The effect of climate protests on issue framing in UK press coverage
Final Project Report

Dario Siegen
dario.siegen@hotmail.com

Nikki Wirawan
nikki.wirawan@gmail.com

Samuel Ribanský
s.ribansky@mpp.hertie-school.org

Abstract

In this report, we analyse whether and how environmental protests have changed the way the media reports on climate change. Our research question is therefore defined as follows:

How have climate protest movements in the UK changed climate change coverage in the national press?

To investigate this, we analyse approximately n=5000 media articles from seven different major UK newspapers between 2010 and 2019. In doing so, we use a similar study as our baseline and perform an automated framework analysis. We rely on structural topic modelling (STM) and principal component analysis (PCA) for identifying the change in topical prevalence in the articles. We use 'topics' and 'frames' interchangeably. We advance the topical analysis by accompanying it with a sentiment analysis of the coverage.

We find that the Climate Strike, Fridays for Future, and Extinction Rebellion have influenced the media coverage of climate change. In terms of topics, we observe a twofold increase in the Expected Topic Proportion of topics relating to climate crisis and extreme weather events. We observe that these follow a significant increase in the coverage of the respective climate protests. With respect to sentiment, we observe a more positive sentiment following the signing of the Paris Agreement and negative sentiment during the coverage following the climate protests. Lastly, we recognise that our results may be biased given that 30% are from the Guardian. Nevertheless, when controlled for the news outlet covariate, our results on topic modelling remain statistically significant to a varying degree.

The exact code applied in our analysis, as well as detailed results, can be found on GitHub.[1]

1. Introduction

Over the last year, environmental movements around the globe have gained momentum. In particular, the climate strikes by Greta Thunberg and many other young people have mobilised millions of people to the streets. In addition, a more radical group of environmental activists, Extinction Rebellion, particularly in England but also internationally, has drawn attention to the climate crisis through non-violent civil disobedience. Our research concerns the impact of these protests on the coverage of climate change in mainstream newspaper outlets.

There is an ongoing debate among scholars as to whether the "doom and gloom" narrative is beneficial or harmful in the pursuit of raising public awareness and support for climate action [6]. Communication science has long agreed that the apocalyptic narrative does not activate people to change behaviour. This is in contrast to the climate activists' latest mobilisation of people through an apocalyptic narrative (e.g. Greta Thunberg "I want you to panic" or Extinction Rebellion "we are in the midst of a mass extinction of our own making"). Therefore, this project focuses on the following research question (RQ):

How have climate protest movements in the UK changed climate change coverage in the national press?

We argue that the doom and gloom narrative of the protesters has influenced both the framing of the issue and the sentiment of the coverage.

To properly unpack the argument, this paper is structured as follows. Firstly, we provide a brief overview of the literature on PCA, STM and sentiment analysis, to explain our methodological approach. Next, we outline our method and introduce the dataset used for this analysis. Thirdly, we present our results and the shortcomings of our research design, offer avenues for further research, and conclude. Lastly, it is important to note that this is not a methodological paper. The purpose of this paper is to apply existing methods onto a research question that has not yet been answered, rather than to propose new methodologies. Nevertheless, we also offer suggestions on how to improve on the existing methods.

The novel contribution of this paper rests in its application of automated framing analysis to identify the dominant

[1] https://github.com/s-ribansky/HertieProject/

frames in a newspaper coverage of a contemporary news topic inductively.

2. Proposed Method and Related Work

We analyse the research question from three angles. Analyses have already shown that British reporting on climate change has increased sharply in absolute terms in recent months ([2]. First, we want to check whether reporting has increased not only in absolute terms, but also relative to overall reporting:

- 1) Has climate change reporting increased in relative coverage in the British press?

However, we do not yet know exactly how these media articles on climate change report on the phenomenon. We would therefore like to examine this in more detail. Our research question is accordingly:

- 2) Has the framing of climate change changed in the British press and if so, how?
- 3) Has the sentiment of the articles changed over time and if so, how?

This paper utilise a mix of inductive and deductive analyses as defined by Boumans and Trilling [4] and applied to a time-series analysis of the framing of the 2015 refugee crisis in Austrian broadsheet and tabloid media by Gruessing and Boomgaarden [7].

In their project, they [7] analysed 5.321 articles pertaining to the refugee crisis that were published by six different Austrian news media outlets. The authors first applied deductive analytical methods, specifically the Principal Component Analysis (PCA), to identify the dominant frames. A PCA analyses how different words (tokens) are related to each other within and across all documents (corpus). More closely related words are then classified into frames and these frames, in turn, can explain parts of the variance of the overall calculation [1].

Principal component analysis transforms the original data, the most frequently used words of all articles, into a smaller number of as uncorrelated as possible linear variables that are interpreted as *"latent attributes of the data"* (frames) [7]. They can be described as *"networks of co-occurring words, constituting the semantic patterns in which words are used, and capturing the underlying structures that provide meaning to a text"* [7].

Automated framing analysis is a relatively novel undertaking and is in the phase of being tested on large data sets with easily predictable results [4]. Previously, framing analyses were labour intensive, as they required a large number of annotators to read through text corpora and identify the frames. With automated framing analysis, researchers can now explore large textual datasets requiring less time and money.

In applying this technique to our dataset of UK articles mined from the selected news outlets, we replicate the technique used by Greussing and Boomgaarden [7] in order to answer RQ2. We then manually analyse and categorise the results of the PCA.

We supplement this analysis using STM estimates. To analyse whether there was an observable change in the framing of climate change coverage following the kick-off of Fridays for Future and Extinction Rebellion protests, we look at the variance in the expected topic proportion of various topics over time. We use the **stm** package to i) identify the topics dominant throughout the coverage and ii) analyse how the relevant topics developed over time and whether there was an observable change following the first protests. We assume the beginning of Q2 2014 as the breaking point. This means that we would expect an observable and robust change in framing and sentiment from Q2 2018 onwards.

To identify the topics and assess their relevance for our research purposes, we use topical prevalence models that identifies the expected share of a topic in a document. To avoid spurious results, we introduce the variable 'medium' as a covariate. Next, we introduce the variable 'date' as an additional covariate. 'Date' describes the day an article was written. As the **stm** function works with continuous and factor variables, we converted the 'date' variable into a continuous numeric variable using the **as.POSIXct** function. Consequently, we estimated two models: **date stm prevalence medium date** and **date stm prevalence medium date2**. The former is a model where we stipulate the desired number of topics as K=0. This prompts the model to estimate the most viable number of topics for the given number of articles and covariates on the basis of probability, frex, lift, and store indicators [CITATION].This model yielded us 119 topics, which we used to validate the results of the estimation of 25 most prevalent topics using the second model.

Having ran both models, we closely inspected the topics identified by the model and identified the pertinent topics. These were topics 11 and 13 (relating to extreme weather events), 16(Paris Agreement), 18 (transport – both road and non-road), and 22 (protest movements) for the **date stm prevalence medium date2** model. For the **date stm prevalence medium date**, we identified topics 11 and 30 (protest movements) and 21 and 70 (global and UK climate crisis respectively). It is worth noting that the topics were chosen amid our arbitrary judgment, where we chose topics that i) related to the civic action and ii) a radical outlook on either the problems or solutions to climate change, and iii) mainstream coverage, amid our research question estimating the effect of (i) on (ii) and (iii) respectively. This part of the analysis was prepared and analysed by Samuel, using inspiration from Roberts, Stewart, and Tingley.

With respect to RQ3, we will apply simple dictionary methods to identify the sentiment of the public discourse as portrayed in the media. We first identify the sentiments of the different articles already classified by different frames. We will then look at whether a change was observed in the sentiment over time. A key disadvantage of using dictionaries for sentiment analysis is context specificity, which renders some dictionaries applicable to some contexts, but not to others [4]. We will commence by applying the 'afinn' dictionary, which includes sentiment words identified in Twitter discussion on climate change, then apply the 'nrc' dictionary, which was compiled by using responses from Amazon Mechanical Turk, and compare the results using qualitative methods.

The novel contribution of this paper rests in its application of automated framing analysis to identify the dominant frames in a newspaper coverage of a contemporary news topic inductively. Moreover, this paper then uses deductive methods to apply the frames to the dataset and measure which frames were dominant in which types of articles. While the impact of the climate protests on the quantity of news coverage has been analysed ([2], such a quantitative framing analysis on the deeper impacts of the protests on reporting are lacking.

3. Experiments

This section describes the data used for this project and provide a detailed outline of the host of pre-processing steps that we undertook to transform the collected articles into a dataset that can be analysed using quantitative methods. Next, we introduce the evaluation criteria we used to validate our results. Following the multitude of methods used in this paper, each step of the analysis has its own evaluation methods. We then describe our analysis step-by-step. The R script is available in the GitHub repo provided in the introduction. Lastly, we provide and overview of the results. The analysis of our results and avenues for future research are reserved for subsequent sections.

Overall, our results match our expectation, albeit not the degree of robustness and explicitness that we originally anticipated. We attribute this to the low-quality of the data, where unwanted elements may have distorted our analyses.

Methodologically, we oriented our process very closely to the Greussing and Boomgaarden paper [7] for the PCA and to Roberts, Stewart, and Tingley. However, due to the format of the data exported from Factiva, we had to develop an entirely new line of code to create a usable dataset, the analysis of which can yield meaningful insights. In order to pre-process our data, we rely on standard R functions to remove stopwords, punctuation, uppercase letters, as well as *n-grams* as we deemed fit. This section summarises the detailed process that has been followed up to now.

Table 1	
Name	*Type*
Daily Mail	Tabloid
The Guardian	Broadsheet
The Times and Sunday Times	Tabloid
The Sun	Tabloid
Daily Express	Tabloid
The Telegraph	Broadsheet
The Independent	Broadsheet

Table 1. The seven UK newspaper outlets used for data collection

Data and data treatment

Data: We used the research tool "Factiva" to collect news articles related to media coverage of climate change and global warming issues in the United Kingdom. We applied the following screening mechanisms:

- The articles must mention the terms "climate change", "global warming", "climate crisis", "climate emergency", "climate breakdown" or "global heating" at least three times in either the titles or bodies. We increased the minimum frequency from Greussing and Boom's [7] methodology. This was done to make sure that we filter articles that mainly discuss those terms.

- We analyse articles from a selection of eight media outlets mentioned in Table 1. These outlets are ranked in the top 10 media with the highest readership according to the UK's newspaper and readership statistics in 2018.

- The articles were published in a period between January 2010 to September 2019. We wanted to make sure that we could capture both the effect of the Paris Agreement in 2015, the Extinction Rebellion protests in 2019, and also the trend before, between, and after those events.

For the news agencies and location, we filtered manually using the dropdown menu provided. We also switched on the duplicate option to "similar". Notably, we include both non-partisan and traditionally partisan news outlets, for both types shape the public discourse, making the analysis of their content desirable for understanding public attitudes. The result showed that there are **5321** articles that were published within the specified time period.

The Factiva license provided by the Hertie School of Governance does not include the possibility to export articles in an XML format, which would have made pre-processing, including the allocation of document variables ('docvars'), much simpler. Consequently, we exported the articles in a HTML format. This form of exporting data yielded us dozens of PDF files containing approximately 100 articles per file. Next, we separated the articles into .txt files, with one article per file.

After separating the articles, we imported them as individual text files using the "readtext" function. In that process, we defined text content and assigned the articles' metadata and docvars. While conceptually a routine operation, the low level of comparability of the individual document rendered this process more complex than initially envisioned. More specifically, we encountered the following three issues that we attempted to solve:

1. **Extracting metadata from lines:** The first issue we encountered even before beginning the treatment of our data was that the metadata of the articles (author, source, article title, word count, etc.) were contained in the head sections of the structured .txt documents, separated by line rather than node or column separators. However, standard R tools for the treatment of .txt documents either extract the metadata from the document name (docnames) or directly from the file based on columns or nodes. Given that the metadata in our documents were located in the body of the text and separated by line separators, we had to run several lines of code to separate the whole documents by line, extract the metadata, and then merge lines that form the body of the article back together to re-assemble the actual data point to be used for our analysis.

2. **Harmonising metadata from lines:** Following from the preceding paragraph, an additional issue that emerged was the lack of harmonisation between the format of the different .txt files. For example, some .txt documents would not contain the name of the article, whereas other documents would include extra lines containing variables that we could not assign meaning to. We also identified several outlet-specific variables that were present in articles from one outlet, but not others. For example, articles from The Guardian contained a specific 'copyright' line in the heading section, which would be located on the same line as the Factiva-assigned outlet code (i.e. GRDN for The Guardian) for other outlets. This further hindered our ability to apply a simple string of code.

3. **Harmonising text bodies:** The last main issue that spread across our dataset was that some of the .txt documents' text bodies were not following the general pattern of paragraphs but rather included odd symbols or no text body at all . Ultimately, we solved this problem by manually inspecting and treating the problematic files, e.g. clearing the "noise", such as website links, copyright statements and other terms that don't belong in the text body. However, due to the time intensity of this type of data treatment, we dropped the worst cases of inconsistent documents from our analysis.

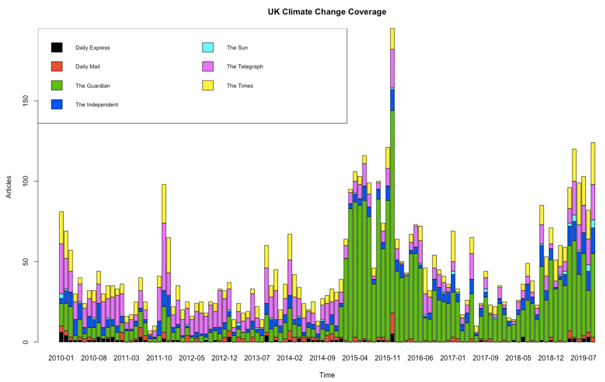


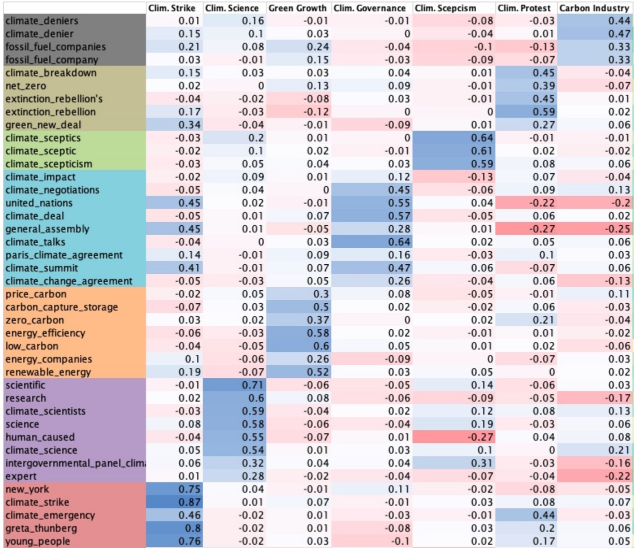
Figure 1. The number of UK articles mentioning climate change terminology at least three times over time (per month and medium.

For the training set, we imported 300 single media articles as .txt files. This process resulted in a data frame with the following variables: "title", "author", "wordcount", "day","month", "year", "time", "medium", "source", "language", "source2", "text". We then created a data corpus using the function "corpus" and created token (function "token") variables that we would use for the PCA.

The Principal Component Analysis: In their analysis, Greussing and Boomgaarden [7] adopted a number of measures to identify the most dominant terms in reporting. These include normalizing letters, eliminating stop words, and stemming. Using tf-idf (term frequency-inverse document frequency) values, the most important terms were then identified [9]. From the remaining 89 terms, 8 frames were then identified using a PCA and their temporal distribution on the data set was examined.

Therefore, in line with Greussing and Boomgaarden [7], we removed punctuation, hyphens, numbers, single letters, stop words, and performed stemming (see Fig. 1 for illustration). Furthermore, we lower-cased all letters. For that process, we mostly used the packages "tm" [5] and "SnowballC" [3].

For the PCA, we included all terms that appeared at least 500 times in the test corpus (300 articles) and identified the 500 most frequent tokens. However, our training set still includes certain bug-tokens such as "get" or "pic.twitter.com". These most frequent tokens still require



	Clim. Strike	Clim. Science	Green Growth	Clim. Governance	Clim. Scepcism	Clim. Protest	Carbon Industry
climate_deniers	0.01	0.16	-0.01	-0.01	-0.08	-0.03	0.44
climate_denier	0.15	0.1	0.03	0	-0.04	0.01	0.47
fossil_fuel_companies	0.21	0.08	0.24	-0.04	-0.1	-0.13	0.33
fossil_fuel_company	0.03	-0.01	0.15	-0.03	-0.09	-0.07	0.33
climate_breakdown	0.15	0.03	0.03	0.04	0.01	0.45	-0.04
net_zero	0.02	0	0.13	0.09	-0.01	0.39	-0.07
extinction_rebellion's	-0.04	-0.02	-0.08	0.03	-0.01	0.45	0.01
extinction_rebellion	0.17	-0.03	-0.12	0	0	0.59	0.02
green_new_deal	0.34	-0.04	-0.01	-0.09	0.01	0.27	0.06
climate_sceptics	-0.03	0.2	0.01	0	0.64	-0.01	-0.01
climate_sceptic	-0.02	0.1	0.02	-0.01	0.61	0.02	-0.02
climate_scepticism	-0.03	0.05	0.04	0.03	0.59	0.08	0.06
climate_impact	-0.02	0.09	0.01	0.12	-0.13	0.07	-0.04
climate_negotiations	-0.05	0.04	0	0.45	-0.06	0.09	0.13
united_nations	0.45	0.02	-0.01	0.55	0.04	-0.22	-0.2
climate_deal	-0.05	0.01	0.07	0.57	-0.05	0.06	0.02
general_assembly	0.45	0.01	-0.05	0.28	0.01	-0.27	-0.25
climate_talks	-0.04	0	0.03	0.64	0.02	0.05	0.06
paris_climate_agreement	0.14	-0.01	0.09	0.16	-0.03	0.1	0.03
climate_summit	0.41	-0.01	0.07	0.47	0.06	-0.07	0.06
climate_change_agreement	-0.05	-0.03	0.05	0.26	-0.04	0.06	-0.13
price_carbon	-0.02	0.05	0.3	0.08	-0.05	-0.01	0.11
carbon_capture_storage	-0.07	0.03	0.5	0.02	-0.02	0.06	-0.03
zero_carbon	0.03	0.02	0.37	0	0.02	0.21	-0.04
energy_efficiency	-0.06	-0.03	0.58	0.02	-0.01	0.01	-0.02
low_carbon	-0.04	-0.05	0.6	0.05	0.01	0.02	-0.06
energy_companies	0.1	-0.06	0.26	-0.09	0	-0.07	0.03
renewable_energy	0.19	-0.07	0.52	0.03	0.05	0	0.02
scientific	-0.01	0.71	-0.06	-0.05	0.14	-0.06	0.03
research	0.02	0.6	0.08	-0.06	-0.09	-0.05	-0.17
climate_scientists	-0.03	0.59	-0.04	0.02	0.12	0.08	0.13
science	0.08	0.58	-0.06	-0.04	0.19	-0.03	0.06
human_caused	-0.04	0.55	-0.07	0.01	-0.27	0.04	0.08
climate_science	0.05	0.54	0.01	0.03	0.1	0	0.21
intergovernmental_panel_clim	0.06	0.32	0.04	0.04	0.31	-0.03	-0.16
expert	0.01	0.28	-0.02	-0.04	-0.07	-0.04	-0.22
new_york	0.75	0.04	-0.01	0.11	-0.02	-0.08	-0.05
climate_strike	0.87	0.01	0.07	-0.01	0.03	0.08	0.07
climate_emergency	0.46	-0.02	0.01	-0.03	-0.01	0.44	-0.03
greta_thunberg	0.8	-0.02	0.01	-0.08	0.03	0.2	0.06
young_people	0.76	-0.02	0.03	-0.1	0.02	0.17	0.05

Figure 2. Seven Topics found in the data and their loading values.

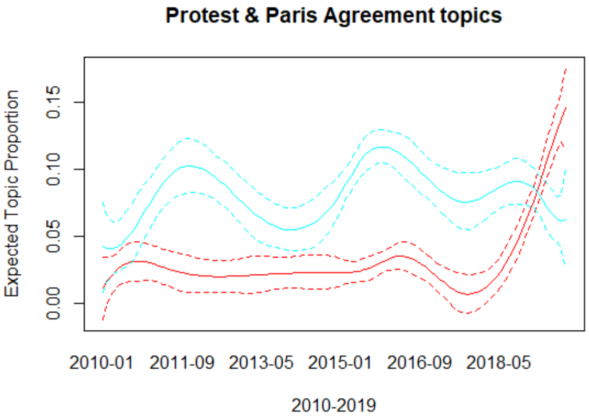


Figure 3. Comparison of the expected topic proportion of 'Topic 22' associated with civil protests and 'Topic 16' associated with the Paris Agreement as estimated by our **dtm stm prep date media2** model. The expected topic proportion of the protest topic is in red.

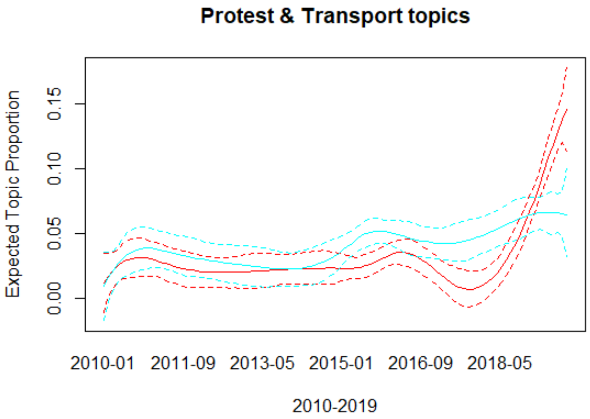


Figure 4. Comparison of the expected topic proportion of topics associated with civil protests and climate and transport as per our **dtm stm prep date media2** model. The expected topic proportion of the protest topic is in red.

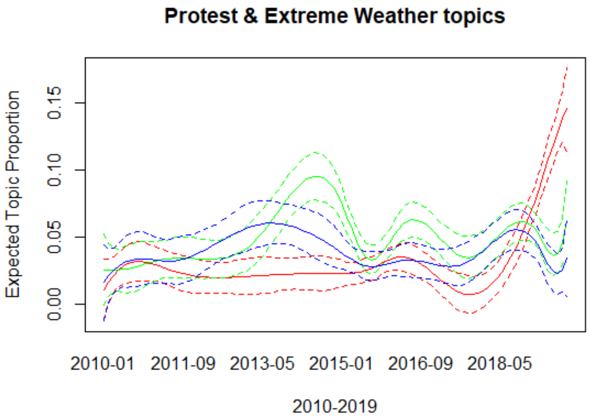


Figure 5. Comparison of the expected topic proportion of topics associated with civil protests and extreme weather occurrences (droughts, flooding) as per our **dtm stm prep date media2** model. The expected topic proportion of the protest topic is in red.

us to look at the data more closely to further improve its quality. Nevertheless, some tokens already catch the eye, such as "fossil", "johnson", "extinction", "thunberg" or "rebellion". For the final analysis, we will select only the relevant tokens (as in, those which remain after rigorous cleaning).

Evaluation of the PCA: We would like to get at least similar values for the explanation of the variance of the found frames as our model study (16 percent). At this point, it is hard to estimate whether we will reach this goal, as the data still requires better pre-processing. But we are getting close to it, the framework for the analysis stands.

Furthermore, we apply qualitative methods to explain the changes in dominating frames over time. We execute an additional framing analysis of the media releases of the protesters or a similarity metric of the same. We also analyse the news sentiment overtime. We observe whether the sentiment changes in relation to the emergence of climate movements and conferences and inspect if there are also differences in sentiments between different media outlets.

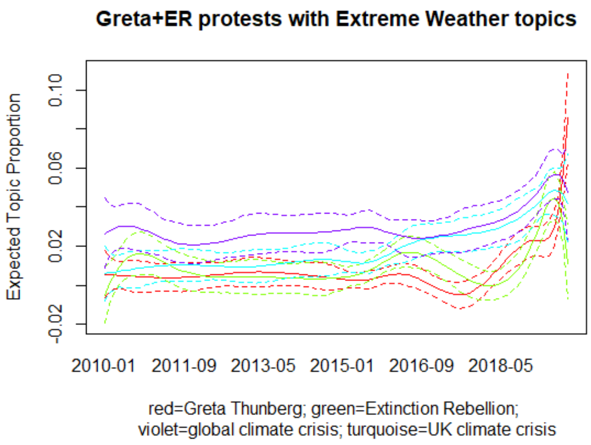


Figure 6. Comparison of the expected topic proportion of topics associated with civil protests and extreme weather occurrences as identified by the first **dtm stm prep date media** model.

Experimental details: Finding all the applicable functions to transform the raw text into a data frame and eventually into token data frames proved demanding amid the aforementioned issues. Once we had the normalized token data frame, we applied the tf-idf function to calculate the individual token's "importance" scores. We then applied the principal() function from the "psych" package to test for the presence of latent variables [8]. In doing so, we defined the number of frames as 5 (RC1 to RC5). The most logical number of frames is still to be determined, this depends on the variation the different frames are able to explain. We also tested the prcomp() function, but found its suitability for the analysis of categorical variables to be limited compared to principal(), which is designed for categorical variables specifically.

The STM analysis and evaluation: For the STM analysis, we used the **estimateEffect** function to estimate the share of individual topics as a function of their prevalence in the n=5152 articles collected. The topical prevalence was estimated for both stm models and plotted against the *'date'* variable (controlling for 'medium') for an easier interpretation of the results of our model. When plotting the results of the **estimateEffect** model, we specified the topics described above. In terms of the validity, the relationship between the topical proportion estimate and *date* is at least $p<0.05$. Consequently, we consider the results as statistically significant. However, because the available methods do not allow for testing the relationship between individual topics, the link between the coverage of the protests and the doom and gloom framing and framing is subject to our qualitative analysis.

The sentiment analysis and evaluation: In regards to sentiment analysis, we apply two different well-known sentiment dictionaries, "bing" and "afinn". Both dictionaries analyze the sentiment per word. The "bing" lexicon categorizes words in a binary fashion into positive and negative categories while The "afinn" lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment[**?**]. To apply both dictionaries we used "get-sentiments" function that we obtain from installing "tidytext" package. We then created two new dfms, one for each lexicon and apply the sentiment value by using "inner-join" function. This resulted in valuation of every word in every news article in rows (provided that the words exist in the dictionaries). Entries that contain words that are not exist in the dictionaries were removed from the rows.

For "bing" sentiment analysis, we then counted the frequency of each word appear in the articles. Then, we multiplied the frequency by +1 if it was a positive word and -1 if it was negative, producing the "sentiment value" for each word in each article. We then aggregated the result by summing the sentiment values of all words within an article, producing the "sentiment score" per article. This scores were then aggregated per media outlet per month by summing all the individual article "sentiment score". We further aggregated the data into yearly trend and later plotted both the monthly and yearly sentiments into a stacked bar charts.

Similarly, for the "afinn" sentiment analysis, we utilised the same technique. The difference is that we did not need to multiple the sentiment value by frequency since adding up the rows already did the process.

To conclude, the graphical interpretations of our results confirm our intuition regarding the relationship between civil action and media coverage as a function of topical prevalence and sentiment.

Results: Figure 2 presents a PCA table with a rough calculation of our tokens (stemmed and non-stemmed). So far, the results suggest that we have applied the right methods.

Table 2 gives an overview of the 10 most frequent stemmed tokens across the 300 test documents. The token number 341, *schools*, is an example for the data preprocessing that we still need to improve. The tokens "fossil" and "fuel" suggested that collocations were present in the corpus. A collocation analysis gave us a clearer idea of the common collocations, which we then applied to out final analysis.

Our stm model shows that the physical protests were followed by an increased coverage of the protests. Figure 3 compares the expected topic proportion of Topic 16 associated with the 'Paris Agreement' topic to the expected topic proportion of Topic 22 associated with the protest topic. Figure 4 shows that following the increased protest cov-

Selected Tokens			
Number	*Token*	*Number*	*Token*
105	”movement”	309	”students”
317	”children”	325	”strike”
347	”schools”	362	”activists”
371	”extinction”	488	”rebellion”
489	”greta”	500	”climatestrike”

Table 2. These are a few non-stemmed selected tokens related to climate movements.

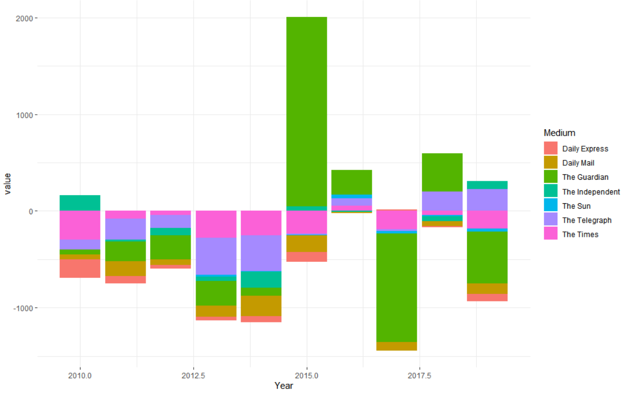


Figure 7. Accumulated Sentiment Score (bing Lexicon).

erage, historically highest proportion of the transport topic is also observed. While the evidence is circumstantial at best, we know from manually inspecting the articles associated with the topic that they indeed accredit the higher salience of the issue to civil action. Figure 5 compares the relationship between the expected topic proportion of the protest topic and the topics covering extreme weather events. It shows that while the extreme weather events were either held constant or in decline, their expected proportion is showing an increasing trend following the topics. Lastly, based on the **date stm prevalence medium date** model, we observe a trend of a rapidly increasing proportion of climate crisis topics in newspaper coverage, which coincides with the increased proportion of the expected topic proportion of the protest topics.

Table 3 shows a collection of manually selected tokens that hint towards a ”climate movement” frame of some sort. Our final analysis will be able to transform this conjecture into more tangible results.

Figure 3 and 4 show the barplots for the aggregate sentiment analysis using ”bing” and ”afinn” lexicon respectively. In a simple generalization, ”bing” sentiment shows how frequent positive and negative words are being written into the news article while ”afinn” shows the intensity of the sentiments. As we can see the results are quite different between the two lexicons.

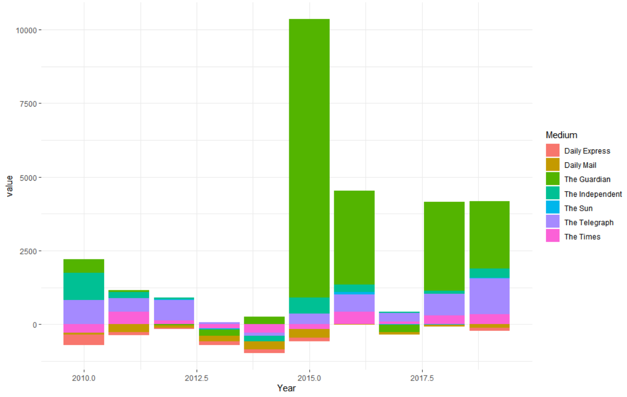


Figure 8. Accumulated Sentiment Score (afinn Lexicon).

4. Analysis

In our analysis, the content pre-processing of the original dataset left us with n 5100 article. The actual number of articles varied with individual methods, as different algorithms also employ their indigenous pre-processing techniques on the bases of preset criteria. Overall, our analysis provides interesting and useful insights for the understanding of the relationship between civil action and the framing of an issue that the civil action pertains to in newspapers. This analysis benefits from its complexity by painting a broad picture of the scale of the issue coverage, the framing of the issue coverage, and the sentiment of the issue coverage.

Nevertheless, it also has several shortcomings. First, despite efforts to improve the quality of the date, the sentiment, PCA, and STM analyses suffered from the low quality of the data. This is most evident in some of the topics and principal components including character structures that had no interpretative meaning. Second, the PCA is a method used primarily for numerical analyses. Hence its validity is subject to an ex-post qualitative assessment. Lastly, the STM would have benefited from the possibility of analysing the relationship between the individual topics. While our qualitative assessment through the manual observation of selected articles proved sufficient in determining the validity of the topics, this might become more problematic when a thematically broader or a larger set is analysed.

We can draw several conclusion from our sentiment and stm analyses:

- In the early study periods (2010 - 2014) most newspaper outlets present the climate change related news with words that have a more negative tone. This trend continues in 2015 with one big exception, The Guardian that cumulatively writes the news using positively-toned words. We believe that this is caused by a sudden increase in the number of articles written in the year that driven by the Paris Agreement.

- In the later periods (2016-2019) the sentiments are more fluctuating, but cumulatively, the intensity of the sentiments are leaning toward a more positive side, except in 2017 when it seems a bit balanced.

- From the two plots we could notice that there is a huge difference between the ”bing” and ”afinn” results. ”bing” result shows a tendency for media outlet to write news that is ”negative” while ”afinn” presents a significantly more positive sentiment. There are couple of possible reasons of why it happened:

  - ”bing” lexicon assigns sentiments to 6786 different words, while ”afinn” has a lot less words in their lexicon (2477).
  - 1598 out of 2477 words in ”afinn” lexicon (0.645) and 4781 out of 6786 words in ”bing” lexicon (0.704) are assigned as having a negative sentiment. However, in ”afinn” lexicon, the most negative words (score -5 and -4) are mostly consist of informal words that are highly unlikely to be used in formal newspaper articles (e.g. curse words). On the other hand, the most positive words (+4 and +5) are significantly more common to be found in the articles.

- Therefore, we think that the ”bing” lexicon is the more objective sentiment analysis.

- We are tempted to attribute the significant increase in positive sentiment in 2015 to the Paris Agreement and the more negative sentiments in 2019 to the Climate Strikes. However, we do not have sufficient evidence to prove it; thus, a more detailed analysis should be done in the future

- For topical prevalence, we observe a clear relationship between civil action and the prominence of certain issues as a function of topical prevalence. More specifically, we can see that the *Fridays for Future* and *Extinction Rebellion* protests resulted in the subsequent increase in the coverage of issues that these groups advocate on: extreme weather events and carbon-intensive transport. Moreover, we also observe the reduced prevalence of the Paris Agreement topic, suggesting a move away from mainstream a (for the moment) fringe discourse.

- However, we also remain aware that 30% of all articles used in for our analysis were sourced from The Guardian. While we controlled for the ’medium’ covariate, this represents a large share of the dataset nevertheless and could have potentiallly introduced some degree of bias to our paper.

- Lastly, with respect to methodological approaches, the *large-n* topic model **dtm stm prep medium date** awards us an important insight into the potential granularity in that it identifies different types of topics talking about either budget or extreme weather events. We consider this a validation of the appropriateness of the stm model and its usefulness for future analyses.

5. Conclusions

In this paper, we looked at the effect of civil action on newspaper coverage of the issues that the civil action pertains to. We found that there is indeed a significant effect. Moreover, we investigated the effect of the recent climate-related protests on climate-related newspaper reporting. We found that the coverage was indeed gloomier and associated with more alarmist and negative language as opposed to the period preceding the action.

We also identified several shortcomings of our research, including low quality of the data and over-reliance on qualitative methods during several steps. Nevertheless, we believe out paper yields useful insights for policymakers with respect to understanding the relationship between civil action and newspaper reporting.

With respect to further research, added value could be gained from understanding how exactly (it at all) the relationship between civil action an media coverage impacts policy-making on a given issue and whether an alarmist strategy is also an effective one. However, this research endeavour will have to be postponed until more data is available on the voting patterns on the issues investigated in this paper.

6. Contributions

The teamwork element of this research project was found beneficial by all members of the group. The group took decisions collectively, resulting in an efficient and fair share of the overall workload. Samuel focused the calculations of the statistical analysis, focusing on the coding and interpretation of the Structural Topic Modelling (STM). Dario focused on the main body of the code and the processing of the data, and executed the the Principal Content Analysis (PCA). The latter included the activities listed in Section 5. Nikki was responsible for the initial collection of the documents. This included defining the search queries. He also attended to the sentimental analysis of the coverage. Overall, it has been a very agreeable collaboration.

7. References

References

[1] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber. *Multivariate analysemethoden*. Springer, 2016.

[2] L. Barasi. Guest post: Polls reveal surge in concern in uk about climate change, May 2019.

[3] M. Bouchet-Valat. Snowballc: Snowball stemmers based on the c 'libstemmer' utf-8 library, 2015.

[4] J. W. Boumans and D. Trilling. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital journalism*, 4(1):8–23, 2016.

[5] I. Feinerer. Introduction to the tm package. text mining in r., 2015.

[6] R. Gifford and L. A. Comeau. Message framing influences perceived climate change competence, engagement, and behavioral intentions. *Global Environmental Change*, 21(4):1301–1307, 2011.

[7] E. Greussing and H. G. Boomgaarden. Shifting the refugee narrative? an automated frame analysis of europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, 43(11):1749–1774, 2017.

[8] W. Revelle. Package 'psych'. procedures for psychological, psychometric, and personality research, 2015.

[9] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.