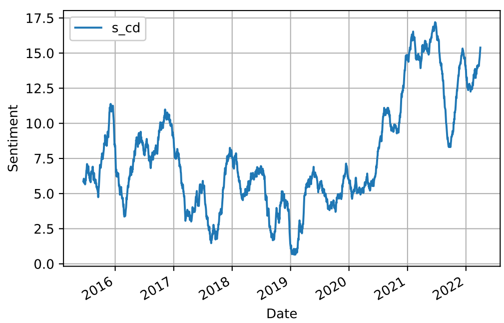

Figure 8: Measured daily average Sentiment of "Environmental Crime and Disaster"

Table 3: Summary table of monthly average sentiment for sentiment indices for general "Climate Change" (sCC, period mean = 7.44) and "Environmental Crimes and Disasters" (sCD, period mean=7.44).

Time frame	Event	sCC	sCD
12-2015	COP-21 (Paris Agreement)	10.42	10.42
11-2016	Trump Election	9.76	9.12
7-2017	U.S. Paris Agreement withdrawal	4.48	4.48
12-2018	COP-24	3.52	3.52
12-2019	COP-25 + AU Bushfires	6.47	6.47
11-2020	Biden Election	10.71	10.70
11-2021	COP-26	13.27	13.27

5 Conclusion

This paper concerns the textual analysis of Climate Change News obtained from GDELT, creating indices with aim of quantifying the relative share and sentiment of Climate Change-related articles in global news. These indices capture global trends and climate change-related events, ranging from September 2013 until April 2022.

It was found that the general Climate Change Index captures the effects of large environmental catastrophes, international climate conferences or agreements, and, US elections and political decisions. Even though the scale and impact of US politics are vast, the fact that all data used is strictly in the English language may be a source of bias, leading to some over-representation of such events.

The methodology applied to create the indices relies on the labeling system, provided by the GDELT GKG 1.0 database, and on two neural language models to approximate the measured sentiment. Despite the heavy reliance on an unsupervised learning approach, the model used for advanced filtering still holds a respectable performance, indicative of future capabilities of these kinds of models when fine-tuned on domain-specific language.

Even though the scale to which the sentiment indexes are calculated was limited by time, the end results still deliver a sentiment index that includes 12 times more articles than the most information-dense index presented in the literature. This allows for the indices to be reported on a daily basis, which opens doors to higher frequency analysis of economic events. Future works can no explore the high-frequency effects such as the presence of volatility clusters between green equity (or green bond) markets and global climate change news.

Another possible, future application of this index would be to see if the finding of Noailly et al., 2021, with respect to assets under management of US renewable energy exchange-traded funds, are also found in global green bond exchange-traded funds.

References

Barbaglia, L., Consoli, S., & Manzan, S. (2021). *Forecasting with Economic News* (SSRN Scholarly Paper No. ID 3698121). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3698121

Bingler, J. A., Kraus, M., & Leippold, M. (2021). Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures. https://doi.org/10.2139/ssrn.3796152

Bodas-Sagi, D., & Labeaga, J. (2016). Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*(Special Issue on Big Data and AI). Retrieved June 27, 2022, from https://www.ijimai.org/journal/bibcite/reference/2536

Bohr, J. (2020). "Reporting on climate change: A computational analysis of U.S. newspapers and sources of bias, 1997–2017". *Global Environmental Change*, *61*, 102038. https://doi.org/10.1016/j.gloenvcha.2020.102038

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021, September 1). *Business news and business cycles* (SSRN Scholarly Paper ID 3446225). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3446225

Dahal, B., Kumar, S. A. P., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.*, *9*(1), 24. https://doi.org/10.1007/s13278-019-0568-8

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Distilbert-base-uncased-finetuned-sst-2-english · Hugging Face. (n.d.). Retrieved June 30, 2022, from https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

Engle, R. F., Giglio, S., Lee, H., Kelly, B. T., & Stroebel, J. (2019, January 17). *Hedging climate change news* (SSRN Scholarly Paper ID 3317570). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3317570

Gentzkow, M., Kelly, B. T., & Taddy, M. (2017, February 15). *Text as data* (SSRN Scholarly Paper ID 2934001). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.2934001

Gentzkow, M., & Shapiro, J. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, *78*(1), 35–71. https://doi.org/10.3982/ECTA7195

Gentzkow, M., & Shapiro, J. M. (2006). Media Bias and Reputation. *journal of political economy*, 37.

Ghanem, D., & Smith, A. (2021). What Are the Benefits of High-Frequency Data for Fixed Effects Panel Models? [Publisher: The University of Chicago Press]. *Journal of the Association of Environmental and Resource Economists*, *8*(2), 199–234. https://doi.org/10.1086/710968

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201

Kölbel, J. F., Leippold, M., Rillaerts, J., & Wang, Q. (2020, June 1). *Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure* (SSRN Scholarly Paper ID 3616324). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3616324

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global data on events, location, and tone. *ISA Annual Convention*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (tech. rep. arXiv:1910.13461) [arXiv:1910.13461 [cs, stat] type: article]. arXiv. https://doi.org/10.48550/arXiv.1910.13461

Noailly, J., Nowzohour, L., & van den Heuvel, M. (2021). *Heard the news? environmental policy and clean investments* (CIES Research Paper series No. 70-2021). Centre for International

Environmental Studies, The Graduate Institute. https://EconPapers.repec.org/RePEc:gii:ciesrp:cies_rp_70
Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Retrieved June 23, 2022, from https://aclanthology.org/D13-1170
Valhalla/distilbart-mnli-12-9 · Hugging Face. (n.d.). Retrieved June 30, 2022, from https://huggingface.co/valhalla/distilbart-mnli-12-9
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need [arXiv:1706.03762 [cs]]. https://doi.org/10.48550/arXiv.1706.03762
Comment: 15 pages, 5 figures
Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBert: A pretrained language model for climate-related text. *arXiv:2110.12010 [cs]*. Retrieved February 2, 2022, from http://arxiv.org/abs/2110.12010
Williams, A., Nangia, N., & Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference [Number: arXiv:1704.05426 arXiv:1704.05426 [cs]]. https://doi.org/10.48550/arXiv.1704.05426
Comment: 10 pages, 1 figures, 5 tables. v2 corrects a misreported accuracy number for the CBOW model in the 'matched' setting. v3 adds a discussion of the difficulty of the corpus to the analysis section. v4 is the version that was accepted to NAACL2018
Yin, W., Hay, J., & Roth, D. (2019). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach* (tech. rep. arXiv:1909.00161) [arXiv:1909.00161 [cs] type: article]. arXiv. https://doi.org/10.48550/arXiv.1909.00161
Comment: EMNLP2019 camera-ready, 10 pages
Yuan, L., Tay, F. E. H., Li, G., Wang, T., & Feng, J. (2021). Revisiting Knowledge Distillation via Label Smoothing Regularization [arXiv:1909.11723 [cs]]. Retrieved June 29, 2022, from http://arxiv.org/abs/1909.11723
Comment: CVPR2020 Oral, codes: https://github.com/yuanli2333/Teacher-free-Knowledge-Distillation

A Appendix - GDELT theme labels included per topic

- "Climate Change":
`WB_567_CLIMATE_CHANGE,`
`ENV_CLIMATECHANGE,`
`UNGP_CLIMATE_CHANGE_ACTION,`
`UNGP_CLIMATE_CHANGE_ACTION, WB_1773_CLIMATE_CHANGE_IMPACTS,`
`WB_574_CLIMATE_CHANGE_ADAPTATION,`
`WB_959_CLIMATE_CHANGE_LAW,`
`WB_747_SOCIAL_RESILIENCE_AND_CLIMATE_CHANGE,`
`WB_1839_OZONE_LAYER_DEPLETION_AND_CLIMATE_CHANGE,`
`WB_575_COMMUNITY_BASED_CLIMATE_ADAPTATION,`
`WB_1750_CLIMATE_CHANGE_ADAPTATION_IMPACTS`
- "Climate Science":
`WB_2197_ENVIRONMENTAL_ENGINEERING,`
`ENV_CARBONCAPTURE,`
`WB_571_CLIMATE_SCIENCE,`
`WB_1784_ENVIRONMENTAL_INFORMATION_MANAGEMENT,`
`WB_399_INNOVATION_FOR_GREEN_GROWTH`
- "Green Energy":
`ENV_WINDPOWER,`

ENV_SOLAR,
ENV_HYDRO,
ENV_BIOFUEL,
ENV_GEOTHERMAL,
WB_525_RENEWABLE_ENERGY

- "Green Economy":
  WB_476_GREEN_GROWTH,
  WB_2674_GREEN_JOBS,
  WB_1100_SUSTAINABLE_GROWTH,
  WB_1856_TRADE_AND_THE_ENVIRONMENT,
  WB_791_TRANSPORT_IMPACT_ON_THE_ENVIRONMENT

- "Environmental Crimes and Disasters":
  MANMADE_DISASTER_ENVIRONMENTAL_DISASTER,
  WB_1831_ENVIRONMENTAL_CRIME_AND_LAW_ENFORCEMENT,
  WB_2916_ENVIRONMENTAL_LAW_ENFORCEMENT,
  WB_2915_ENVIRONMENTAL_CRIME,
  MANMADE_DISASTER_MARITIME_ENVIRONMENTAL_DISASTER,
  SELF_IDENTIFIED_ENVIRON_DISASTER,
  WB_1936_ENVIRONMENTAL_CRIME_ENFORCEMENT

- "Environmental Policy and Management": WB_901_ENVIRONMENTAL_SAFEGUARDS,
  WB_849_ENVIRONMENTAL_LAWS_AND_REGULATIONS,
  WB_1785_ENVIRONMENTAL_POLICIES_AND_INSTITUTION,
  ECON_DEVELOPMENTORGS_UNITED_NATIONS_ENVIRONMENT_PROGRAMME,
  WB_1782_ENVIRONMENTAL_AGREEMENTS_AND_CONVENTIONS,
  WB_1783_ENVIRONMENTAL_GOVERNANCE,
  WB_2307_ENVIRONMENTAL_MANAGEMENT_AND_MITIGATION_PLANS,
  WB_2306_ENVIRONMENTAL_IMPACT_ASSESSEMENT,
  WB_1376_ENVIRONMENTAL_OFFSETS,
  WB_157_ENVIRONMENTAL_WATER_USE_AND_CATCHMENT_PROTECTION,
  WB_1378_PAYMENT_FOR_ENVIRONMENT_SERVICES,
  WB_2195_ENVIROMENTAL_IMPACT_ASSESSMENT,
  WB_598_ENVIRONMENTAL_MANAGEMENT,
  WB_526_RENEWABLE_ENERGY_POLICY_AND_REGULATION,
  WB_1057_SUSTAINABLE_FOREST_MANAGEMENT

- Climate Finance:
  WB_1847_CLIMATE_FINANCE,
  WB_1844_MARKET_BASED_CLIMATE_CHANGE_MITIGATION,
  WB_573_CLIMATE_RISK_MANAGEMENT,
  WB_1849_PUBLIC_CLIMATE_FINANCE,
  WB_1850_PRIVATE_CLIMATE_FINANCE,
  WB_1838_CLIMATE_RISK_SCREENING,
  WB_582_GREENHOUSE_GAS_ACCOUNTING

B Enlarged Images

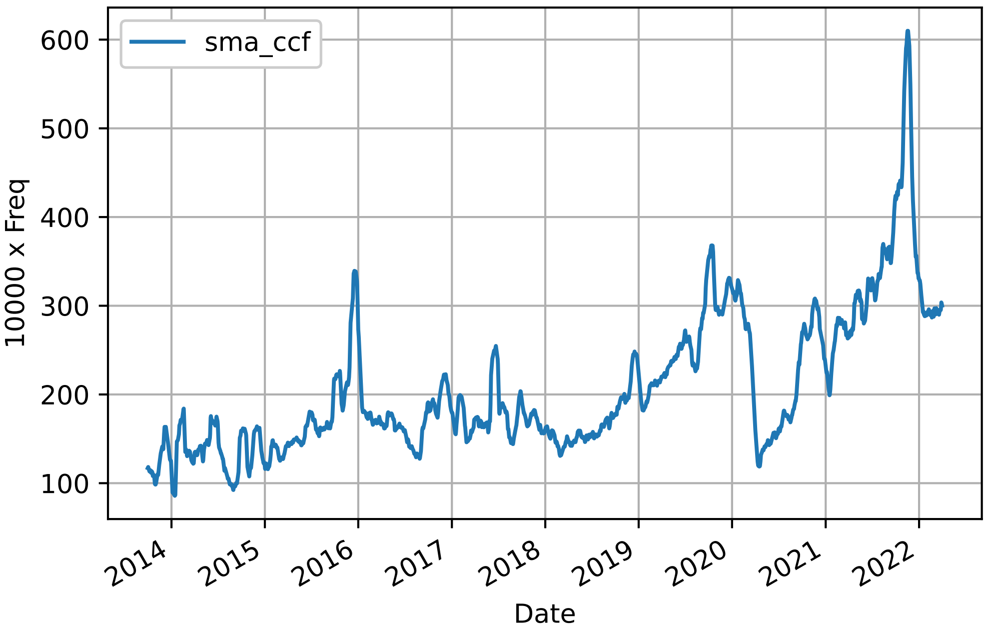

Figure 9: Relative Frequency of publication of Climate Change related articles from September 2013 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), Trumps Election and withdrawal from Paris Agreement (November 2016 and June 2017), COP 24 (December 2018), Australian Bushfires (September 2019 - Jan 2020), COP 2019, California Wildfires (February 2020) and COP 26 (November 2021).

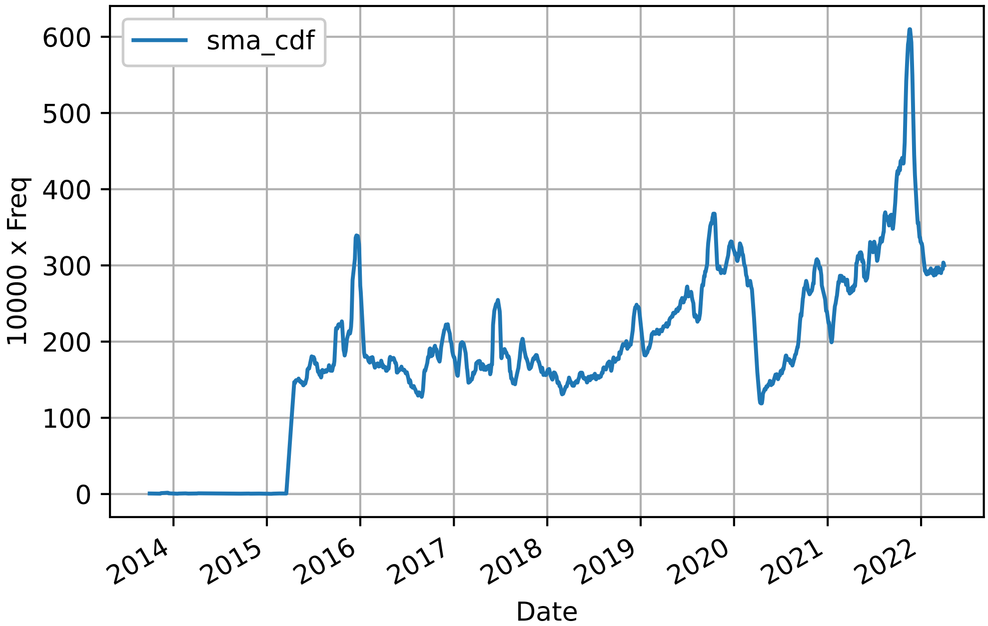

Figure 10: Relative Frequency of publication of Enviromental Crimes and Disasters, related articles from March 2015 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), Trumps Election and withdrawal from Paris Agreement (November 2016 and June 2017), COP 24 (December 2018), Australian Bushfires (September 2019 - Jan 2020), COP 2019, California Wildfires (February 2020) and COP 26 (November 2021).

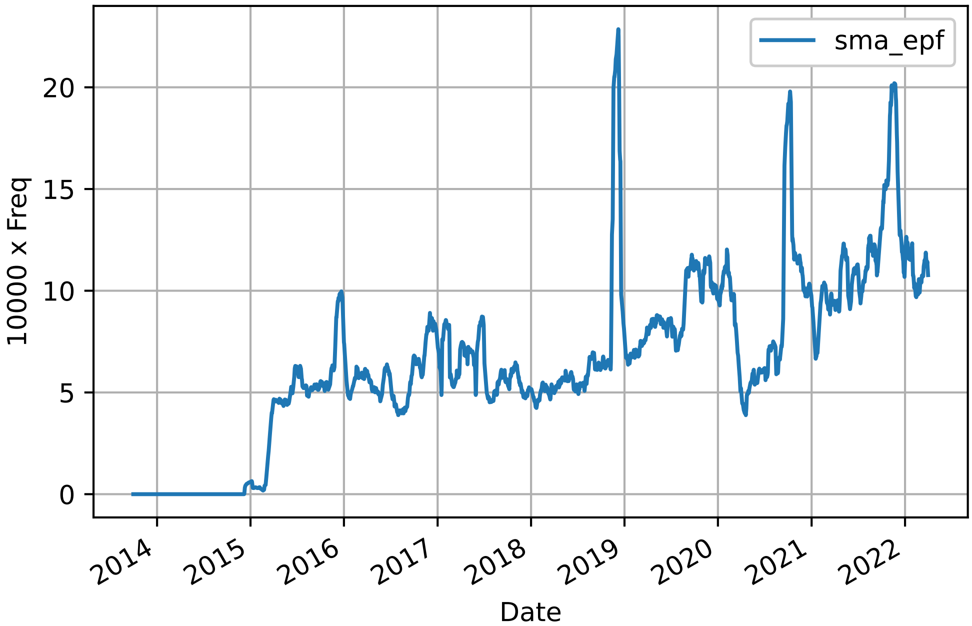

Figure 11: Relative Frequency of publication of Environmental Policy related articles, from April 2015, until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and debates (June-November 2020) and COP 26 (November 2021).

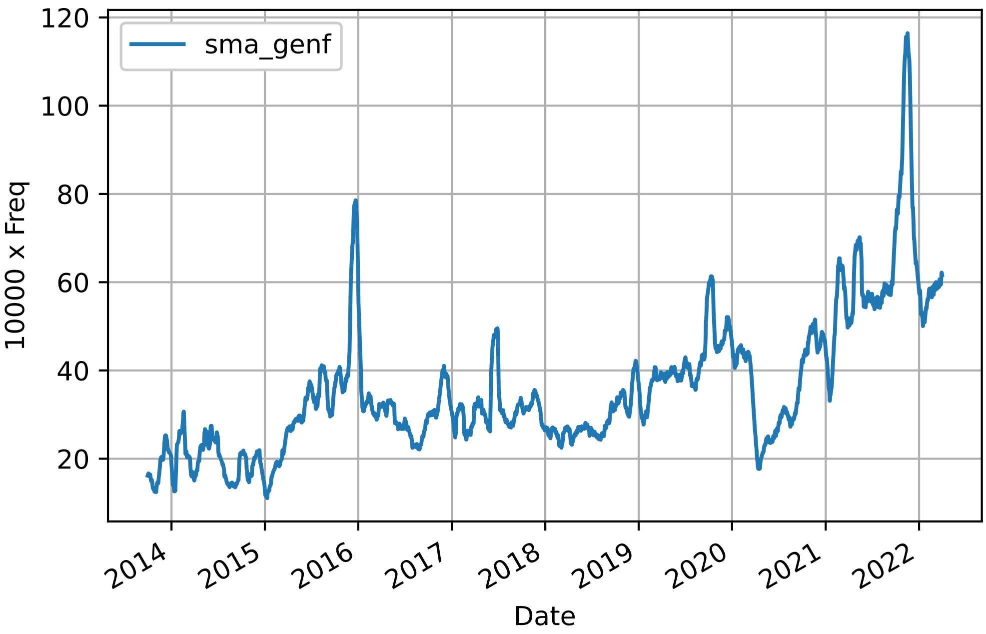

Figure 12: Relative Frequency of publication of Green Energy related articles from September 2013 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and debates (June-November 2020) and COP 26 (November 2021).

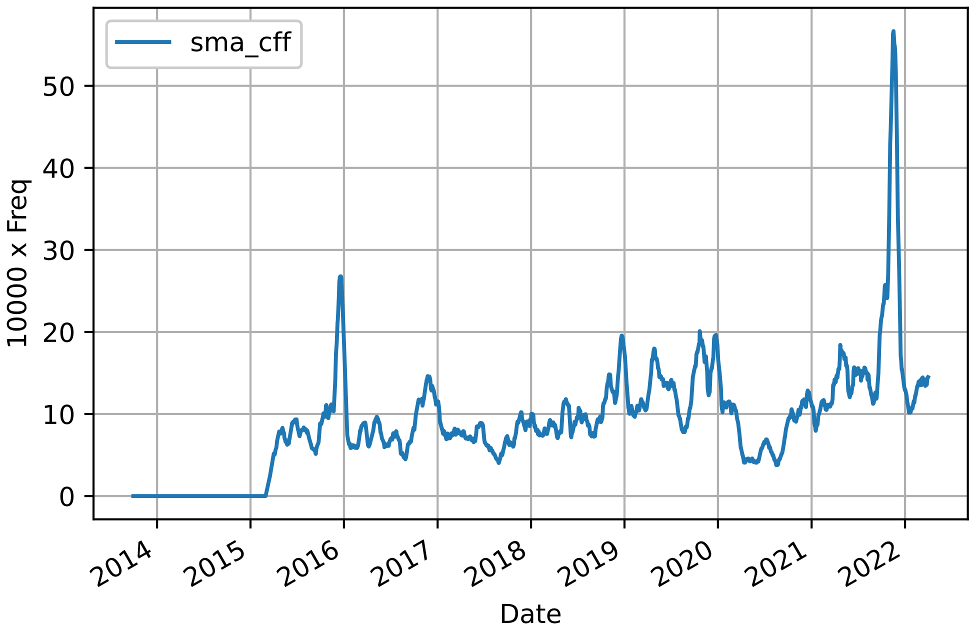

Figure 13: Relative Frequency of publication of Climate Finance related articles from September 2013 until April 2022

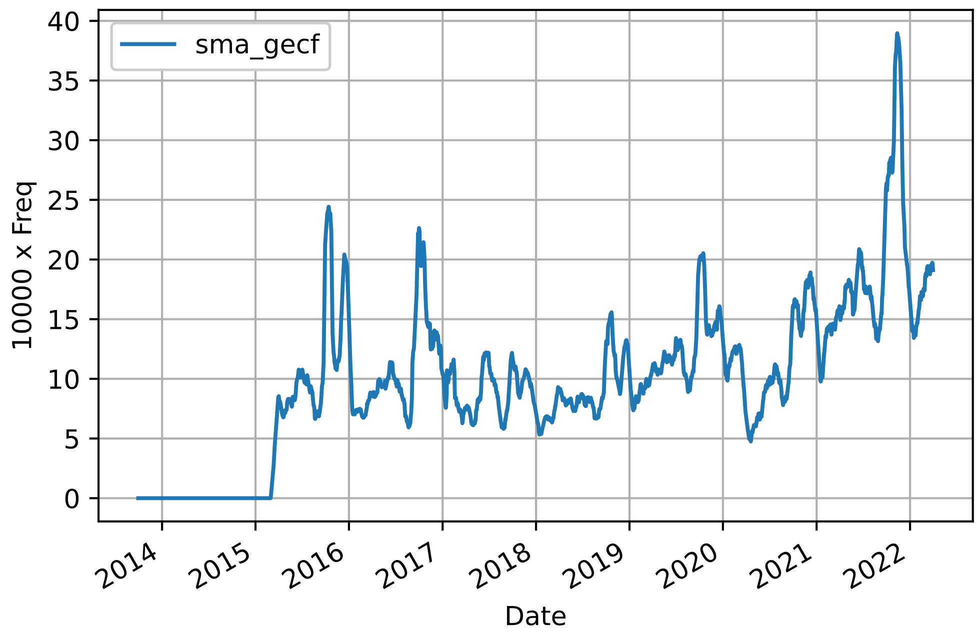

Figure 14: Relative Frequency of publication of Green Economics related articles from March 2015 until April 2022. The most visible peaks are related to the COP21 and Paris Agreement (December 2015), COP 24 (December 2018), American Elections and Debates (June-November 2016) and COP 26 (November 2021).

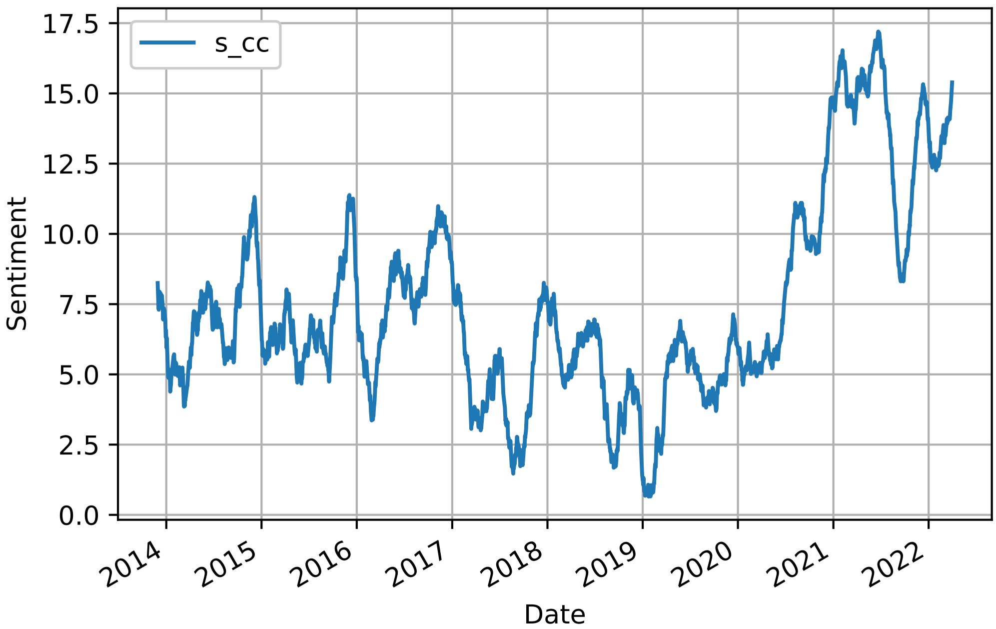

Figure 15: Measured daily average Sentiment of 'Climate Change' news.

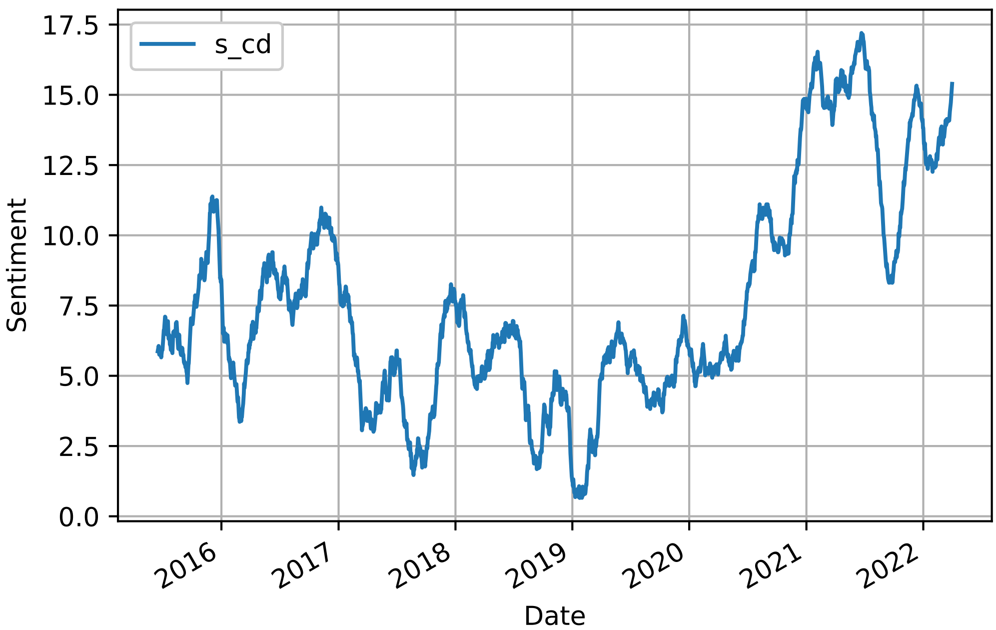

Figure 16: Measured daily average Sentiment of "Environmental Crime and Disaster"