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

Dario Siegen
dario.siegen@hotmail.com

Nikki Wirawan
nikki.wirawan@gmail.com

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

Abstract

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.

We want to analyse whether and how these new 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 over 10.000 media articles from eight 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. Calculations such as the principal component analysis (PCA) and sentiment analysis allow to describe and analyse the content of the articles quantitatively. With our results, we want to contribute to the literature on social movements and climate change communication. We hope to draw some conclusions about the impact of the dominant media frames by combining them with the results of survey studies over the same period. This midterm report explains our data and method in more detail and summarises our progress in this project. The next steps are defined at the end of the document.

1. Proposed Method

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 look at this 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?

In this midterm report, we focus on the preparation of the data and the calculation for RQ2, as we consider these calculations to be the most pressing ones for our project. In the next section, we shall explain the method in more detail.

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

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

In their project, they [7] analysed 10.606 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, before mapping the frames back onto the articles and plotting the frequency of individual frames against time.

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.

Therefore, 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.

Table 1	
Name	*Type*
BBC	Broadsheet
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. These are the eight UK newspapers that we analyse.

2. Experiments

Methodologically, we oriented our process very closely to the Greussing and Boomgaarden paper [7]. 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 meanigful 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.

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 **10.584** 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 preprocessing, including the allocation of document variables (’docvars’), much simpler. Consequently, we exported the articles in a PDF format. This form of exporting data yielded us several tens 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.

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.

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.

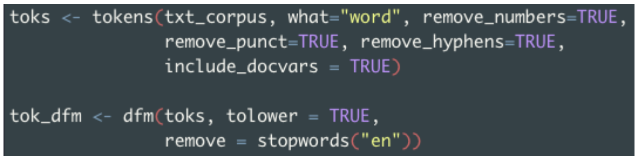
Harmonising text bodies: The last main issue that spread across our dataset was that some of the .txt documents’ text bodies were not following the general pattern of paragraphs but rather included odd symbols or no text body at all . Ultimately, we solved this problem by manually inspecting and treating 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 will consider simply dropping the worst cases of inconsistent documents from our analysis.

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.

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]. After a manual selection from the remaining terms, the Kaiser-Meyer-Olkin (KMO) criterion was applied and terms with a value below .5 were eliminated. 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 us to look at the data more closely to further improve its quality. Nevertheless, some tokens already catch the eye,



```
toks <- tokens(txt_corpus, what="word", remove_numbers=TRUE,
               remove_punct=TRUE, remove_hyphens=TRUE,
               include_docvars = TRUE)

tok_dfm <- dfm(toks, tolower = TRUE,
               remove = stopwords("en"))
```

Figure 1. Example for the code used to pre-process the tokens.

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 method: 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.

Apart from that, we also want to find qualitative ways to explain the changes in dominating frames over time. One possibility would be an additional framing analysis of the media releases of the protesters or a similarity metric of the same.

Experimental details: Finding all the applicable functions to transform the raw text into a data frame and eventually into token data frames proved to be time demanding due to the issues we encountered described earlier. 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.

Results: So far, we have two PCA tables with rough calculations of our tokens (stemmed and non-stemmed). The details can be found in the GitHub repository (Midterm-script” branch). So far, the results hint towards the fact that we have applied the right code. However, to really reach meaningful results, we have to further improve the quality of our data. As some tokens appear in multiple different forms (e.g. ”school” and ”schools”), stemming might be the right approach.

Table 2				
”can”	”get”	”fossil”	”fuel”	”compani”
”right”	”thing”	”lot”	”left”	”us”
”one”	”realli”	”like”	”talk”	”money”
”last”	”week”	”young”	”campaign”	”challeng”
”http”	”news”	”take”	”live”	”elect”
”trump”	”high”	”power”	”support”	”see”

Table 2. Top 30 stemmed tokens in frequency.

Table 3			
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 3. These are a few non-stemmed selected tokens related to climate movements.

Table 2 gives an overview of the 30 most frequent stemmed tokens across the 300 test documents. The token number 21, ”http”, is an example for the data pre-processing that we still need to improve. The tokens ”fossil” and ”fuel” hint towards the fact that some strong collocations are in the corpus. A collocation analysis will give us a clearer idea of the common collocations. Like the authors of our model study [7] chose to proceed in their analysis, we could analyse some tokens as collocations.

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.

3. Future work

Overall, we found our analysis to be helpful in identify frames. However, our ability to extrapolate meaningful results was hindered by the poor quality of data and the need for further refinement thereof. We intend to address these issues ahead of the execution of the final analysis.

Our main task for the coming weeks will therefore be to further improve the quality of the data and analyse it according to our model article (pre-processing, Kaiser-Meyer-Olkin calculation, PCA). In addition, we will perform a simple sentiment analysis [10]. A more concise interpretation of the results will be the most important task to come.

Furthermore, linking our findings to public polls in the UK over the same time period, we might be able to substantiate the hypothesis that the climate protests with their ”doom and gloom” narrative not only impacted media coverage, but also public opinion [2]. In doing so, it is particularly important that we find ways to identify the direction of causality.

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.

[10] L. Young and S. Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231, 2012.