IDS.131 Proposal:
The Evolution of the News Narrative on Climate Change

Adam Block Helena Caswell Dylan Lewis Disha Trivedi

March 29, 2021

1 Introduction

Print and televised media reporting on climate change influences the public perception of climate change [Antilla, 2008], which in turn affects support for systemic policies to reduce greenhouse gas emissions and for individual actions to mitigate climate change, such as purchases of hybrid vehicles [Chen and Zhao, 2019]. Even with the rise of online news and social media in the past decade, approximately 68% of Americans get their news often or sometimes from television [Center, 2020]. At the same time, US public opinion on climate change has shifted significantly in the past decade, rising from only 49% of the population believing that global warming would harm the U.S in 2008 to 62% believing the same in 2018 [YPC, 2020, Ballew, 2020].

Based on this decade-long rise in public opinion that climate change will harm the U.S., we seek to investigate the simultaneous changes in news coverage of climate change to ask the following research question: What environmental, social, and political factors influence the sentiment and frequency of top American English-speaking news media coverage of climate change? If identified, these factors serve an important role in shaping news coverage that correspondingly shapes all-important public opinion about climate change's risk and related policy and personal choice to offset it.

To conduct this investigation, we will analyze a 2009-2020 dataset of television climate change coverage from MSNBC, CNN, and Fox News and study environmental, social, and political factors that may drive changes in these television networks' sentiment or frequency of climate change coverage. To do so, we will analyze approximately 10 years of data from three television stations across the political spectrum: CNN, FoxNews, and MSNBC. Some of these environmental, social, and political factors we will investigate are:

- natural disaster type and frequency (wildfires, hurricanes, flooding)
- extreme weather and temperature variation (extreme heat or cold, extreme snow)
- political events and terms (presidential terms or campaigns)

Based on related work, we expect that such factors will be correlated with, and possible contributors to, changes in television media coverage of climate change.

2 Related Work

There is ample evidence that the environmental, social and political factors which we propose to study do influence news media coverage and opinions on climate change. Such evidence stems from recent computational sentiment analyses of social media discourse on climate change. In a review of related work on environmental factors, we found that public perception of climate risk has been shown to be strongly influenced by concurrent extreme weather events [Kim and Marlon, 2020]. We also found work from Ruz et al. 2020 that indicates that natural disasters are strongly correlated with negative sentiment associated with environmental discourse on social media. In the same work, Ruz et al. 2020 found that social movements, particularly those around climate and environment, were associated with heterogenous sentiment changes on Spanish-speaking Twitter media. Sentiment analysis and Bayesian classifier use on such data also enabled

network mapping between related patterns of environmental and natural disaster-associated words [Ruz, 2020].

Political affiliations and media leanings also influence political coverage and events. A study by the Feldman et al. on 2007-2008 survey data of climate change coverage on Fox News, CNN, and MSNBC found a negative association between Fox News viewership and acceptance of global warming and a positive relationship between CNN and MSNBC viewership and acceptance of global warming [Feldman and Leiserowitz, 2011]. The authors also found that Fox News takes a more dismissive tone towards climate change than CNN or MSNBC and interviews a greater ratio of climate change deniers than believers. This partisan news divide maps to beliefs about climate change, with 95% of liberal Democrats believing global warming is happening compared with 41% of conservative Republicans.

We seek to expand upon this work with a computational sentiment analysis and frequency analysis for a larger array of factors over a longer timespan. In some instances, such studies already exist for other media spheres or geographic areas. Dahal et al. 2019 conducted such computational analysis on U.S.-based users; Twitter data in order to gauge public opinion on climate change. Their sentiment analysis demonstrated a strong negative correlation between Twitter discourse on climate change in response to extreme weather or political events [Dahal and Li, 2019]. Similar work has been conducted on British and Spanish Twitter data, where sentiment analysis found a more negative association with climate change in Britain than in Spain and also determined positive sentiments attached to discussion of renewable energy on both countries [Loureiro and Alló, 2020].

Much like the above studies Jost et al. 2019 conducted a sentiment analysis, albeit non-computational, of Canadian print and televised media to determine the positive or negative associations attached to the 'change' portion of 'climate change' [Jost, 2019]. Their findings indicated that sentiments attached to 'change' varied with political factors that influence, block, or drive change.

Such findings authoritatively demonstrate the environmental, political, and social factors influence climate change discourse in text-based social media. These findings correspondingly dovetail with and motivate our investigation of environmental, political, and social factors' influence on American English-speaking television media. Some of the methods used in the above papers (Bayesian classifiers, topic modeling, etc.) may provide rich avenues for exploration beyond the methods we later propose in Section 5 (Methods).

3 Data

For this analysis we have decided to use a dataset with television transcript snippets related to climate change coverage across CNN, MSNBC, and FoxNews from July 2009-January 2020, from the GDELT Project[1]. This dataset includes the time of day and date of the mention, the station, the show, and a snippet of the transcribed audio. Although the dataset includes similar data from shows associated with the BBC, due to the relative paucity of the BBC data (the BBC snippets extends only from 2017-January 2020), as well as our primary focus on the American TV media landscape, we exclude these data from consideration. In the resulting dataset, we have 19,304 snippets from CNN, 25,865 snippets from FoxNews, and 26,429 snippets from MSNBC for a total of 71,598 snippets of climate change coverage.

3.1 Exploratory Analysis

3.1.1 Text Preprocessing Pipeline for the Climate TV News Corpus

Before we can either explore or analyze our data, we need to clean it. The first point to note is that there is missing data from CNN from the month of October 2009; after removing this month from consideration, we are left with a somewhat standard Natural Language Processing (NLP) pipeline. While our data have ancillary features including URLs to sources, the name of the show, and some identifying information, we are primarily concerned with the date of airing and the text snippet. We follow standard NLP practice by removing punctuation and numbers, converting all letters to lower-case, and tokenizing the data. Finally, we lemmatize each word, i.e., convert each word to its root, in order to better match similar semantics (for example, we might hope that "go," "going," and "went" are all treated as the same word). As we are working

[1] Link to Climate TV News Dataset

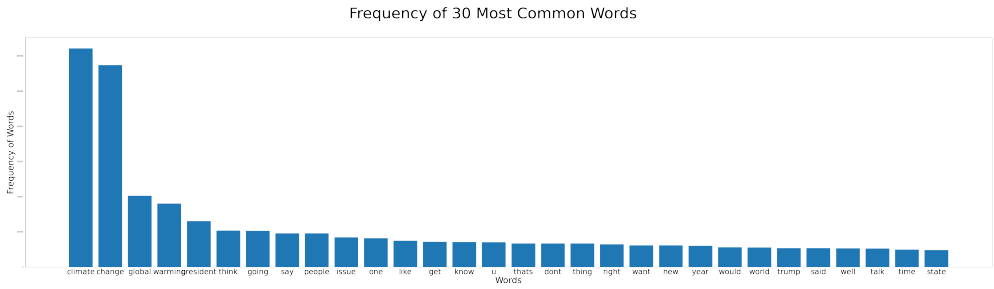

Figure 1: Frequency of the 30 most common words in snippets coming from all sources.

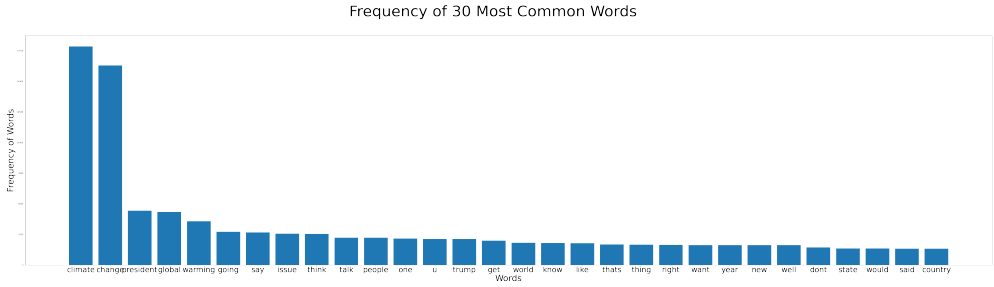

Figure 2: Frequency of the 30 most common words in snippets coming from CNN.

with relatively short snippets, we do consider the structure of the sentence to be of lesser importance and thus tokenize by single-word chunks as opposed to more complicated n-grams. Finally, we use the NLTK package in Python [Bird et al., 2009] to remove standard stopwords in order to better distill the signal in each snippet. In our analysis, we treat each snippet, or sometimes a group of related snippets, as what is typically referred to as a document in NLP literature. We use the terms interchangeably, but we do make clear the distinction when using a single snippet vs. a group of snippets. Beyond devising the pre-processing pipeline, we consider other exploratory steps.

3.1.2 Identifying and Removing Stopwords

As part of the exploratory phase, from Figures 1-4 above, we identify words common to many of the snippets that we deem too ubiquitous to contribute to any signal in the differentiation among documents. These words are referred to as stopwords. Using a max document frequency threshold of 20% (i.e. terms which appear in greater than 20% of the snippets in the entire dataset), we find that the corpus-specific stopwords are **{climate, change, global, warming}**, so we additionally remove these words from all snippets in the corpus. We will remove any other words that do not contribute to understanding the semantics of what is being said or to the differentiation of the signal that we come across in further analysis to further eliminate noise.

3.1.3 Comparing Climate-Change Word Frequencies between 2019 + 2020 and 2009

Since snippets of our dataset range from July 2009 - January 2020, we analyze the change of term frequencies between snippets in 2019 + 2020 compared to snippets from 2009. Since 2009 and 2019 + 2020 represent the largest time difference of our data, we use the snippets of these years for a coarse term frequency change analysis between the ends of our dataset. We are interested in identifying what terms used in climate coverage were said more in 2019 + 2020 than in 2009. We choose to combine all snippets from 2019 and 2020 for this comparison since there is only one month of snippets available for 2020. To do this analysis, we normalize the word frequency of each word by dividing the occurence of a word by the total word count for each year respectively, yielding two normalized relative word frequencies, one for 2019 + 2020 and one

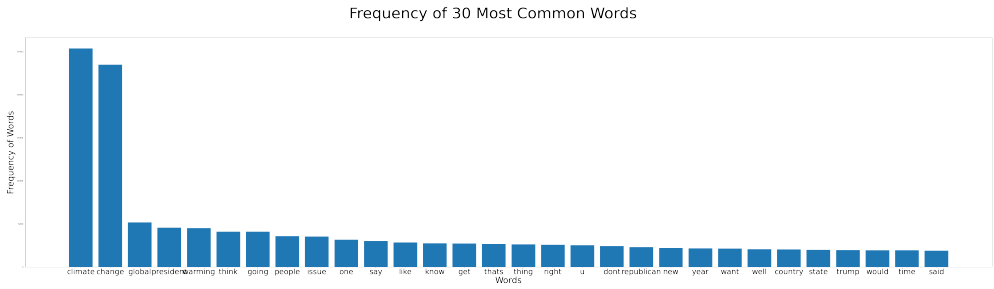

Figure 3: Frequency of the 30 most common words in snippets coming from MSNBC.

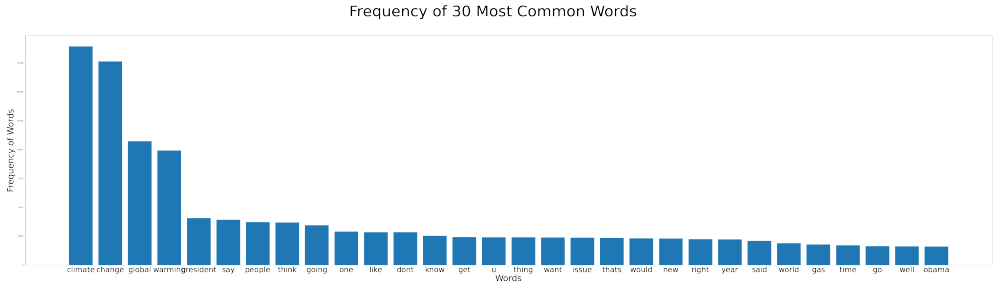

Figure 4: Frequency of the 30 most common words in snippets coming from FOX News.

for 2009. We take the difference between these two numbers to enable us to see specifically which words were said more in 2019 + 2020 than in 2009. This normalized relative word frequency difference measure effectively identifies vocabulary that was said more in 2019 + 2020 than in 2009. From Figure 5, we notice that the US president during 2019 + 2020, **Donald Trump**, is mentioned the most (appears as **trump** in Figure 5). Along with Donald Trump, other words and phrases related to politics are mentioned more so in 2019 + 2020 than in 2009 including **candidate, democrat (democratic), voter, biden**, and the **Green New Deal** (appears as **green, new, deal** in Figure 5). These words likely show up more in the TV News of 2019 + 2020 because Joe Biden was a leading democratic candidate for the 2020 presidential election and democratic policies in regards to climate change are typically environmentally-friendly (including the Green New Deal). Lastly, we observe words indicative of widespread crisis being used to describe climate change more so in 2019 + 2020 than in 2009 including **crisis**, **issue**, **everywhere, people, emergency, threat, impacted, need**, and **national**.

4 Methods

To address our research question on what environmental, social, and political factors drive changes in climate coverage, we present our preliminary hypotheses and methods of analysis.

4.1 Identifying Coverage Topics and Modeling Media Trends

4.1.1 Uncovering Topics using LDA

We are interested in specific topics that can found in the data. A common method for topic analysis is Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. We will apply LDA to our textual data to find specific words associated to particular common topics of discussion.

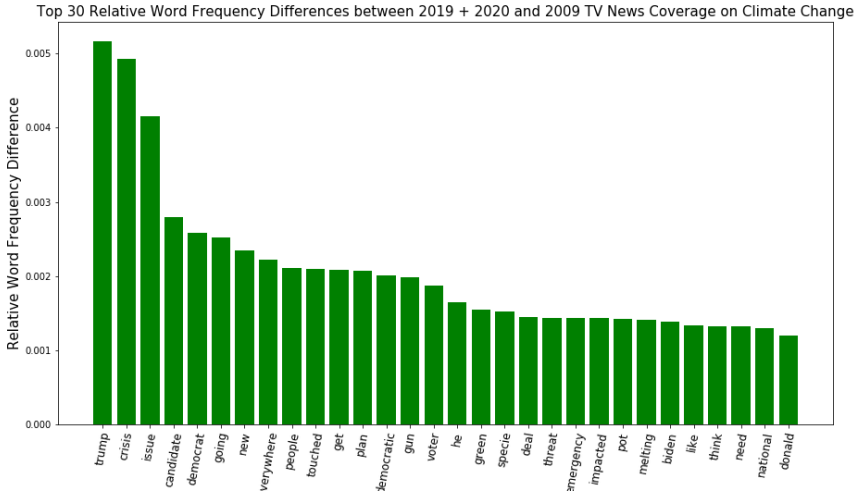

Figure 5: Relative Word Frequency Differences between 2019 + 2020 and 2009 for TV News Coverage on Climate Change

4.1.2 Modeling Media Trends

Having found such words from topics using LDA, we will study how the frequency of these terms changes over time. We hypothesize that some topics found will roughly correspond to natural disasters like fires, hurricanes, or earthquakes; it will be relevant to determine how media mentions of said disasters vary with time. We expect some degree of seasonality, as the media mentions likely increase when a disaster occurs. At the same time, we suspect that certain disasters, such as fires, are mentioned much more recently, due to their increasing prevalence in California and Australia.

We will fit a linear model for the trends associated to several different key words associated to each topic; we will then adjust for seasonal or monthly variation to determine whether the prevalence of these words is indeed increasing. We are also interested in the ways that different sources describe climate change. Thus, we will compare the frequency of different key words across the three different channels by testing whether the frequency differences are significant. We will use data from NOAA's billion dollar weather and climate disasters to identify the largest US natural disaster type (wildfire, storm, etc.) and timing [NOAA, 2020].

4.2 Change-Point Detection

We hypothesize that presidential administration has an effect on climate coverage. As such, we will test for a difference in the distribution of frequencies between the 2009-2016 time range and the 2017-present time range. We will do this both as an aggregate analysis as well as on a channel-by-channel basis (with appropriate correction of family-wise error rate due to the fact that we are testing several hypotheses). Time permitting, we will also use methods from the change-point detection literature to find and test for significance other times where the frequencies have changed. This is a somewhat subtle problem, as our hypotheses of certain change-points depend on the data themselves; this issue has been addressed in works such as [Jewell et al., 2019, Tibshirani et al., 2016].

4.3 Term Importance by Year and Network

We use term frequency-inverse document frequency (TF-IDF), which is a statistical measure of how important a word is to a snippet, or a subset of snippets in the corpus. We use TF-IDF to identify the words most important to each year from the entire corpus vocabulary. Additionally, we aim to find subset of words important to each network. We aim to gain a broad overview of what was being discussed over time and by who.

4.4 Sentiment Analysis

To understand the role sentimental language has in the coverage of climate change, we investigate how words of non-neutral sentiment (positive and negative sentiment) have changed over the years and if the use of sentimental language differs between TV networks. We will first identify the polarity words by tagging the words in our corpus with an associated sentiment score using the NLTK sentiment analysis framework.

4.4.1 Identifying Sentimental Words by Year and TV Network

After separating words by sentiment (positive and negative), we use TF-IDF similarly to to yield the set of sentimental words important for each year respectively. We plan to generate visualizations of these words over the years to yield an overview of what sentimental words were being used in coverage of climate change in different years.

4.4.2 Change in the Use of Sentimental Language over Time

We use hypothesis testing to test if the overall use of sentimental language has changed over the years in the coverage of climate change, to gain insight into how climate coverage is being discussed over time rather than what is being discussed.

4.4.3 Use of Sentimental Language Between Networks

We investigate if there is a statistically significant difference in the use of sentimental language in the coverage of climate change between the TV networks by looking at normalized amounts of sentimental words used across the snippets for each of the networks. We will use the ANOVA and Tukey's Range hypothesis tests to assess the hypothesis that there is a difference between the use of sentimental language between the networks. We select these tests as we will be conducting multiple hypothesis tests when comparing the networks pair-wise and these tests allow us to limit the probability of Type I Error.

5 Timeline

We plan to meet weekly to update on progress and plan next steps. Below is a timeline of our analysis and write up:

Table 1: Timeline.

March 29	Project Proposal Due
March 30-April 9	Finalize Analysis Topics and Methods based on Continued Exploratory Analysis
April 12-19	Begin Sentiment and Time Series Analysis
April 19	Intermediate Status Report Due
April 19-May 3	Continue Sentiment and Time Series Analysis
May 3-10	Finish Analyzing Data and Create Final Data Visualizations
May 10-17	Create Presentation and Write Report
May 17	Project Presentation
May 20	Project Report

References

Climate Change in the American Mind: National Survey Data on Public Opinion (2008-2018) . *Yale Program on Climate Change Communication (YPCCC) George Mason University Center for Climate Change Communication (Mason 4C).*, 2020. doi: 10.17605/OSF.IO/JW79P.

Liisa Antilla. Self-censorship and science: a geographical review of media coverage of climate tipping points. *Public Understanding of Science*, 19(2):240–256, 2008. doi: 10.1177/0963662508094099.

Leiserowitz A. Roser-Renouf-C. Rosenthal S. A. Kotcher J. E. Marlon J. R. Lyon E. Goldberg M. H. Maibach E. W. Ballew, M. T. Climate change in the american mind: Data, tools, and trends. *Environment: Science and Policy for Sustainable Development*, 61(3):4–18, 2020. doi: 10.1080/00139157.2019.1589300.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Pew Research Center. 86% of Americans Get News Online from Smartphone, Computer or Tablet, 2020. URL https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/.

Mrinal Ghosh Yong Liu Chen, Yubo and Liang Zhao. Media coverage of climate change and sustainable product consumption: Evidence from the hybrid vehicle market. *Journal of Marketing Research*, 56(6): 995–1011, 2019. doi: 10.1177/0022243719865898.

Kumar Sathish Dahal, Biraj and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):1–20, 2019. doi: https://doi.org/10.1007/s13278-019-0568-8.

Edward W. Maibach Connie Roser-Renouf Feldman, Lauren and Anthony Leiserowitz. Climate on cable: The nature and impact of global warming coverage on fox news, cnn, and msnbc. *The International Journal of Press/Politics*, 17(1):3–31, 2011. doi: 10.1177/%2F1940161211425410.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a Change in Mean after Changepoint Detection. *arXiv preprint arXiv*, 2019. doi: 1910.04291.

Ann Schwebel Shoshana Jost, François Dale. fix. *Environmental Science and Policy*, 96(1):27–36, 2019. doi: https://doi.org/10.1016/j.envsci.2019.02.007.

Ballew M. Lacroix K. Leiserowitz-A. Kim, L. and J. Marlon. How does the american public perceive climate disasters? 2020.

Maria L. Loureiro and Maria Alló. Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the u.k. and spain. *Energy Policy*, 143(1):111–90, 2020. doi: https://doi.org/10.1016/j.enpol.2020.111490.

NOAA. Billion Dollar Weather and Climate Disasters, 2020. URL https://www.ncdc.noaa.gov/billions/time-series.

Henríqueza Pablo A. Mascareño-Aldo Ruz, Gonzalo A. Sentiment analysis of twitter data during critical events through bayesian networks classifiers. *Future Generation Computer Systems*, Volume 106(1):92–104, 2020. doi: https://doi.org/10.1016/j.enpol.2020.111490.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.