Title
Climate Change in American Media

Author
Broad, Joseph

Publication Date
2024

Climate Change in American Media

By

JOSEPH BROAD
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Political Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Christopher Hare, Chair

Amber Boydstun

Benjamin Highton

Committee in Charge

2024

To my wife, Kathryn, for her endless support, patience, and sacrifice that made this dream possible.

ACKNOWLEDGEMENTS

I would like to thank the following people who were integral to the completion of this dissertation. Dr. Christopher Hare, whose guidance and support inspired me to push my methodological skills beyond what I thought I was capable. Thanks to Dr. Amber Boydstun and Dr. Benjamin Highton for their constructive feedback, encouragement, and faith in me from the beginning. I am grateful to all of my peers at UC Davis, who made the political science program a place of warmth and belonging.

My sincerest apologies to my family for years of neglect, missed family functions, postponed holidays and other "rain checks." To my parents, Leonard and Teri, I cannot thank you enough for the support and encouragement you have given me. You helped me overcome so many hurdles along the way.

I would especially like to thank my wife, Kathryn. I could not have completed this dissertation without her love and support. Thank you, Kathryn, for years of encouragement. You inspired me to never give up on this dream, from on our first date 15 years ago and now still. From the bottom of my heart, thank you.

ABSTRACT

In spite of climate change being considered an existential threat to the human species, this issue receives relatively little attention in U.S. news media. In this manuscript I provide an overview of attention to the issue of climate change in media communications and address shortcomings in methods for measuring media attention to this issue. Results indicate that attention is lower than previously reported and increasing at a slower rate than previously reported.

This manuscript shows that the standard text classification method in climate change communications literature overestimates news media attention to climate change by a factor of two to three. A Support Vector Machine (eSVM) model enriched with features from an experimental Latent Dirichlet Allocation (LDA) topic model was trained on pre-labeled data ($N \approx 50,000$). This model produced substantially higher climate change story classification accuracy (F1 scores) compared to the industry standard (Boolean classification) and showed better performance than other text classification alternatives.

Applying an the eSVM text classification model on a novel database of 1.1 million news stories distributed on the front page of the New York Times (1996-2023) and via Twitter by a diverse set of news content creators (2007-2023), this manuscript provides a comprehensive analysis of climate change attention across different domains and platforms. Results from machine learning classification (eSVM) indicate that news media attention to climate change is increasing over time but at a far slower pace than previous literature suggests.

In this manuscript I show that the inflation of climate change attention in past literature is due to the diffusion or permeation of climate change as a relevant consideration in other policy topics. Using an experimental LDA topic model, I analyze the network of associations of climate change with other topic considerations, including energy, agriculture, health, the economy, and others. The experimental "guided" LDA was trained using prior information about the structure of policy topics in news media: the model was fit to produce a topic structure that closely mirrors the Comparative Agendas Project's topic coding schema at the four-digit level by selecting highly informative keywords from pre-labeled data (tf-idf) as central to each topic. Results indicate that the issue of climate change is becoming a more relevant consideration in a greater number of policy topics over time; however, this finding holds only for "prestige" and "niche" news sources and not for "new" media sources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Abstract

In this manuscript I use experimental methods and a novel data set to compare various text classification models, evaluate attention to the issue of climate change and pollution between 1996 and 2023, and perform automated content analysis of climate change articles to examine how climate change is associated with other major topics in U.S. news media.

1.2 Introduction

In spite of climate change being considered an existential threat to the human species (Huggel *et al.*, 2022), this issue receives relatively little attention in U.S. news media. News media attention can influence the U.S. public's perceptions or perceived importance of the issue of climate change and can translate to local, state and federal policy agendas. The amount of attention an issue receives is not always proportional to the seriousness of an issue; this seriousness-attention gap can result in important issues such as climate change being overlooked by the public and policymakers in favor of hot topics or other pressing issues.

Climate change effects are expected to be widespread, impacting extreme weather events, immigration, agriculture, wildlife and more. Solutions to prevent, mitigate or reverse the effects of climate change are equally diverse: ranging from technical innovation (e.g., vehicle emissions standards, alternative and renewable energy sources, carbon capture technologies), reduction in consumption of energy and goods, and even restructuring local zoning laws.

The impacts of and potential solutions to climate change are highly diverse and yet the associations between climate change and other important political issues in U.S. media has not been studied in climate change communication literature.

Climate change is an obscure topic and the public's awareness of the breadth and severity of its impacts requires an information environment that reflects the reality of the issue. In this manuscript I explore U.S. news media attention to climate change and how (or whether) news media associates climate change with other important political topics.

1.3 Background and Theory

Climate change was rarely covered on the New York Times front page during the 1990s and 2000s. Stories on this issue made up less than half of a percent of all New York Times stories during this period. Front page of climate change was also rare in scope: when the New York Times covered climate change, it focused primarily on policy shifts relating to vehicle emissions standards and transitioning away from consumer goods that required high energy or produced byproducts harmful to the environment.

Uncertainty was often a central theme in early front page coverage of climate change: the unknown costs, risks, and even potential benefits of climate change frequently underpinned discussion of this topic and undermined meaningful policy changes to curtail its causes and effects. For example, in 1997 the New York Times front page included a story that solar storms may be causing global warming (Broad, 1997) and that the uncertainty surrounding climate change impeded meaningful policy action on this issue (Stevens, 1997). In the early 2000s, the New York Times front page even included stories discussing the potential benefits of climate change Krauss *et al.* (2005), despite clear signals from the international political and scientific communities that climate change posed substantial risks to human ecosystems if not addressed (the international response to the 1992 United Nations Framework Convention on Climate Change, for example).

It wasn't until 2010 that the New York Times front page included a report directly linking climate change with extreme weather and other visible impacts (Gillis, 2010). By the mid-2010s the New York Times began to frequently report on the visible impacts (Gillis, 2016), economic costs (Davenport, 2014), and certainty of anthropogenic climate change (Gillis, 2014). Additionally, the New York Times Front Page began to expose individuals and organizations for their deliberate efforts to undermine the public's confidence in climate change science, including the originator of the dubious solar storm hypothesis that the New York Times had promulgated 18 years earlier (Gillis & Schwartz, 2015).

What might explain the shift in climate change news coverage from scarce and uncertain to infrequent and certain? Construal Theory suggests that the temporal and spatial proximity of the costs or risks of an issue are proportional to its perceived importance. As the costs of climate change become clearly visible and immediate, we might expect the news media to give greater attention to this issue. The relationship between perceived proximity of costs and perceived importance has been corroborated in climate change communication literature, as well as other dimensions such the expected severity of climate change costs (Rickard *et al.*, 2016; Spence *et al.*, 2012).

By extension of Construal Theory, issue proximity may also be associated with increased commitment to pro-environmental policies to curb the effects of climate change. Communicating the costs of climate change that directly impact an issue public — a group that prioritizes one particular policy area — may also increase commitment to pro-environmental policies and practices. For example, coverage of climate change impacts on foreign countries or even U.S. coastal cities may not motivate farmers in Iowa to shift toward more sustainable and eco-friendly agricultural practices; however, this issue public may express greater commitment to policies to mitigate the effects of climate change when they are expressed in terms of threats to crop yields and business sustainability (Haden *et al.*, 2012).

Communications associating the effects and causes of climate change with issue publics' issue priorities have the potential to mitigate the effects of partisanship (Haden *et al.*, 2012). Climate change has been a highly polarized (and polarizing issue since) for decades (Bayes & Druckman, 2021; Egan & Mullin, 2017; McCright & Dunlap, 2011); but the potential for issue associations to supersede partisan polarization may be critical to making meaningful progress on this issue by building consensus across the ideological spectrum. Yet the association of climate change with other issues has not been explored in climate change communication literature. The potential for issue associations to increase support for climate change mitigation and prevention policies and for these associations to cut across party lines makes issue association content analysis of news media coverage a crucial addition to climate change communications literature.

In addition to the temporal and spatial proximity of climate change, past content analyses of climate change news stories have focused heavily on debates on the evidence for climate change and the concerted efforts to undermine climate science (Boussalis & Coan, 2013; Brulle, 2019). Other literature has focused on general themes of optimism and pessimism in climate change communication (Johns & Jacquet, 2018). But given the polarized nature of the issue of climate change, additional news media attention to the evidence of climate change (or denial thereof) is unlikely to motivate the public (or issue publics) to change their policy preferences or actions (Egan & Mullin, 2017).

Climate change issue associations can also be construed as containing an element of personal efficacy: associating climate change with other issues may activate issue publics on issues for which personal efficacy was already high. Scholars have examined themes of efficacy in addressing the climate change threat, suggesting that portraying the issue as one that can be ameliorated through individual efforts may increase commitment to climate solutions (Feldman *et al.*, 2017; Haden *et al.*, 2012).

Whereas federal policy faces continual gridlock, individuals may have greater efficacy to impact change at the local level and the policies to mitigate the effects of climate change are dispersed over several policy areas. Thus, performing content analysis on climate change issue associations in U.S. news media provides an opportunity to examine how the discussion of climate change has dispersed across topics and how climate change solutions may be dispersed across policy areas.

In this manuscript I evaluate U.S. news media attention to the issue of climate change between 1996 and 2023 and analyze shifting trends in how the issue of climate change is associated with other important political issues such as the economy, health, agriculture, and others. Using experimental methods in machine learning and a novel data set comprised of tweeted articles collected from Twitter and various news sources, I compare various text classification models, evaluate attention to the issue of climate change between 1996 and 2023, and perform automated content analysis of climate change articles to examine how climate change is associated with other major topics in U.S. news media.

1.4 Structure

The second chapter demonstrates improvements over the climate change communications literature's standard corpus selection and topic classification methods (i.e., how news articles are chosen for evaluation and how they are classified as climate change articles). In climate change and environmental communication literature, keyword "hits" (or "Boolean classification") is the method most frequently used for corpus selection, topic classification, and measuring media attention.

A "hit" refers to the appearance of a specified keyword or phrase in a news story; news stories are considered to be about a topic such as climate change when the text contains one or more "hits" in its headline or content. Some scholars use Boolean searches on articles' keywords through services such as LexisNexis to select a corpus of documents about that topic for evaluation (Feber *et al.*, 2017). Some use Boolean searches on story headlines or content to classify documents as belonging to the topic of interest (Holt & Barkemeyer, 2012; Liu *et al.*, 2011).

While classifying articles as belonging to a topic using hits is the industry standard in climate change communication literature, this approach may inflate the number of both false positives and false negatives (King *et al.*, 2017). As King *et al.* (2017) show, the ability of scholars to build a comprehensive list of keywords and phrases related to a topic of interest is limited and should be supplemented with machine learning techniques; this limitation can result in articles of interest being classified as some other topic. For example, selecting a set of articles with "climate change" in the article's keywords may not identify earlier articles where "global warming" was the preferred term used to discuss this issue.

Using keyword hits in headlines and content to classify climate change articles increases the prevalence of false positives as climate-change-related keywords can be mentioned in articles focused primarily on other topics. For example, climate change keywords may appear briefly in news articles covering political debates, weather reports, natural disasters, fires, news roundups, discussions on the era of misinformation, or even sources for increased public anxiety and hopelessness.

Appropriately identifying articles' primary topic with minimal error is critical to our understanding of the relationship between media attention and both public opinion and political agendas. Media attention's associations with shifts public opinion (Althaus & Tewksbury, 2002; Iyengar, 1987; Iyengar & Simon, 1993) and political agendas (Jones & Baumgartner, 2005a; Van Aelst & Walgrave, 2016) are well-documented. However, methods that produce substantial classification errors may confound our understanding about the level of attention climate change receives in news media, its effects on public opinion and policy, and the causal effects driving these observations.

The second chapter provides an overview of more sophisticated supervised and unsupervised machine learning approaches to classify articles as belonging to the climate change subtopic.[1] This chapter will also compare the Boolean "hits" classification approach to other text classification methods such as Naive Bayes (NB) and machine learning methods including Support Vector Machines (SVMs), enriched SVMs and Neural Networks.

These and similar techniques have been used in the social sciences to classify the topics of online communications such as message boards (Inkpen & Razavi, 2014), news stories

[1]The subtopic "climate change" as discussed in this manuscript is defined as the Comparative Agendas Project subtopic 705: Air Pollution, Global Warming and Noise Pollution. While this subtopic includes air and noise pollution stories, these stories make up an insignificant fraction of the training data used to build a text classification model to predict climate change stories.

(Silla & Freitas, 2011), tweets about the Supreme Court (Sandhu *et al.*, 2019), tweets to and by members of congress to evaluate "follow the leader" effects (Barberá *et al.*, 2019), pieces of legislation to evaluate government agendas (Hillard *et al.*, 2007), and even sections in elections manifestos as one measure of how political parties prioritize certain policy topics (Verberne *et al.*, 2014).

Beyond the use of terms as predictors, researchers also use numeric or probabilistic representations of words, terms and phrases to classify stories into topics. Because the terms in a corpus (a "dictionary") may number in the hundreds of thousands or millions, given the size of the corpus, text classification models may have a vast number of predictors. SVMs and NNs typically use terms as topic predictors, which result in models with hundreds of thousands or millions of parameters. Feature reduction techniques are applied in text classification to simplify models and reduce computational demand by replacing the dictionary with perhaps only hundreds of reduced (or "enriched") features.(Chen & Li, 2016; Dogru & al., 2021).

Enriched features are simply the output of feature reduction techniques meant to reduce model dimensionality by assigning topic probabilities or dimensional representations to terms. Imagine all terms in a dictionary lined along a climate change dimension, where high values indicate the term is highly associated with climate change and low values indicate low association with this topic. Through feature reduction processes terms such as "global warming" and "greenhouse gas" are assigned high values on our imagined climate change dimension while terms like "finance" or "health" may have a relatively lower climate change dimension score. While feature reduction techniques were designed to produce replacement features to simplify models, machine learning scholars opted to add the features as predictors for an increase in classification accuracy (Chen & Li, 2016; Dogru & al., 2021; Inkpen &

Razavi, 2014; Nassif & Fahkr, 2019; Yan & Zheng, 2020).

There is no one "true" model that best classifies text data for every use-case. In the second chapter, I compare the performance of these methods using pre-labeled data and discuss the intuition for each approach. The prelabeled data include the New York Times front page 1996-2006 data set (Boydstun, 2014) as well as about 20,000 randomly selected articles tweeted during the 2007-2023 period coded under the same Comparative Agendas Project topics schema (described in further detail below).

The third chapter addresses three additional shortcomings in climate change communication literature. As discussed above, several scholars measure attention in terms of keyword "hits": climate change attention is measured as the number of news articles mentioning keywords like "climate change," divided by the total number of news stories published during the study period (Liu *et al.*, 2011). In contrast, climate change attention is measured here as the percentage of news articles predicted to be primarily about climate change using the highest performing machine learning model from Chapter 2, a guided LDA-enriched Support Vector Machine model. I perform a bootstrap analysis to analyze the estimated news media attention to climate change between 1996 and 2023 while accounting for potential error in the climate change classification model.

Second, there is a prevalence for researchers to select stories from leading print news media as a representative sample for all news media (Johns & Jacquet, 2018). Leading print news media have been shown to have similar topic coverage to other media sources, in the aggregate (Vargo & Guo, 2017). While there may be general similarities in the coverage of topics in the aggregate, it's important to note that more niche or 'fringe' issues, such as climate change, which seldom garner front-page attention, often exhibit significant differences in how they are reported across various sources. This project improves upon existing research by diversifying the set of examined sources. These sources represent a more diverse set of news media content creators[2] in terms of intended audience, which is appropriate for an analysis of a "fringe" issue like climate change.

The novel data analyzed in this chapter consists of about 1 million stories tweeted by various news media content creators (New York Times, Wall Street Journal, NPR, Atlanta Black Star, The Root, Daily Wire, Christianity Today and Inside Climate News) from 2007 through 2023[3], and New York Times front page stories from 1996 through 2023. While attention to climate change is increasing overall, attention was lower for perceived right-leaning sources; attention to climate change was unexpectedly low among some left-leaning sources (@TheRoot, @ATLBlackStar, @LaOpinionLA).

Finally, this research improves upon past research by analyzing stories that were deemed sufficiently important by each source to warrant distribution via Twitter. Researchers tend to select corpora (or bodies of documents) that include any article published by a source, or any published article containing selected keywords (Liu *et al.*, 2011). This approach ignores the importance of the front-page news generating process in which issues compete for finite agenda space (McCombs & Zhu, 1995). Faced with spatial constraints, news media sources provide a signal for issue priorities in selecting stories that are worthy of the front page. Using tweeted stories provides a weaker issue priority signal compared to front page coverage (given that spatial constraints are relaxed) but also allows for the evaluation of "fringe" issues

[2]The "news media content creators" term was selected as an encompassing umbrella term to include both news sources, like the New York Times, and content creators that comment on the news, like the Daily Wire.

[3]The October 2022 purchase of Twitter by Elon Musk and subsequent attempts to discredit certain news organizations caused some disruptions in the usage of Twitter to disseminate news. For example, NPR halted its use of Twitter in April 2023.

that infrequently appear on the front page, like climate change (and allows for comparison across several sources, many of which do not have a print version). However, the novel data examined here include both front page- and twitter-coverage, allowing a direct comparison of these media distribution types for the New York Times.

The fourth chapter evaluates the association of climate change with other issues over time and across sources. With the novel data described above and an experimental approach in topic modeling, the fourth chapter provides a unique and novel contribution to climate change communications literature by using a guided (or keyword-assisted) Latent Dirichlet Allocation with a topic structure closely mirroring the Comparative Agendas Project's topic coding schema at the four-digit level (major topic as well as minor topic). This chapter includes automated content analyses in two directions: the first direction evaluates how other issues are associated with climate change in articles primarily about climate change (measured as the distribution of topics in climate change articles as estimated by the keyword-assisted topic model discussed below) while the second direction evaluates the topics in which climate change is discussed as a secondary issue.

The fourth chapter improves upon existing literature with an experimental guided Latent Dirichlet Allocation (LDA) topic model. The key feature of the guided LDA is a researcher-constructed term-topic dictionary that assists the LDA in finding the term-topic probabilities that best fit the data and impose a flexible topic structure on the data (Eshima *et al.*, 2020; Jagarlamudi *et al.*, 2012; Jones *et al.*, 2021; Meng & al., 2019). By using a term-topic dictionary, automated content analyses are highly replicable: topic labels are defined prior to running the model and the data are fit to selected terms (see Appendix B). Consistent with past research in attention and attention diversity, findings in this chapter indicate that issue associations expand with greater attention to this topic (Boydstun, 2008). It is unclear whether the expansion of topic associations is due to the costs of climate change becoming clear or due to coverage of action on climate change mitigation across dispersed policy areas but the findings suggest that the issue of climate is permeating different policy areas and reaching people in new ways.

The primary goal of this manuscript is to explore whether discussions of climate change in U.S. news media are dispersed over policy topics in U.S. news media and, if so, whether there are meaningful changes in climate change issue associations over time and differences across sources. In the second chapter, I compare the performance of machine learning classification techniques to identify articles that are primarily about climate change and other topics using pre-labeled news articles coded according to the CAP schema. The third chapter uses the best-performing machine learning model to predict the primary topic for about 1.1 million articles appearing on the New York Times Front Page (2007-2023) or tweeted by the set of news media content creators (2007-2023) in order to evaluate news media attention to climate change over time and across sources. The fourth chapter uses a keyword-assisted LDA topic model to estimate documents' major topic mixtures according to the CAP coding schema, analyzes the shifting diversity of climate change associations between 1996 and 2023, evaluates changes in specific issues associated with cliamte change over time, and evaluates differences in issue associations between sources.

CHAPTER 2

MODELING CLIMATE CHANGE ATTENTION

2.1 Introduction

The earliest example of automated content analysis and text classification was developed to combine a vast and complex set of content-specific rules on computer systems using punched paper cards. In spite of technological barriers, MIT's General Inquirer System was able to evaluate emerging themes in folklore across cultures and discriminate between real and simulated suicide notes, just to name two examples (Stone & Hunt, 1963).

In Stone & Hunt (1963), the author describes how a Psycho-Sociological Dictionary consisting of a series of rules associating parts of speech with simpler categorical elements outperformed human judgement in distinguishing between real and simulated suicide notes. The dictionary included tags to simplify terms into psychological themes such as weakness, strength, authority, danger, and death; the dictionary also included emotive themes (e.g., affection, anger, love, distress) as well as persons, objects, environments, evaluations and several other categories. The term-category dictionary was complex but its application allowed for simple and powerful analyses that not only classified texts more accurately than human judgement but led to discovery of the patterns and combinations of categories that made accurate, automated classification possible.

While complex and not without its flaws, the dictionary and paper punch card approach in Stone & Hunt (1963) was a testament to the early potential of automation in content analysis and text classification. There have been many advances in the fields of automated data wrangling, information retrieval, text classification and content analysis in the 60 years since MIT's General Inquirer was developed. However, modern techniques are rarely used in climate change communication literature.

Scholarly research on how much (and how) news media discusses climate change has historically been much simpler than the General Inquirer system. This chapter provides an overview of the methods currently used in climate change communications literature as well as their shortcomings. I propose several open source (at the time of writing this manuscript) alternatives to help researchers strengthen and automate their research pipelines, including methods in data wrangling (webcrawling, API requests), text classification (Naive Bayes, Support Vector Machines, Neural Networks) and information retrieval (word embeddings, Latent Dirichlet Allocation and its variants).

This chapter also compares the performance of some of these models in predicting the major topic (and minor topic 705, "Air Pollution, Global Warming, and Noise Pollution") assigned to stories in the New York Times Front Page data set (Boydstun, 2014) and a novel data set of stories tweeted from various online news content creators. Professor Boydstun's data set contains front-page stories from the New York Times between 1996 and 2006. Articles were assigned major and minor topics based on a modified Comparative Agendas Project topic coding schema (Jones & Baumgartner, 2005b). These data were supplemented with about 20,000 randomly selected front page New York Times stories and articles tweeted by various news media content creators between June 2007 and April 2023. These stories were also assigned a major topic using the Comparative Agendas Project topics schema, as well as an indicator for whether the story was primarily about the issue of climate change.

In this chapter I show that Boolean classification — the standard classification method in climate change communications research — has systematically overestimated news media attention to the issue of climate change. Machine learning text classification and feature reduction methods available since the 1990s and early 2000s substantially outperform Boolean classification (and perform on par with recent advances in Neural Network text classification). Additionally, data selection methods typically used in climate change communications research may introduce unnecessary limitations and can be avoided using open-source data wrangling tools. Given the severity of the issue of climate change, an unbiased measure of climate change attention is critical to our understanding of the relationships between media attention, public opinion and policy action.

The information provided here is not meant to be a comprehensive guide to performing automated data wrangling and machine learning research. Instead, I hope that the following might spark researchers' interest to consider adapt some of these techniques in their own work. To that end, I have provided R and Python code to replicate the research presented here and in later chapters. (See Appendix A.) As a final note: the techniques here are widely applicable to any policy topic, especially if the researcher is using the Comparative Agenda Project policy topic coding schema. While the focus of this manuscript directly relates to the coverage of climate change in American news media out of personal interest, my hope is that it can help researchers investigate coverage of other policy topics in news media.

2.2 Climate change attention research

If not addressed, climate change poses serious threats various facets of human life (and potentially human life) (Calvin *et al.*, 2023). If unmitigated, climate change will continue to impact United States immigration (Peluso & Harwell, 2001), agriculture (Ortiz-Bobea *et al.*, 2018), economies (Kompas *et al.*, 2018), housing (Krayenhoff *et al.*, 2018), health (Mitchell *et al.*, 2016), racial justice (Schlosberg & Collins, 2014) and other policy topics. The international community has collaborated on climate research for decades to anticipate, mitigate, and adapt to climate changing impacts. A recent Intergovernmental Panel on Climate Change (IPCC) report, for example, emphasizes the breadth of direct costs of climate change as well as positive feedback loops that accelerate and exacerbate expected costs of climate change that rise to the level of existential threat (Climate Change, 2022, p. 1985).

Given the level of academic and international attention climate change receives, why should it matter whether U.S. media gives this issue any attention? In short, higher media attention to climate change may translate to issue salience and climate action. The effect of media attention on issue salience and public issue priorities has been well established in communications and public opinion research (Baumgartner *et al.*, 2008; Iyengar, 1987; McCombs & Valenzuela, 2020; McCombs & Shaw., 1972; Schattschneider, 1960; Weaver *et al.*, 2004, 2004). The more an issue is discussed in news media, the more likely it is to be listed among the public's topic priority concerns. As Cohen (2015) succinctly put it, news media does not tell the public what to think but can tell us what to think *about.* There also appears to be reciprocal effect wherein news media attention influences political agendas such as congressional hearings, which in support maintained attention. (Boydstun *et al.*, 2014b; Walgrave & Hardy, 2017)

The influence of media attention on political agendas holds for issues like climate change, as well. Liu *et al.* (2011) find a correlation between the share of New York Times attention to climate change and congressional attention to the issue. Most of the climate change communications literature (Table 2.1) provides an optimistic outlook: climate change coverage is increasing in domestic media markets as well as international media markets. The overall outlook on climate change coverage suggests an optimistic outlook on public attention and potentially policy action.

Yet, there are several issues relating to standard research conventions in climate change communications research that may misrepresent past and current levels of climate change attention. The central focus of this chapter is the limitations of data collection, classification, and content analysis methods that are standard or common in climate change commutations literature. These data collection methods may introduce unknowable errors into estimates of attention while the classification methods inflate measures of attention with a systematic upward bias. Additionally, common methods used in content analysis are often opaque or task-specific, making collaboration in this field especially difficult. Below I discuss these limitations before turning to alternatives for researchers in this field.

2.2.1 Issues in data selection and collection

Nearly all of the articles listed in Table 2-1 examine data purchased from online third-party news data aggregation and hosting services, such as LexisNexis or Factiva. These services are extremely easy to use as researchers can purchase published news articles meeting user-specified criteria such as source, publication date, and content (or headline) keywords. This method for data collection is particularly appealing because of its minimal skill requirements. However, it is accompanied by several notable drawbacks. These include concerns regarding the quality of the data and any subsequent analyses as well as issues pertaining to reproducibility.

Table 2.1: Climate Change Media Attention Literature Review

	Classification	Sources	Period	N
Hase et al. 2021	[cc, gw, gg]	Print (global)	2006-2018	71,674
Bohr 2020	[cc, gw, gg]	Print (US)	1997-2017	78,599
Keller et al. 2020	[cc, gw, gg]	Print (IMD, English)	1997-2016	18,224
Chinn et al. 2020	[cc, gw, gg]	Print (US)	1985-2017	1,793,439
Vu et al. 2019	[cc, gw, gg]	Print (worldwide)	2011-2015	37,670
Reber 2019	[cc, gw, gg]	Print (US, GRBR, GER)	2014	3,896
Boussalis et al. 2018	SVM modeling	Press releases US cities	2014-2017	76,249
Saunders et al. 2018	[cc]	Print (GRBR)	1997-2017	6,884
Johns & Jacquet 2018	[ocean, pollution]	Print (US)	2001-2015	169
Barkemeyer et al. 2017	Not specified	Print (worldwide)	2007-2009	2,600,000
Belfer et al. 2017	[cc, gw, gg]	Print (US, CAN, ASTL, NZ)	1997-2015	92
Wagner & Payne 2017	[cc, gw, gg]	Print (IRE)	1997-2011	517
Feber et al. 2017	(animal welfare)	Print (UK)	2014	23,811
Brüggemann & Engesser 2017	[cc, gw, gg], authors	Print (US & Europe)	2011-2012	747
Duan & Takahashi 2017	[cc, gw, gg]	Print (US)	2012-2015	841
Boussalis & Coan 2016	[cc, gw]	Conservative think tanks	1998-2013	16,028
Boussalis et al. 2016	[cc, gw, gg]	Print (US & RUS)	1980-2014	11,131
Jang & Hart 2015	Not specified	Twitter users	2012-2014	5,700,000
Ford & King 2015	[cc], (adaptation)	Print (US & CAN)	1993-2013	23,146
Schäfer & Schmidt 2014	[climate]	Print (ASTL, IMD, GER)	1996-2010	44,448
Schmidt et al. 2013	[cc, gw, gg]	Print (worldwide)	1996-2010	152,125
Elsasser & Dunlap 2013	[cc, gw, gg]	townhall.com	2007-2011	203
Holt & Barkemeyer 2012	Not specified	Print (worldwide)	2005-2008	24,000,000
Liu et al. 2011	[cc, gw, gg]	NYT	1969-2005	4,197
McComas & Shanahan 1999	[cc, gw, gg]	Print (US)	1980-1995	376

Note: Terms in brackets "[...]" indicate keywords specified in the publication. The set of keywords [cc, gw, gg] includes "climate change" (cc), "global warming (gw), and"greenhouse gas" (gg). Terms in parentheses "(...)" indicate topics from which unspecified terms were drawn.

First, there is an "emphasis problem." News media aggregation and hosting services tend not to provide indicators for articles that were shared on the front page of news sources' print version. Acquiring data in this way provides researchers with all published news stories from a source but the data lack critical information about how the news source chooses to highlight or emphasize published stories. Unless the researcher augments purchased data with external information about story emphasis, analyses of these data inherently overlook the importance of the front page news generating process, as well as its impact on public opinion and policy (Boydstun, 2008; Boydstun, 2013).

Second is a "numerator problem." When using these services to estimate topic attention as the number of articles on a given topic over all articles, researchers do not have full control over the numerator (i.e., the articles that are considered relevant to the topic of interest). Third-party news hosting services tend to use Boolean classification to identify relevant data for the researcher but some may use more sophisticated methods that are unknown to the researcher. Elsasser & Dunlap (2013), for example, analyze townhall.com articles using the site's search engine. It is unknown what search engine optimization algorithm the townhall.com website used at the time of this research. It may have been simple Boolean searches, something more complex, or it may be based on staff-entered article keywords and search terms, which may be unreliable.

Third, related to the "numerator problem," is a "denominator problem." The core issue here is that the stories counted in the numerator undergo qualitative screening processes that are not applied to all articles in the denominator (Belfer *et al.*, 2017). This discrepancy becomes problematic when calculating attention, which is measured as the percentage of the total number of stories that are primarily about the topic of interest. Perhaps as a cost-saving measure, researchers purchase only the text data deemed relevant to the topic of interest and calculate attention using the total number of articles available (but not purchased) from that source during the specified time period as the denominator. Upon qualitative screening, researchers often find articles outside the scope of the research (e.g., native advertising or sponsored content). This qualitative screening step cannot be applied to data not retrieved. This approach also provides no prior information to the researcher about rates of story irrelevance conditional on topic relevance, which would allow the researcher to estimate the denominator (i.e., it is not clear if data not retrieved are irrelevant at the same rate as retrieved data).

Finally, financial barriers may prevent researchers from reproducing academic research. The number of stories collected is often proportional to cost of third-party news hosting services, and the budget for many researchers may be zero or near-zero. The cost of acquiring news data may create a barrier to replicability of existing research. Given that the size of the retrievable sample is dependent on the research budget, the reliability of findings derived from these data may also correspondingly hinge on the researcher's financial resources. The Data section below provides an overview of free and open-source alternatives to data collection and information retrieval.

2.2.2 Issues in automated classification

Boolean classification is an exceedingly popular method compared to text classification alternatives in climate change communications research. Boolean classification consists of simple researcher-specified decision rules to determine the primary topic of news stories. Nearly all of the research listed in Table 2-1 uses Boolean classification to classify climate change stories and most researchers in this subfield used the same set of decision rules: if the story headline or content contains any one of the set of key phrases [“climate change,” “global warming,” “greenhouse gas”], then the primary topic of the story is classified as “climate change.” In most applications of Boolean classification, a single instance of a keyword term or phrase (i.e., a single “hit”) is considered sufficient to treat the article as giving attention to the topic of climate change.

While simple and straightforward, the Boolean classification approach relies on two assumptions that fail to meet common standards of news media attention under scrutiny.

First, in using Boolean classification, the researcher assumes that the selected keywords sufficiently identify the *primary topic of interest* in news stories. That is, a single hit of "climate change" or "global warming" anywhere in a story's headline or content classifies the story as *primarily about* climate change. Second, researchers assume that the selected keywords sufficiently cover the breadth of the topic of interest while sufficiently discriminating between the topic of interest and all other topics. While the standard set of key phrases applied in climate change communication research is comprehensive for the topic of climate change, it is possible for stories to discuss the causes and implications of a warming climate without mentioning any of the key phrases.

While this section is critical of Boolean classification, note that no single text classification method is best-suited for every scenario and even the best-suited method for a particular scenario will be imperfect. In the words of statistician George E. P. Box, "All models are wrong but some are useful." The usefulness of Boolean classification relative to alternatives can and should be tested directly. Researchers should be aware of the assumptions underlying Boolean classification and, more importantly, test these assumptions. I address each assumption and how it fails below.

2.2.2.1 Primary topic assumption

First, Boolean classification assumes that the selected keywords sufficiently identify the *primary topic of interest* in news stories. However, a news story that mentions a topic may not necessarily be *primarily about* (or even relevant to) that topic. Researchers that implement and discuss quality assurance measures report high rates of false positives (Type I errors) when using Boolean classification (i.e., many stories are classified under the climate change topic but are in fact not relevant to climate change). One notable example is Liu *et al.* (2011), who manually reviewed a random sample of New York Times stories classified using the standard Boolean keyword selection method. Upon manual review, Liu et al. find that only 64 percent of stories with at least one climate change keyword "hit" were actually relevant to climate change.

Other researchers have taken a machine learning approach to examine the accuracy of Boolean text classification. Boussalis *et al.* (2016), for example, train a Support Vector Machine (SVM) model to classify climate change stories using prelabeled data. Boussalis et al. find that only 20 percent of texts with at least one climate change keyword "hit" would be considered primarily about climate change. Based on review by Liu *et al.* (2011) and Boussalis *et al.* (2016), the research listed in Table 2-1 might have inflated the degree of news media attention to the topic of climate change by a factor of 1.5 to 5.

To demonstrate the importance of ensuring the validity of Boolean classification, consider news stories that include a climate change keyword in passing while the primary focus of the story is another topic. Elsasser & Dunlap (2013), for example, found that about seven percent of stories containing climate change keywords were covering the outcomes of the 2006 Academy Awards ceremony, in which *An Inconvenient Truth*, a climate change documentary, was competing for best feature-length documentary. In 2018, an Atlanta-based print news source, *The Root*, began its coverage of a poultry shortage with context to explain the public's fervor for fast food, which included the looming threat of climate change: *"Climate change is real, the political discourse is trash, and people in the U.S. and U.K. just want some damn fried chicken to soothe their souls"* (Branigan, 2018). The remainder of the story focuses on the lengths to which some fast food customers went to obtain a chicken sandwich during the poultry shortage. Despite the climate change "hit," a researcher would be hard pressed to defend this story's classification as a climate change story.

To avoid false positives such as the story above, some researchers have elected to raise the bar on Boolean classification by requiring two or more "hits" in story headlines or content (e.g., Bohr, 2020). Even so, it is common for news stories to mention climate change keywords two or more while focusing primarily on another topic. For example, a 2022 New York Times front page story covered the wide-range implications for President Biden's legislative agenda should Republicans win the House majority (Baker, 2022); this story had five climate change hits but equal attention was given to about a dozen other policy topics.

Stories like the above are quite common among documents with at least one or more hits from the set of climate change-related keywords:

— *G.O.P. Hopes Climate Fight Echoes Health Care Outcome* (5 hits) (Mitchel, 1997)
— *Bush Dismisses Idea that Kerry Lied on Vietnam* (4) (Sanger & Bumiller, 2004)
— *Emphasis Shifts for New Breed of Evangelicals* (3) (Luo & Goodstein, 2007)
— *Tick and Mosquito Infections Spreading Rapidly, C.D.C. Finds* (5) (McNeil, 2018)
— *Biden Condemns Russia as Threat to the World in U.N. Speech* (5) (Tankersley *et al.*, 2022)
— *Inside the Global Race to Turn Water Into Fuel* (3) (Bearak, 2023)

Because the stories above are not primarily about the topic of climate change, they should not be classified as such and should not be treated as news media attention to the topic of climate change. When estimating news media attention to climate change, researchers must be aware that Boolean classification inflates the false positive rate, which may confound and obfuscate our understanding of the causal relationships between media attention to climate change and events, public opinion, and political agendas.

That said, climate change keyword hits in stories primarily about topics other than climate change may provide valuable information about how important political topics permeate public consciousness by appearing as secondary considerations in other policy discussions. This topic is explored in greater detail Chapter 4.

2.2.2.2 Sufficient breadth assumption

The second Boolean classification assumption is that the selected keywords sufficiently cover the breadth of the topic of interest and sufficiently discriminate between the topic of interest and all other topics. Failure to select a sufficiently comprehensive set of keywords will lead to inflated false negatives (Type II errors) and underestimate the level of attention that climate change receives from news media sources.

In general, researchers perform poorly in identifying a sufficiently comprehensive set of keywords to classify stories to a topic of interest (King *et al.*, 2017). Individuals generally have poor recall ability, making ad-hoc creations of sets of Boolean key terms inconsistent and unreliable across individuals. King et al. (2017) run an experiment asking several individuals to create ad-hoc lists of keywords relevant to a topic and find substantial variation in keyword selection across individuals for the same topic. Note, however, that the results of this experiment is generalizable only insofar as keyword selection is ad-hoc.

Researchers should be wary of Type II errors from Boolean classification, as a general rule. In climate change communications literature, however, the standard set of key phrases for Boolean classification is well-accepted and widely applied, as illustrated in Table 2-1.[1] Boolean classification for climate change stories is indeed less prone to Type II errors than Type I errors; however, the prevalence of Type II errors is rarely tested in climate change communication literature.[2]

While relatively rare, Type II errors may occur in Boolean classification of climate change stories. News stories may discuss the causes (such as deforestation, carbon dioxide and methane emissions) and implications (such as rising sea levels, ecosystem disruption, and increasing propensity for extreme weather) of a warming climate without mentioning any of the standard key phrases. See the previous sentence, for example. For additional examples, see:

— *Old Ways of Life Are Fading as the Arctic Thaws* (Myers *et al.*, 2005)[3]
— *With Something for Everyone, Climate Bill Passed* (Broder, 2009)
— *Rising Sea Level Tied to Faster Melt* (Naik, 2013)
— *Scientists Report the Planet Was Hotter than Ever in the First Half of 2016* (Joyce, 2016)
— *Where Are America's Winters Warming the Most? In Cold Places* (Popovich & Migliozzi, 2018)

Boolean classification requires a trade off between selecting a sufficient number of terms to reduce false negatives and sufficiently discriminating terms to reduce false positives. As noted above, no classification model is without its flaws and all are subject to this bias-variance

[1]The set of key phrases in climate change communication literature consists of "climate change," "global warming," and "greenhouse gas."

[2]See the Results section later in this chapter for observed rates of false negatives and positives across classification methods.

[3]Note: The standard set of keywords for Boolean classification of climate change articles do not appear in the purchased, truncated article text; the keywords do appear in the full text, however. This further illustrates the power of machine learning methods to identify articles about climate change by identifying other terms that are frequently associated with the topic of climate change.

trade off but alternative text classification methods have fewer and less severe limitations compared to Boolean classification. Researchers have at their disposal relatively simple alternatives (and quality assurance measures) to reduce false positives and false negatives through Naive Bayes and machine learning methods discussed later in this chapter.

2.2.3 Issues in automated content analysis

The gold standard of content analysis is human-led content analysis (Belfer *et al.*, 2017; Brüggemann & Engesser, 2017; Elsasser & Dunlap, 2013; Feber *et al.*, 2017; Johns & Jacquet, 2018; Wagner & Payne, 2017). Despite recent advances in large language models and natural language processing, humans continue to outperform machines in interpreting textual meaning, context, humor, sarcasm, satire, and nuance. However, human-led content analysis is challenging to scale. Online content creation experienced a massive expansion in recent years and will likely continue to grow, especially with recent advances in AI-generated content in news media. It is more important now than ever for climate change communications researchers to redouble efforts to develop methods (or use developed methods) that approach human-level abilities to retrieve valuable information and examine trends in text content.

Some climate change communications researchers have attempted to automate or partially automate content analysis to enable research pipelines to scale with ever-expanding creation of content. At the time of writing this manuscript, methods in automated content analysis have greater limitations than some researchers might admit. Researchers should pursue automation in content analysis but should be aware of limitations in the following methods.

2.2.3.1 Keyword decision trees or dictionary methods

Stone & Hunt (1963) used complex decision trees involving keywords in content analysis: MIT's General Inquirer system used a large set of linguistic rules and manual parts-of-speech tagging to produce automated content analysis. For example, the system was able to identify the specificity of post-mortem requests in order to distinguish between real and simulated suicide notes; real suicide notes were highly specific in their requests to pay the bills for example while simulated notes were more general.

Boolean-driven decision trees have also been used in climate change communications content analysis research. Jang & Hart (2015), for example, used Boolean classification to identify climate change stories and applied a series of Boolean decision rules to examine trends in frames related to the veracity, impacts and causes of climate change. Boolean approaches to content analyses face the same challenges as Boolean classification: the rules established may inflate false positives and false negatives in frame identification. Jang et al. (2015) dismiss this concern as they evaluate the relative frequency across frames but these scholars ignore the inevitability of variance in Type I and Type II errors across frames. That is, any differences in estimated rates of referring to climate change as a "hoax" versus "impact" frames, for example, could be due to measurement error.

Type I and Type II errors aside, Boolean or dictionary decision rules are notoriously difficult to maintain. There are several dictionaries available for estimating text sentiment (positive, negative or neutral) including thousands of terms (e.g., Hu & Liu, 2004). Such dictionaries are often (but not always) created and maintained manually, rather than through applied inference methods. The amount of manual work required to maintain rules for complex topics may quickly become too cumbersome to manage. Dictionary and decision tree-based methods may be more appropriate for sentiment analysis or other analyses where the underlying dimensions producing observed data are relatively few in number (e.g., the negative to positive dimension in sentiment analysis). However, probabilistic methods have shown greater accuracy over rules-based methods in approaching the gold standard in sentiment analysis (Miyato *et al.*, 2016).

2.2.3.2 Statistical topic clusters as frames

Several climate change communications researchers have applied statistical topic-term clustering models (or "topic models") on text documents and have treated the topic model features as “frames” (Bohr, 2020; Boussalis *et al.*, 2016; Boussalis & Coan, 2013, 2016; Keller *et al.*, 2020; Vu *et al.*, 2019). The features extracted from topic models have been shown to be useful in identifying trends in textual data and classifying text (Chen & Li, 2016; Inkpen & Razavi, 2014). However, there are substantial challenges that must be overcome to justify treating topic model features as frames.

Framing a topic consists of emphasizing a particular dimension of that topic to make the dimension more salient, noticeable, meaningful or memorable (Entman, 1993). A framing analysis of climate change news stories might identify the prevalence and analysis of misinformation. Such an analysis might distinguish between messages that frame climate change as a hoax and messages that denigrate climate deniers, despite the similarities in key terms used. Consider the following two messages, the first from former President Trump and the latter from President Biden:

— *“The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive”* (@realDonaldTrump, Nov. 6 2012).
— *“Nobody can deny the impact of climate crises, at least nobody intelligent can deny the impact of the climate crisis anymore”* (President Biden, Sep. 3 2023).

Human readers might easily identify the former message as an example of climate change denierism and the latter as a denigration of denierism. However, a topic model would likely struggle to identify themes of denial in the example of denierism. Unless the body of text (“corpus”) used to train the model had several text entries similar to the first statement above, a topic model would more likely identify themes related to global warming, international relations and manufacturing rather than a “denierism” or “misinformation” topic.

Latent Dirichlet Allocation (LDA) is a commonly selected topic model for automated content analysis, which can identify the portions of documents that belong to latent topics in a body of text. The LDA topic model uses an unsupervised approach to cluster terms based on their relative frequency of appearing in similar contexts (i.e., in the same documents) as all other terms. The model assumes that in a body of documents there are latent topics from which terms are randomly drawn in order to generate the observed texts in that body;

through an iterative process, the model estimates terms' topic probabilities that would be most likely to produce the observed body of documents (Blei *et al.*, 2003).[4]

The LDA model produces estimates of topic distributions per story (i.e., the share of each document belonging to each latent topic), referred to here as "topic model features," which do not fit the established definition of frames in communications literature (Entman, 1993). LDA and other topic models are considered a "bag of words" models: they are agnostic toward the order of terms in texts and therefore ignore the nuance and complexity required of framing analysis. Topic features may be able to identify news stories with a meaningful amount of attention given to, perhaps, a "hoax" topic; but this approach cannot distinguish between "climate change is a hoax" and "calling climate change a hoax is idiotic." The mere presence of frame-related keywords is not equivalent to the presence of a frame.

The assertion that features derived from LDA topics should not be labeled as "frames" extends beyond mere semantics and is rooted in more substantive considerations: referring to topic model features as "frames" may insinuate that topic model features and issue framing share the same effects on individual-level opinion (e.g., Tversky & Kahneman, 1981) and policy outcomes (e.g., Baumgartner & Boydstun 2008) established in communications research. Topic model features are not frames and the effects of topic features on public opinion, agendas, and policy outcomes have not been tested empirically, and so should not be referred to as frames.

This is not to say topic modeling cannot be used in content analysis. In fact, topic features from an experimental LDA derivative are used in this chapter and Chapter 3 for text classification as well as Chapter 4 in a novel approach to content analysis. However, researchers must be careful not to conflate the mere presence of frame-related keywords as the frame itself. Hase *et al.* (2021), for example, refer to topic model features as *themes* rather than frames to compare trends in climate change communications content across sources in various countries. Reber (2019) refer to topic model features strictly as *topic prevalence*. In

[4]See the next chapter for a fuller description of the Latent Dirichlet process.

the fourth chapter, topic model features are also referred to as *issue associations* since the topic model is constructed so that latent topics are centered around various policy issues.

In the following section, I review alternative methods to data collection and management. This discussion will lead into alternative methods for text classification as well as useful applications of Boolean decision rules and topic model features. At the end of this chapter, I review the performance of alternative classification methods.

2.3 Data

The remainder of this chapter compares model performance in predicting the major topic (and minor topic 705, "Air Pollution, Global Warming, and Noise Pollution") assigned to stories in the New York Times Front Page data set (Boydstun, 2014) and supplemented with data retrieved from Twitter and various news media websites. Professor Boydstun's data set contains front-page stories from the New York Times between 1996 and 2006. Articles were assigned major and minor topics based on a modified Comparative Agendas Project topic topic scheme (Jones & Baumgartner, 2005b). These data were supplemented with a random sample of about 20,000 news stories tweeted by a diverse set of news media content creators between 2007 and 2023, manually coded using the same topic scheme.

2.3.1 Data collection

Data collection procedures for this project consisted of four custom programs built in Python: 1) a program built to retrieve recent tweets using the open-source Twitter API, 2) an extension of the first program that collects tweets older than what is accessible via the Twitter API, 3) a web crawling program that navigates to tweeted links and retrieves the full textual content of the tweeted story, and 4) a fail safe program that collects the content of missing stories and broken hyperlinks using Wayback Machine, a nonprofit organization that maintains archives of internet postings long after they have been archived, relocated, or deleted by the original authors.[5]

2.3.1.1 New York Times Front Page, 1996-2006

The New York Times front page (NYTFP) data (Boydstun *et al.*, 2014b) consist of about 30,000 stories published in the New York Times between 1996 and 2006. Professor Boydstun led a team of researchers in classifying each of these stories according to their primary topic in the Comparative Agenda Project topic scheme. Researchers read the headline and first three paragraphs of each story to classify at the two-digit (major topic) and four-digit (minor topic) level, as well as a six-digit subtopic level.

2.3.1.2 Supplemental data (NYTFP and tweeted stories), 2007-2023

New York Times front page stories published between January 2007 and April 2023 were collected using the NYT developer API. Additional story information was retrieved using the custom programs built in Python described above. These data were supplemented with news stories tweeted by various online news content creators between 2007 (or the earliest date the sources began tweeting) and 2023. These data were also collected using an API with supplemental data collection efforts. The combined database totaled to about 1.1 million rows after screening and processing the data. Collected data represent nearly all of the stories tweeted by the various source examined (see Appendix A for a fuller description of data wrangling methods).

The full database (N = 1,085,260) is examined in greater detail in Chapter 3 and 4. These stories were from a diverse set of news content creators, selected specifically to increase the diversity in how climate change is discussed in order to bolster the robustness of the classification model on external data. In addition to stories tweeted by the New York Times (@nytimes), this research examines tweeted news stories from ideologically diverse sources including the perceived left-leaning National Public Radio (@NPR), perceived center-right Wall Street Journal (@WSJ), and right-leaning The Daily Wire (@realDailyWire). The

[5] See Appendix A for more information and sample code.

selected sources are also racially diverse in terms of sources' intended audience: stories include those published by black-owned magazines The Root (@TheRoot) and The Atlanta Black Star (@ATLBlackStar) and the Spanish-language newspaper based in Los Angeles, LA Opinion (@LAOpinionLA). Stories also include those tweeted by Christianity Today (@CTmagazine) for representation of the largest religious affiliation in the United States.

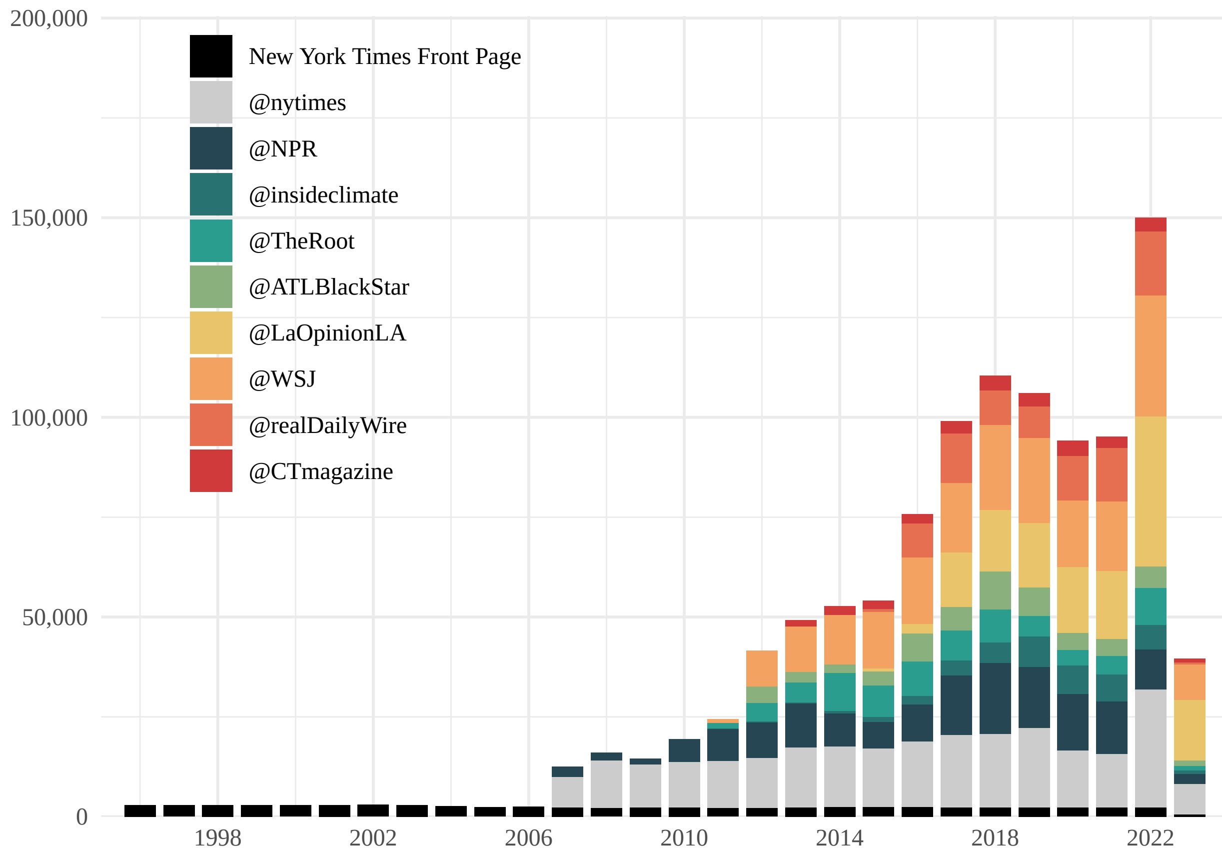

Figure 2.1: New York Times and tweeted stories (Jan. 1996–Apr. 2023)

Note: Figure 2.1 illustrates the number New York Times front page stories and tweets containing valid links to source stories aggregated by year. Valid links direct users to stories on the source's webpage that are active or accessible via Wayback Machine and were not identified as an advertisement, self-promotion, or multimedia.
Source: New York Times front page data set, 1996-2006 (Boydstun); data collection using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023).

As mentioned in Chapter 1, New York Times front page coverage of climate change was rare between 1996 and 2006. There were just 124 stories in this time period that were manually classified as climate change stories. To ensure that the machine learning models had sufficient positive observations for model training, stories from Inside Climate News (@insideclimate) were also collected for analysis. Inside Climate News is a niche news organizations focused strictly on the issue of climate change and other environmental issues. These stories also provide information to help the machine learning programs identify the characteristics that discriminate stories that are primarily about climate change from stories that merely mention or discuss climate change.

2.3.2 Data screening

Text processing and story screening rules were built into the data collection process. Stories were first sorted into different types according to their relevance to this research. Tweeted links to advertisements such as New York Times's "wirecutter" promotions, to self-promotional materials (e.g., "Subscribe to our newsletter!"), and to multimedia such as audio, video, and interactive graphics were removed from analysis (as these links were irrelevant or had no text data to analyze). Tweeted hyperlinks leading to stories outside the source's domain were also dropped (e.g., @nytimes tweets linking to ap.com were dropped) as the data wrangling programs were not built to handle all possible tweeted web domains.

These data wrangling procedures address several of the drawbacks of data purchases from third-party news hosting services:

- Purchased data include stories published rather than the stories that were deemed sufficiently important to distribute via Twitter (again, not all published stories are tweeted). The data collected and analyzed here include stories that the sources chose to emphasize as either front page news or in a tweet.

- The “numerator problem” and “denominator problem” are eliminated, as the same pre-processing and screening steps are applied to all data rather than only the data in the numerator when measuring attention to a given topic, like climate change.
- Having access to these data allows researchers to apply more robust classification techniques to identify climate change stories, rather than delegating this classification task to sources’ search engines or using Boolean classification.
- Issues relating to replicability are mitigated as sample code to generate these data has been provided in supporting appendices. Any interested researcher can essentially copy-paste and run that code to generate the same data analyzed here.

Figure 2.1 illustrates the count of all tweeted stories following screening and fail safe procedures described above. The figure shows the massive expansion in the amount of content to analyze, reaching an annual total of 150,000 tweeted stories in 2022 (note that the total for 2023 is much lower as the study period ends on April 12, 2023, when NPR left Twitter after being labeled as state-sponsored media by Twitter’s new owner). One limitation of textual analysis alone (as opposed to visual and auditory analysis) is visible in the number of tweets collected in 2020 and 2021: during COVID-19 pandemic, sources like @nytimes and @WSJ frequently tweeted links to interactive COVID tracker applications, which were dropped due as these applications did not contain text data for analysis. Attention to climate change during these years may be overestimated when analyzing text data alone, as omitted data were not random with respect to the major topic of the tweeted link.

Table 2.2: Summary of New York Times and Twitterverse Training Data (1996-2023)

Source	No. CC	Total	% CC	Min. Year
New York Times Front Page (Boydstun)	123	27,974	0.4%	1996
New York Times Front Page	8	596	1.3%	2007
@nytimes	47	5,171	0.9%	2007
@NPR	20	3,206	0.6%	2007
@WSJ	10	3,671	0.3%	2011
@TheRoot	1	1,704	0.1%	2011
@ATLBlackStar	1	1,347	0.1%	2012
@insideclimate	267	597	44.7%	2012
@CTmagazine	0	604	0%	2013
@realDailyWire	14	1,629	0.9%	2015

Note: Table 2.2 indicates the number of observations per source in the training data as well as the number of articles manually classified as climate change articles.
Source: New York Times front page data set, 1996-2006 (Boydstun); data collection using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023).

2.3.3 Data processing

Text processing may be more critical to model performance than model selection. In text classification tasks, the terms in the text are model predictors; best practices include processing the text in order to omit confounding and uninformative predictors. Extensive cleaning procedures were applied to the text analyzed in this report. For each source, the text was separated by paragraph and then by sentence. The most common paragraphs and sentences were reviewed for each source to identify patterns of sentences and paragraphs that should not be included in analysis (e.g., "Associated Press contributed to this story."). Additional text processing procedures were applied to simplify and and clarify terms (e.g., consolidating acronym formats like "E.P.A." and "EPA").

Processing time for machine learning text classification models increases with additional features and documents; it is common practice to remove term suffixes through word stemming and simplify terms to their roots through lemmatization. Stemming and lemmatization are helpful feature reduction tools to simplify models and improve model training time and accuracy. Some scholars in communication and linguistics, however, argue that stemming and lemmatization are not critical steps for classification accuracy given enough data and time, which can be tested directly (Manning *et al.*, 2009; Scharkow, 2013). After testing different data processing steps, it was determined that models best fit these data when using lemmatization but not stemming, which oversimplified terms and reduced model accuracy.[6]

Other feature reduction techniques used for this analysis included the removal infrequent terms (appearing less than 300 times in the 50,000 pre-labeled stories for model training and testing) and the removal of frequent but uninformative or possibly confounding terms such as verbs that might appear in any news story (e.g., "say", "comment", "announce"). The R package `quanteda` was used to identify bigrams, concatenations of words that appear together frequently in a corpus (Benoit *et al.*, 2018). Bigrams may contain valuable information that each word or unigram in a bigram does not contain alone. For example, "climate" and "change" have very different meanings apart than when together as "climate change." Concatenating terms in this way adds additional high-information predictors to text classification models.

2.3.4 Data selection and labeling

For the task of comparing different classification methods, about 20,000 stories were randomly selected from the full database for manual topic classification. The tweet, headline and content of each tweeted story was read to classify stories at the two-digit (major topic) level. These articles were also classified according to whether they were primarily about the topic of climate change or not. A separate random sample of about 230,000 stories were used to train a guided topic model discussed below.

[6]See Barberá & al. (2019) for a review of text data processing procedures and considerations.

2.4 Methods

It bears repeating that no one method in text classification is best-suited for all text classification tasks. Additionally, all automated text classification methods are error-prone approximations of manual coding with extensive quality control, the gold standard in text classification. That said, automated text classification using Boolean keyword selection is among the least useful models tested below in the task of identifying articles that are primarily about climate change. Keywords must be used in classification tasks but there are more sophisticated researcher-keyword relationships that can be built into a research pipeline with automated elements.

If all models are wrong, then why automate? It also bears repeating that the sheer magnitude of news media content constantly being generating is staggering. The database of human-coded stories in this chapter includes about 50,000 rows and spans nearly 40 years. But these data represent only four percent of the full database examined in the following two chapters (and the full data examined in the next two chapters is a small fraction of the full internet news catalog). The amount of text data being created increases by the day and recent advances in AI-generated content may serve to accelerate content creation online. Automated classification and content analysis are necessary tools in the face of automated content creation. The discussion below provides a brief introduction to the intuition behind several text classification and content analysis methods that improve over Boolean keyword classification and analysis.

2.4.1 Naive Bayes classification

Due to its speed, simplicity, and relatively high accuracy, the Naive Bayes (NB) classifier is typically used to produce a baseline accuracy benchmark to compare against more complex machine learning text classification methods. Unlike traditional user-defined dictionaries and rules for classifying texts, the Naive Bayes Classifier creates a probabilistic dictionary using pre-labeled data. The NB classifier uses two characteristics to estimate documents' topic probabilities. The first characteristic of the NB classifier is the prior probability of observing a topic, $P(t)$. Prior topic probabilities are calculated as the frequency of documents labeled as topic t divided by the size of the training data. The second characteristic is the product of probabilities of observing terms or features[7] (w) in the collection of documents labeled to topic t, $P(w_k|t)$.

The speed of NB classification is due to a perhaps oversimplifying assumption that all terms in a document are selected independently (an assumption not required in other methods). Nonetheless, this method is simple, interpretable, and robust. Scharkow (2013) implemented an NB classifier with eight topics in German news and found that automated text classification was nearly as reliable as human classification for certain topics. Additionally, the predictive performance of this method (and machine learning methods) is partially conditional on the clarity and separability of the text data. Scharkow (2013) showed that topic-conditional NB predictive performance is related to intercoder reliability: predictive performance suffers on topics where intercoder reliability is low, such as topics with confounding terms or concepts (e.g., crime versus controversy). This work suggests that there may be a cap on the performance of automated classification depending on semantic similarities between topics. However, machine learning models tend to outperform NB models in text classification tasks.

2.4.2 Machine learning methods

There is a wealth of literature of machine learning text classification techniques, all of which are too complex for a full description of the methods as the NB section above. Two supervised machine learning techniques lead in text classification accuracy: Support Vector Machines (SVMs) and Neural Networks (NNs).[8] SVMs have been used in the social sciences for a variety of text classification tasks: Hillard *et al.* (2007) and Purpura *et al.* (2008) used

[7] Usage of "feature", "term" and "token" in this manuscript refers to a single word, processed non-word or collection of words combined into a single feature.

[8] Hvitfeldt & Silge (2021) is an especially helpful introduction to text analysis using Neural Networks, Joachims (1998) for an introductory guide to Support Vector Machines and Kuhn & Johnson (2013) for SVM application in R.

SVM models to classify congressional bills to evaluate government agendas. Verberne *et al.* (2014) applied an SVM model to partitions of elections manifestos to operationalize political parties' policy priorities. Researchers have also found Neural Networks to perform well in classifying the primary topic in news stories (Nassif & Fahkr, 2019; Yan & Zheng, 2020). NNs are also especially effective in classifying text sentiment; for example, Sandhu *et al.* (2019) evaluate the sentiment of tweets about the Supreme Court in order to gauge public opinion about this institution.

2.4.2.1 Support Vector Machines

Support Vector Machines identify decision boundaries or hyperplanes that best separate classes. If data are well-separated by predictors, SVMs identify the boundary decisions on predictor dimensions that maximizes the width or margin of that boundary between classes. For example, if a predictor is a "climate change" topic dimension, SVMs identify the boundary at which a story's "climate change" topic proportion indicate that a story is primarily about climate change as well as the width of that boundary. For example, perfectly separable data might indicate that non-climate change stories have a climate change topic proportion of about one percent on average while climate change stories have ten percent on average. With a logistic regression, any climate-change-topic coefficient between one and ten would perform equally well. SVMs address by estimating the width of the decision boundary.

In cases in which the classes are not perfectly separable, Cortes & Vapnik (1995) developed a cost parameter that penalized observations in the training set that were within the margin or on the wrong side of the support vector. Lineary SVM models were estimated using the `LiblineaR` package in R (Fan *et al.*, 2008; Helleputte *et al.*, 2022). Non-linear Support Vector Machines are also popular but not necessary (or recommended) when the number of features is large, as in text classification (Fan *et al.*, 2008).

SVM was introduced as a method for text classification by Joachims (1998). SVM text classification uses "features" or one-hot representations of terms instances in documents as predictors of the outcome variable. Text data structured with each feature or term representing its own predictor can be extremely sparse, as vocabularies may range in the tens or hundreds of thousands while only a small fraction of those features appear in a single document. Joachims (1998) compared the SVM approach to other conventional methods of classification at the time and showed that SVM consistently and considerably out-performed other methods in terms of precision, recall and computational rigor. He found that SVMs perform well even with high dimensionality and handle sparse vectors well.

2.4.2.2 A note on Neural Networks and feature embeddings

Neural Network text classification models typically rely on feature embeddings trained on external text data sources, such as Wikipedia. A Bidirectional Long Short-Term Memory Convolutional Neural Network (BiLSTM CNN) was used to predict major topics following Hvitfeldt & Silge (2021); this model had relatively low performance compared to SVM classification techniques. One explanation for poor CNN performance here is that the Comparative Agendas Project topic scheme may not map well onto pre-trained word embeddings. Word embeddings are trained on massive amounts of text data and reduced to an embedding matrix of just 300 dimensions (Pennington *et al.*, 2014); each term is represented as a vector in this embedding matrix. It is possible that the 300 dimension embedding matrices available online[9] oversimplifies highly related topics, such as Agriculture and Environment, International Affairs and Defense, or Macroeconomics and Banking and Finance.

Another possible explanation for low CNN performance is that the program simply did not have sufficient data or time to train for optimization of various model specifications. Future application of NNs for the purposes of CAP classification may require custom feature embeddings applied to data larger than the N $\approx$ 50,000 training data used here.

[9] https://nlp.stanford.edu/projects/glove/

2.4.3 Enriched features and dimensionality reduction

Recent advances in classification techniques have shown potential improvements using "enriched" document features from feature reduction techniques such as Latent Dirichlet Allocation (LDA) document features and Global Vector (GloVe) as predictors (Chen & Li, 2016; Dogru & al., 2021). SVMs and NNs typically use terms as predictors: either as a binary presence of the term in a document, the number of appearances of the term in a document, or the term counts weighted by their prevalence and topic exclusivity in a corpus (term frequency-inverse document frequency). Using terms as predictors yields highly sparse data where there may be hundreds of thousands or millions of predictors and most predictors contain mostly zeroes.

Enriched features are simply the output of feature reduction techniques meant to reduce model dimensionality by assigning topic probabilities or dimensional representations to terms. For example, a term like "global warming" might be assigned a high probability of belonging to a latent topic labeled "climate change" and low probabilities for all other latent topics; through the LDA process, topic probabilities are estimated for every term in the corpus and each document can be represented as a mixture of topics. Say an LDA is applied to a corpus with the number of latent topics (K) set to 3; a document can be represented as a Dirichlet probability distribution (or topic mixture) over the three topics, where the three always sum to 1 (e.g., 0.7, 0.2 and 0.1).

Rather than replace the hundreds of thousands of term predictors, researchers in computational linguistics and natural language processing opted to add the features as predictors for an increase in classification accuracy (Chen & Li, 2016; Dogru & al., 2021; Inkpen & Razavi, 2014; Nassif & Fahkr, 2019; Yan & Zheng, 2020). SVM models can be simplified (or enhanced) by using document-level topic distributions from Latent Dirichlet Allocation (LDA) instead (or in addition to) terms and features (Inkpen & Razavi, 2014). Feature embeddings, pre-trained numerical representations of features and phrases, have similarly been used to enrich NNs, simplifying models and improving their performance (Nassif &

Fahkr, 2019). As discussed further in the Results section, the climate change communications literature can benefit from adopting one or more of these techniques.

2.4.3.1 Latent Dirichlet Allocation

LDA is an unsupervised machine learning topic model that produces topic distributions within text documents and topic probabilities for terms in a corpus, allowing researchers to use LDA output for content analysis, examine word similarity in the context of inferred topics and improve classification accuracy. The LDA model is referred to as a "bag of words" model, in that it is agnostic to the order of words in a document. Documents are treated as random mixtures of latent topics: documents are collections of terms randomly drawn from a Dirichlet distribution representing K latent topics.

The goal of the LDA model is to find the term-topic probabilities that maximize the likelihood of producing a topic distribution in a corpus.[10] The model uses Bayesian inference (Gibbs sampling) to update probability terms' latent topic probability distributions until convergence or max iterations have been reached. The output of the process is a matrix terms' probabilities of being drawn from each topic, as well as the topic distribution for each document in the corpus.

LDA is an incredibly powerful tool that is well established in natural language processing literature and beyond. But it is not without its drawbacks. First among these is the task-centric nature of LDA: it is designed to produce topics for a corpus and has limited application beyond that corpus and research question. LDA topics must be assigned topic names in an ad-hoc manner, making it difficult for researchers to compare findings based on LDA topics. Similarly, it is difficult to apply prior information about a domain to a new or external corpus using LDA (applying insights from past research about the New York Times front page database and the Comparative Agendas Project topic scheme, for example). There are supervised machine learning versions of LDA (sLDA) that draw Dirichlet priors conditional on pre-labeled data (Glenny & al., 2019; Lakshminarayanan & Raich, 2011; Mcauliffe &

[10] See Blei *et al.* (2003) for a full review of the LDA process.

Blei, 2007). While helpful for classification tasks, its application for content analysis and emerging theme analysis faces the same limitations as LDA, particularly an absence of prior information about the structure of latent topics in the body of documents.

2.4.3.2 Guided Latent Dirichlet Allocation

This report makes use of an experimental derivative of LDA that exploits researchers' prior domain knowledge to develop replicable topic structures (Eshima *et al.*, 2020; Jones *et al.*, 2021). The key difference between the Guided LDA and past iterations is that the user can structure the expected topics along with words central to each topic. LDA, very briefly, uses an iterative process to determine which words are frequently used in the same context and therefore belong in the same topic. In LDA, initialization parameters (or starting values) for terms' topic probabilities for that sorting process are random. Term-topic probabilities are then updated and optimized through Gibbs sampling, which clusters words that appear in similar contexts.

Guided LDA in contrast enables researchers to use non-random, weakly informative prior term-topic probabilities to be updated and optimized in the Gibbs sampling process. In other words, guided LDA begins with terms in a more likely starting point than a random starting point. The sorting process then begins with a selected set of terms with higher probabilities of belonging to their appropriate topics. For example, "greenhouse gas" might start with a higher probability of belonging to a "climate change" topic. Then, through the iterative sorting process, terms used in similar contexts to "greenhouse gas" will also be assigned higher probabilities of belonging to a "climate change" topic. With regular LDA, "greenhouse gas" is assigned a random probability of belonging to any K number of topics. Through the iterative process, a "climate change" topic *might* form but this method requires the researcher to find that topic and label it appropriately based on the highest probability terms in that unlabeled topic.

Because prior information supplied to the Guided LDA process is weakly informative, this method of establishing non-random initialization parameters is non-deterministic. However, the method also allows the researcher to set deterministic priors to prevent the LDA process from producing topics that are hyper-specific to the corpus or research task. For example, the researcher might set topic probabilities for certain terms to zero to prevent certain terms from being assigned to certain topics, even though the terms are frequently used in similar contexts. These deterministic elements should be used sparingly, and were used sparingly for this research. For example, the prior probability that the term "war" belonged to the "Energy" topic was set to zero to assist the program in delineating between nuclear energy and the minor topic, "Arms Control and Nuclear Nonproliferation." Without this human-led intervention, there would be confounding term-topic probabilities in the CAP topics *Energy* and *Defense* due to the common usage of terms like "nuclear" in these two major topics.

While the researcher can supply the LDA process with prior information such as topic structure, this approach grants a great degree of flexibility. In addition to updating the topic-term probabilities through Gibbs sampling, this process also allows the researcher to set any K number of topics on top of the established topic structure. This flexibility enables the LDA process to identify any topics that might not have been accounted for in the process of developing the topic structure and prior term-topic probabilities. Guided LDA also removes the ad-hoc nature of naming LDA topics, provides a coherent structure the topics, and connects the research to past research using the Comparative Agendas Project (or Policy Agendas Project) schema. The `tidylda` package also enables researchers to refit the model with new information, enabling more room for direct collaboration between researchers.

The Guided LDA approach grants enhanced replicability through transparent prior information about terms and topics. In this research, I created a prior term-topic probability table by reviewing and selecting the most frequent and discriminating words per minor topic in the CAP topic scheme.[11] A Guided LDA model with 255 topics based on the CAP topic

[11]See Appendix B for the top five terms for each minor topic in the modified CAP topic topic scheme. Note that each of the top five terms has an indicator for whether it was one of the terms selected for weakly informative priors.

scheme was trained on a random sample of about 230,000 unlabeled stories tweeted by the New York Times, Inside Climate News, The Root and other news content creators examined in these chapters. The K parameter for the number of unguided topics was set to 12 to allow some flexibility in the production of assignments. The model was used to predict topic distributions for stories in the training data described above (2006-2023). The estimated topic distribution for each article was determined using the average of 100 iterations of Guided LDA topic distribution estimations.

2.5 Results

In this section I review the results of the various models described above in terms of classification accuracy on documents' assigned major topic topic code from the CAP topic scheme. The prelabled data were subset into a training set (75 percent of the data) and a hold out test set (25 percent) to test each method's out of sample (OOS) performance. Machine learning methods were trained using five-fold cross validation to maximizing training accuracy while attempting to prevent over-fitting to the training data. SVM models were trained to optimize the cost parameter using coarse and fine tuning.

The first set of results includes model performance on training and test subsets of the data overall, followed by a review of model performance for each major topic. Results indicate that the SVM model without enriched features tends to slightly outperform the enriched SVM on most (but not all) topics. Next, I review the performance of each method in identifying articles that are primarily about the topic of climate change. In the task of identifying climate change articles, a two-stage enriched SVM slightly outperforms a two-stage SVM.

2.5.1 Major topic classification model comparison

Model performance is illustrated in Figure 2.2, which includes the predictive accuracy in major topic classification for both the training and test subsets of the prelabeled data. Results indicate that the relatively simple linear Support Vector Machines are equally suited to predict news stories' primary major topic. The SVM models using terms as predictors showed greater performance in major topic prediction (78 percent OOS accuracy) compared to Naive Bayes classification (72 percent). Enriching the SVM model with guided LDA topic model features appears to have had very little impact on classification accuracy, overall. Given the similarities in the performance of these SVM models, any differences in performance may be due to differences in the optimization of tuning parameters in the training process.

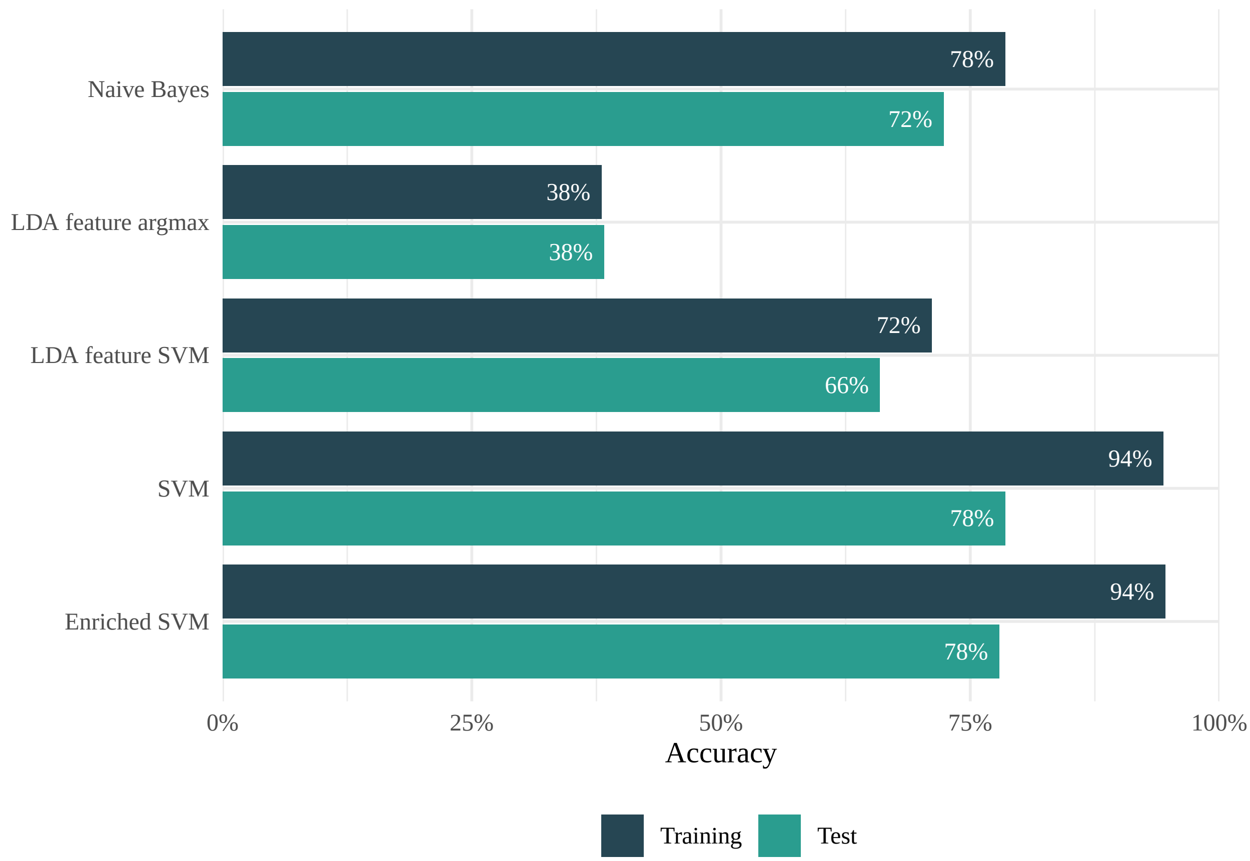

Figure 2.2: Comparison of major topic classification model performance

Note: Figure 2.2 illustrates various models' performance predicting news stories' primary topic according to a modified Comparative Agenda Project codebook.
Source: New York Times front page data, 1996-2006 (Boydstun), New York Times front page data, 2007-2023 (Broad), tweeted articles collected using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023). N = 46,499.

Of note in Figure 2.2 is the performance of the guided LDA features. Classification using the maximum guided LDA feature per story (*LDA feature argmax*) is poor (38 percent), which is unsurprising: stories often cover more than one topic and bulk of the content in a may not be related to the central focus of the story. For example, any one of these stories may choose to discuss a policy topic in terms of its legislative history with most of the content focusing on the actions of governing bodies. However, the moderate performance of major topic classification with SVM using only guided LDA features provides evidence for the validity of these LDA features. It is also a testament to the strength of the guided LDA approach considering that the program was not trained on the same set of stories as the classification model training data: the LDA training data were comprised of a random sample of about 230,000 stories pulled from the full million-row database.

2.5.2 Per-topic classification model comparison

Table 2.3 includes out of sample model predictive accuracy (F1 scores) by each of the 28 major topics in the CAP coding scheme, including an "Other" topic. The SVM model without enriched features tended to slightly outperform the enriched SVM model. Differences between these methods were fairly negligible and may be due to differences in cost optimization in the training process. Given that the SVM tended to outperform the enriched SVM, major topic predictions from the SVM model were added as a predictor in training the two-stage climate change classification model (i.e., the first stage is the classification of the major topic while the second stage is the classification of climate change while considering the results of the first stage).

Table 2.3: Comparison of major topic classification model performance

Major topic	Naive Bayes	LDA feat. argmax	LDA feat. SVM	SVM	Enriched SVM
Macroeconomics	73.6%	33.8%	65.5%	**78.1%**	77.1%
Civil Rights, Minority Issues, and Civil Liberties	59.2%	27.6%	45.0%	**64.7%**	64.2%
Health	78.7%	63.3%	76.1%	**84.2%**	84.0%
Agriculture	57.5%	32.6%	49.1%	**65.8%**	62.9%
Labor and Employment	55.5%	31.2%	43.5%	65.4%	**66.6%**
Education	79.3%	59.1%	70.3%	**82.6%**	81.5%
Environment	69.4%	26.8%	66.5%	**76.1%**	73.2%
Energy	71.2%	41.8%	63.4%	**80.2%**	79.8%
Immigration	56.6%	39.5%	54.2%	66.2%	**72.9%**
Transportation	73.4%	57.8%	62.7%	**74.0%**	73.1%
Law, Crime, and Family Issues	71.6%	38.4%	65.7%	**77.8%**	76.9%
Social Welfare	62.9%	36.2%	60.4%	**77.2%**	76.5%
Community Development and Housing Issues	60.5%	41.6%	56.6%	**69.7%**	68.2%
Banking, Finance, and Domestic Commerce	61.2%	35.1%	53.1%	**66.7%**	64.6%
Defense	73.3%	51.4%	65.9%	78.7%	**78.9%**
Space, Science, Technology and Communications	66.1%	43.2%	57.7%	**68.8%**	67.9%
Foreign Trade	43.3%	7.5%	54.1%	**59.0%**	58.6%
International Affairs and Foreign Aid	77.8%	41.6%	72.5%	**85.4%**	85.0%
Government Operations	82.3%	24.6%	73.9%	**84.5%**	84.5%
Public Lands and Water Management	51.4%	23.7%	30.6%	54.7%	**57.9%**
Arts and Entertainment	76.1%	22.3%	69.3%	**76.5%**	75.9%
State and Local Government Administration	65.7%	19.6%	50.3%	70.8%	**71.1%**
Weather and Natural Disasters	77.7%	59.8%	62.7%	**83.8%**	80.8%
Fires	35.0%	21.1%	21.1%	45.8%	**50.8%**
Sports and Recreation	87.3%	66.5%	75.7%	**88.3%**	84.9%
Death Notices	36.4%	-	3.6%	37.0%	**44.0%**
Churches and Religion	61.8%	26.6%	61.0%	66.4%	**67.9%**
Other	18.0%	2.1%	-	19.0%	**22.2%**

Source: New York Times front page data, 1996-2006 (Boydstun), New York Times front page data, 2007-2023 (Broad), tweeted articles collected using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023). N = 46,499.

2.5.3 Climate change classification model comparison

Figure 2.3 includes the results of the climate change classification models by method, starting with keyword classification. One-hot representations of major topic predictions were used as additional predictors in the SVM models to predict climate change stories. Researchers sometimes report the overall accuracy (the number correctly classified documents divided by the total number of documents) for climate change story classification. This metric of model performance is inappropriate when working with highly imbalanced data, like stories on climate change. Because so few of the stories in the training data are manually classified as climate change (just one percent), most any classification method will have exceedingly high overall accuracy: a method that predicts that *no* stories were primarily about climate change will have an overall accuracy score of about 99 percent. Balanced accuracy is a similarly deceptively scoring metric for highly imbalanced data: this score takes the mean of the recall and the true negative rate. As with overall accuracy, the majority class (non-climate change stories) make up 99 percent of the data and so the true negative rate inflates the apparent performance of Boolean classification.

Instead, it is more appropriate to review the F1 score, precision and recall. The precision score indicates the percentage of the stories classified as climate change that were actually climate change stories. Recall indicates the percentage of climate change stories that were correctly classified as such. The results indicate that Boolean classification is wrong about two-third of the time. Raising the bar to require two keyword hits for Boolean classification yields results that are wrong about half of the time. Boolean classification performs fairly well in terms of recall, but on balance with its precision score indicates that it casts too wide a net and retrieves far too many false positives. The F1 score represents the harmonic mean of the precision and recall scores. Aside from the maximum LDA feature method, Boolean classification had the worst performance in the task of climate change classification.

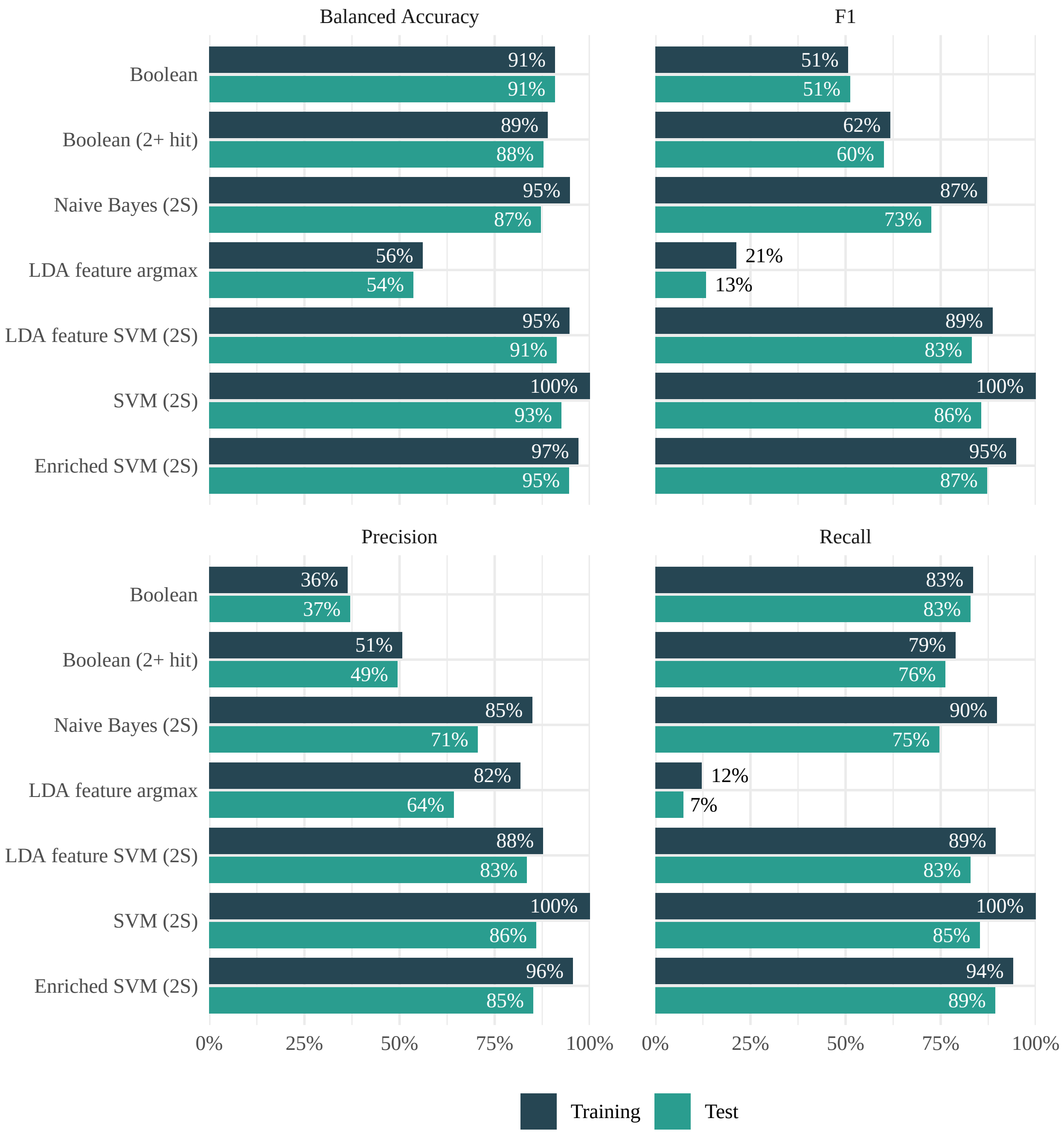

Figure 2.3: Comparison of climate change classification model performance (accuracy)

Note: Figure 2.3 illustrates various models' performance identifying whether news stories are primarily about climate change.
Source: New York Times front page data, 1996-2006 (Boydstun), New York Times front page data, 2007-2023 (Broad), tweeted articles collected using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023). N = 46,499.

The Enriched SVM model had the highest performance, with an 87 percent out of sample F1 score. The argmax LDA feature classification models performed poorly, but showed a higher true positive rate than Boolean classification. Recall for this approach was exceedingly low, indicating that the estimated topic distribution of climate change articles tends to have greater portions of content assigned to topics other than climate change (such as energy sources or even the environment, generally). However, using the guided LDA topic distributions as features in an SVM model performed extremely well, with F1 scores exceeding Naive Bayes and approaching the SVM and enriched SVM F1 scores (again indicating the high validity of the Guided LDA features).

2.6 Discussion

The central focus of this chapter was to demonstrate the limitations of classification and data selection methods that are standard in the climate change communications literature. Nearly all research in this field relied on the same methods that inflate news media estimates attention to climate change. Based on the results presented above, past research using Boolean classification may have overestimated news media attention to climate change by a factor of two to three. Given the relationship between news media, public and policy agendas, inflating news media attention to climate change may give a more hopeful outlook than deserved for the potential for policy action on climate change.

In this chapter I also provided a brief introduction to an experimental guided topic model. This model was trained on data outside the classification training data, using prior information about the Comparative Agenda Project topic scheme and terms that were frequently associated with each of the minor topics in the NYTFP (1996-2006) data. The moderately high classification results using guided LDA features are promising for the potential of this experimental method to apply broadly to research in this field. The validity of these features suggest that, unlike LDA or sLDA, any researcher might use the term-topic dictionary constructed for this task (in Appendix B), replicate the topic model structure on their own text data and compare results directly with those presented in Chapter 4.

In the next chapter, I estimate attention to climate change using the guided LDA-enriched SVM model and perform a bootstrap analysis to evaluate attention to climate change while accounting for potential measurement error. As discussed in the next chapter, not only does the Boolean classification method inflate measures of attention to climate change but does so at an increasing rate as climate change becomes more frequently mentioned in stories on other policy topics or events.

CHAPTER 3

ATTENTION TO CLIMATE CHANGE

3.1 Introduction

Our information environment shapes our political realities (Cacciatore *et al.*, 2016, p. 14; Takeshita, 2006; Zaller, 1992). The issues that we prioritize and the way we respond to real-world events is in large part driven by the information we access.[1] The distribution of news through social media websites has enabled the public to curate their own information environments and audiences are gradually choosing to consume news from niche online sources (Hollander, 2008; Stroud, 2011). This chapter builds on previous political communications and climate change communications research by directly testing whether front page news coverage of climate change is representative of news media attention to climate change on social media (Twitter, specifically).

As discussed in detail in the previous chapter, researchers tend to focus analyses on traditional or "prestige" media, such as the New York Times and the Wall Street Journal. Based on past research, traditional print news media sources tends to be representative of news media across sources and domains (print and television, for example). However, with increasing opportunities for media consumers to select niche news sources, it is all the more important to examine whether attention differs significantly across news sources and domains. Here I compare the New York Times front page and the @nytimes Twitter timeline major topic attention distribution and climate change attention over time. Contrary to expectations, Monte Carlo simulations reveal that climate change attention does not substantially differ

[1]Note: The relationship between media attention and the public's issue priorities is often measured using "Most Important Problem" (MIP) survey data. In these surveys, respondents provide open-ended responses to the prompt: "What is the most important problem facing the country today?" Open-ended responses are then coded into categories. Most Important Problem data tend to closely mirror news media agendas. Unfortunately, responses related to climate change in MIP data are categorized under a broader "environment and climate change" category, making direct comparison of climate change attention and public priorities a difficult task.

between the New York Times front page and the @nytimes timeline. However, the count of @nytimes tweets linking to stories about climate change substantially exceeds the number of climate change stories on the front page of the New York Times, which is relevant for analyses of attention diversity in Chapter 4.

This chapter also improves upon past research in this field by examining attention to climate change across a diverse set of news sources in addition to the New York Times front page. I compare levels of climate change attention across these sources over time, which vary in perceived ideological lean and intended readership. Consistent with prior research, news sources that are perceived as left-leaning sources have higher rates of climate change attention than perceived right-leaning sources. However, not all left-leaning sources have higher rates of climate change attention: contrary to expectations, climate change attention levels are lower for left-leaning Black- and Hispanic-owned news sources.

Expanding on the analyses of text classification model performance in the previous chapter, I examine the full extent of bias in Boolean classification estimates of news media attention to the issue of climate change. As noted in Chapter 2, nearly all researchers in climate change communications literature measure attention in terms of keyword “hits”: the percentage of news stories mentioning phrases such “climate change,” “global warming”, and “greenhouse gases,” divided by all news stories in the corpus. Conducting a mixed effects model, I demonstrate that Boolean classification not only overestimates attention to this issue but amplifies the rate at which attention to climate change appears to be growing. In addition, using a more robust classification method enables additional analyses of attention to climate change as well as secondary attention to climate change (i.e., the permeation of considerations related to climate change into other topics). Attention to climate change as a secondary issue association is analyzed more closely in Chapter 4.

Finally, the novel data used in these chapters improves upon a tendency in climate change communication literature to calculate climate change attention as the number of stories with climate change keyword hits over all published stories. This approach ignores the importance of the front-page news generating process in which issues compete for finite agenda space (McCombs & Zhu, 1995); it also ignores the impact that front-page news can have on public and government policy priorities, for example (Boydstun, 2013). The data analyzed in this chapter includes stories tweeted by various news sources; news media sources do not tweet every article published on their website (or paper) but instead tweet a subset of articles that are presumed to be considered newsworthy (or tweet-worthy) articles. Additionally, news sources often tweet the same story more than once. The data do not remove stories tweeted more than once, as duplication in story distribution is an important indicator of emphasis. Thus, using stories distributed via Twitter grants the opportunity examine news media sources issue priorities under a set of constraints alternative to the spatial constraints of front page news coverage.

3.2 Past research

The International Government Panel on Climate Change (IPCC) estimates that unmitigated climate change may rise to the level of existential threat (Climate Change, 2022). What might explain the low levels of attention it receives in U.S. news media despite the severity of the issue? First, there have been substantial efforts by moneyed interests to obfuscate the issue of climate change and propagate doubt around the veracity, causes or impacts of climate change (Bell & York, 2010; Dunlap, 2013; Elsasser & Dunlap, 2013; Oreskes & Conway, 2010). The doubt surrounding the issue likely prevented journalists from seriously engaging with the issue, or, as noted in the introduction, peddle dubious alternatives to greenhouse gases as primary contributors to climate change (e.g., solar storms).

Several studies indicate that attention to climate change rose substantially in the late 2000s but the increased attention to climate change was news media coverage of the debate about the veracity of climate science (Boykoff, 2011). Through coordinated efforts, climate change counter-movement (CCCM) coalitions produced news media and bogus scientific articles meant to stoke mass skepticism and denial of the existence of climate change (Dunlap

& Jacques, 2013; Elsasser & Dunlap, 2013). Some argue that these efforts directly led to federal climate change policy stagnation (e.g., McCright & Dunlap, 2003). Recent research suggests that funding for CCCM coalitions or "merchants of doubt" peaked in the 2000s and their influence tapered off during the 2010s (Brulle, 2019). Yet this research has shown that even during the CCCMs' relatively quiet period, climate change attention was remained extremely low on American and international media agendas.

Low media attention to climate change is best explained by its position in the later stages of the Downsian issue attention cycle (Downs, 1972). The Downsian issue attention cycle framework conceptualizes five stages of issue attention: the pre-problem stage, in which experts may be aware of the issue but it is not a broad public concern; the alarmed discovery phase depicted by a public fervor to address the issue, followed by realization of the costs to address the issue and then a decline in public interest. Finally, the post-problem stage is characterized by lingering attention and occasional spikes of attention following key developments related to the issue (Boydstun *et al.*, 2014b).

The issue of climate change is well beyond the alarmed discovery phase and well into the final post-problem stage. The issue is now best described as a "creeping" issue, as the severity of the issue is increasing very gradually. Creeping issues have a difficult time garnering attention as shocking events that demand public and policymaker attention are quite rare. Additionally, there is little public demand for attention to climate change: U.S. residents who are highly concerned about climate change make up a small minority of the population (Leiserowitz *et al.*, 2021). What's more, the public benefits from a consumerist system reliant on high greenhouse gas emitting industries, and many solutions to the problem may require substantial buy-in regarding the public's consumer and environmental behaviors (Dunlap & Liere, 1984; Jacques, 2012; McComas & Shanahan, 1999; Shanahan, 1996).

Despite low overall attention to climate change, there may be instances of "media storm" coverage of the issue (Walgrave & Hardy, 2017). Media storms are characterized by sudden spikes in media attention to an issue followed by sustained coverage for a period of time (operationalized as a 150 percent increase in attention followed by 20 percent attention for seven days in Boydstun *et al.*, 2014b). Media storms are significant as they may be helpful or even necessary to overcome "institutional friction," the political or institutional dynamics that slow or prevent policy reform (Jones *et al.*, 2009; Walgrave & Vliegenthart, 2010).

Yet, media storms are very rarely studied in the context of climate change attention, especially for U.S. coverage of the issue. This may be due to climate change coverage failing to meet the criteria for media storms, except on very rare instances. For example, a single climate change media storm was identified in 2009 following a reported error in an IPCC report (Hajer, 2012). However, spikes in climate change coverage have been observed and causal determinants for these spikes have been identified in past research. Analyses below review the prevalence of media storms (or "semi-storms") related to climate change according to New York Times front page coverage, stories tweeted by several news sources, as well as a public engagement-weighted approach to analyzing media storms.

Climate change attention is driven primarily by real-world, non-weather-related "focusing" events. In particular, the media tends to respond to high-profile international events such as conferences, activists, and especially violence. Domestic and international climate change attention is primarily driven by political actors acting, particularly developments regarding international climate change agreements like the Kyoto Protocol and meetings like the Conference of Parties (Liu *et al.*, 2011; Schäfer *et al.*, 2014). Other research has found that issues receive more attention when illegal activity or violence involved (Feber *et al.*, 2017) as well as celebrity involvement (Feber *et al.*, 2017; Lawrence & Boydstun, 2017). The media appear not to increase coverage of climate change in response to weather related events or major scientific publications on the issue (Schäfer *et al.*, 2014) with the exception of the "climategate" conspiracy (Hajer, 2012; Leiserowitz *et al.*, 2013).

Attention is also determined by the perceived ideological lean of news sources. News media focuses on issues important to their intended or expected audience, climate change is highly polarized and polarizing issue (Dunlap *et al.*, 2001; Egan & Mullin, 2017). It is not unexpected, therefore, that the perceived ideological orientation of a news source (or that of its intended readership) would influence its coverage of this issue (Bohr, 2020; Saunders *et al.*, 2018). These findings are largely corroborated below, although this it appears to true only for "prestige" sources and did not hold for certain "new" media sources that may be perceived as left-leaning.

The research presented in this chapter accepts past research into the causal mechanisms of climate change attention but questions whether the level of attention in past research was correct, give the issues with data collection and classification discussed in Chapter 2. The remainder of this chapter examines levels of climate change attention (and other major topic attention) using a higher accuracy classifier relative to the industry standard, Boolean classification. As discussed in greater detail below, past research has combined both primary attention to climate change as well as secondary attention to climate change ("permeation" or "diffusion" of climate change into other topics as a relevant consideration). The topic of climate change diffusion is explored in greater detail in Chapter 4.

3.3 Theory

There is an ongoing debate about the agenda-setting role of online and print media (Althaus & Tewksbury, 2002; Heim, 2013; Meraz, 2011; Sweetser & Wanta, 2008) and the reciprocal effects that these domains have on each other (Vargo & Guo, 2017). Whether inter-media agenda-setting effects are driven by online or traditional media, research consistently finds that agendas correlate strongly in the aggregate across domains. Yet, despite similarities in agendas in the aggregate, we might expect important differences between domains at the topic level (Stern *et al.*, 2020). I expect climate change to receive a greater share of attention online relative to front page news due to the combination of two phenomena: 1) there are fewer constraints for online news coverage that front page news coverage; and 2) climate change is a creeping issue with relatively fewer focusing events compared to other topics.

Highly correlating agendas may be a product of alarm-driven media: the high levels of topic agenda correlations observed in past research are possibly driven by attention to a few topics prone to abrupt developments and real-world events. Front page news production is best explained by the alarm/patrol hybrid model, wherein media attention “lurches” from one topic to another, following “alarms” or breaking developments; attention to a topic may be sustained for a period during which news media patrols the story for developments and in-depth analysis (Boydstun & Russell, 2016). Under the hybrid alarm/patrol model, we should expect traditional news media sources to focus on the same or similar stories that have gripped the nation’s attention (or are likely to grip the nation’s attention).

Attention lurching from major developments or events may be driving strong inter-media topic correlations but observed correlations may mask or distract from meaningful differences at the topic level. It is rare for “creeping” and fringe issues that do not typically have abrupt attention-grabbing events to make front page news. The Earth is warming very quickly on a geological scale but very slowly on a human time scale and so it is difficult to grab the public’s attention and sustain it for any period of time. However, I expect there to be proportionately higher patrol coverage of climate change stories between the rare instances when climate change does make front page news. Analyzing stories distributed via Twitter allows an opportunity to estimate attention in between alarms.

While existing climate change communications literature tends to analyze either front page news, all published stories from a source, or (in some cases) social media posts, this report makes a novel contribution to the literature by analyzing the content of stories that were distributed via Twitter, treating the platform as the “front page of the internet.”[2] This report makes a novel contribution by examining news data distributed via Twitter, rather than examining the contents of tweets alone (e.g., Barberá *et al.*, 2019; Jang & Hart, 2015). Twitter is treated as a medium through which news is distributed, as it was quickly treated by major news sources shortly after its launch in 2006. Twitter’s potential for academic research

[2]Twitter’s role as the front page of the internet was undermined upon NPR’s exit from the platform following Musk’s attempt to label the organization as state-sponsored media. Around the same time, alternative platforms providing similar services (Mastodon and Bluesky) and creation of platforms for niche users (like Truth Social). Future research might evaluate data from these platforms, instead.

of public opinion was identified almost immediately (e.g., Java *et al.*, 2007) and the platform was examined as a platform for news distribution as early as 2010 (Kwak *et al.*, 2010).

This report treats Twitter as a platform for news distribution by following the hyperlink of every tweet to the tweeted story to retrieve and examine story content, a novel approach in climate change communications literature. Using this method of data collection, the spatial constraints of the physical front page of the newspaper are relaxed (but not eliminated): the number of stories tweeted is not a function of physical space on a newspaper (although news sources may have been limited in other ways, such as a cap on the number of daily tweets, which has shifted over time). News sources might also limit tweets desire to avoid follow loss due over-tweeting and timeline flooding.

Using Twitter timelines also has the added benefit of source-weighted emphasis on stories, as sources often tweeted links to the same story multiple times per week to amplify story exposure. Access to Twitter data also enables the potential for weighting story emphasis by public engagement (analyzed below). Twitter timelines, like front page news, signal news sources' issue priorities with spatial constraints relaxed. This is unlike the industry standard of purchasing all published stories, which does not convey story emphasis or priority. Again, not all published stories are tweeted ad so Twitter timelines provide a stronger signal of sources' issue priorities than collecting all stories published from a source, as through a third-party news hosting service.

In sum, shifting from front page stories to tweeted stories comes at the cost of a more opaque signal on source priorities: tweeted stories are not front page news and do not share the same spatial constraints as the front page in printed media. However, tweeted stories arguable send a stronger signal on issue priorities when compared to all stories published by a source. Additionally, analysis of tweeted stories may be appropriate when the research focus is on issues that do not frequently appear on the front page. The effect of this trade-off between signal clarity and breadth of articles analyzed on attention to climate change can be measured directly by comparing the attention to this issue on the front page of New York

Times and the @nytimes Twitter timeline. Given that climate change is considered a fringe issue and has received little attention on the front page in past research, we might expect relaxed spatial constraints to lead to greater attention to climate change.

Hypothesis 3.1: Climate change attention is higher on Twitter than front page news

Past research has consistently shown that news media attention to climate change is increasing in the U.S. and internationally. However, if attention in past research is measured using Boolean classification, observed increases in climate change attention levels may be attributable to the diffusion of climate change into other topic considerations, rather than attention to issue of climate change itself. While the previous chapter showed a high false positive rate of climate change classification when using Boolean hits, it is unclear if differences between this and more sophisticated classification methods are meaningfully distinct over time.

However, it has been theorized (Schattschneider, 1960) and empirically supported (Boydstun, 2013) that issues are likely to receive news media attention the more "angles are at play." As recent IPCC reports have documented a widening scope of both observed and anticipated impacts of climate change, it is reasonable to anticipate a corresponding increase in news media coverage of climate change over time.

Hypothesis 3.2a: News media attention to climate change increased between 2007 and 2023

As discussed in the previous chapter, using "climate change" hits in article keywords, headlines and content to measure climate change attention may increase the prevalence of false positives, as climate-change-related keywords can be mentioned in articles focused primarily on other topics. This manuscript analyzes attention to climate change according to primary topics. Similar to social scientists employing machine learning text classification techniques (Barberá & al., 2019; Hillard *et al.*, 2007; Purpura *et al.*, 2008; e.g., Purpura & Hillard, 2006; Silla & Freitas, 2011; Verberne *et al.*, 2014), each article's primary topic is predicted using machine learning techniques that identify both climate change and non-climate-change articles, covered extensively in the first chapter.

As shown in the previous chapter, classifying stories as belonging to a topic using "hits" inflates the number of both false positives and false negatives (King *et al.*, 2017). The enriched SVM model was shown to have much higher precision and recall compared to Boolean classification. However, the previous chapter did not show whether the upward bias is a difference that makes a difference. Climate change attention even measured with an upward bias is typically very low; it is possible that the difference in methods is insignificant and negligible. To the contrary, I expand on the previous chapter by illustrating that the upward bias from Boolean classification is indeed statistically significant.

Hypothesis 3.2b: Boolean classification significantly inflates estimated attention to climate change

Beyond inflating levels of attention, I expect the Boolean classification to inflate the rate at which attention for climate change appears to have grown over time. Past research measuring attention as Boolean hits has shown that climate change mentions have increased over time, with some spikes in mentions surrounding international policy discussions (e.g., the Kyoto Protocol, United Nations Climate Change Conference) and domestic and international weather events and natural disasters.

Returning to Construal Theory, attention to climate change is expected to increase as the costs of climate change appear to become nearer in proximity (spatial or temporal). The severity of extreme weather events is beginning to clarify the potential costs of climate change (Clarke *et al.*, 2022). I also suspect issue-proximity to have a role in secondary attention to climate change. The costs of climate change are predicted to be pervasive: affecting immigration (Peluso & Harwell, 2001), agriculture (Ortiz-Bobea *et al.*, 2018), economies (Kompas *et al.*, 2018), housing policy (Krayenhoff *et al.*, 2018), health (Mitchell *et al.*, 2016), racial justice (Schlosberg & Collins, 2014) and other policy topics. As the costs of climate change become clearer, I expect mentions of climate change keywords to appear in non-climate change stories more frequently over time.

Hypothesis 3.2c: Boolean classification inflates estimated rates of climate change attention growth

Based on past research, I expect the levels of climate change attention to differ between sources: perceived left-leaning sources (e.g., New York Times and National Public Radio) will tend to have higher levels of climate change attention relative to perceived right-leaning sources (e.g., Wall Street Journal and Daily Wire) (Bohr, 2020; Saunders *et al.*, 2018; Schäfer *et al.*, 2014). There is a tendency for news media to focus attention to topics that appeal to the ideological leanings of their audiences (Budak *et al.*, 2016; Druckman, 2014). Additionally, issue attention reflects policy positions in other forms of political communication; for example, political party manifestos more frequently mentioning environmental terms also tend to align with left-leaning environmental policies (Budge, 2002). Given that public opinion on the issue of climate change has polarized substantially since the 1970s (McCright & Dunlap, 2011; Turner, 2009), is now highly crystallized (Egan & Mullin, 2017), and is a polarizing topic (Feldman & Leiserowitz, 2014), I anticipate that climate change attention levels will differ by news sources' perceived ideological lean.

Hypothesis 3.3a: Climate change attention is higher among left-leaning sources than right-leaning sources

Additionally, I expect issue public news media sources to give higher attention to climate change when it is directly relevant to the source's intended audience or highest priority policy. I anticipate that due to the rise of race-related focusing events in the late 2010s, such as incidents of police brutality and the emergence of the Black Lives Matter movement, there would be an increased likelihood of discussing other topics, like climate change, in the context of racial justice. Research about racial disparities regarding climate change outcomes (Schlosberg & Collins, 2014) and extreme weather impacts (Zanocco *et al.*, 2022) was and distributed on Twitter (e.g., Lambert, 2019) in the late 2010s. I expect that academic research and investigative journalism emphasizing racial disparities in the effects of climate change would, in turn, drive up climate change attention among news sources whose intended audiences are primarily people of color and potentially issue publics for issues related to social justice and racial equity. I explore this hypothesis in greater depth in Chapter 4.

Hypothesis 3.3b: Attention to climate change is increasing more recently and rapidly among Black- and Hispanic-owned news sources

3.4 Data

The data analyzed in this chapter include the full set of stories tweeted between June 2007 and April 2023 as well as stories that appeared on the New York Times front page between January 1996 and April 2023. All New York Times front page stories published between 1996 and 2006 were manually coded according to their primary major topics and subtopics according to the Comparative Agendas Project topic scheme (Boydstun, 2014); these data were supplemented with 2007-2023 front page data retrieved from the New York Times developer API and supplemented with a custom data wrangling program described in the previous chapter (and discussed in detail in Appendix A). About 20,000 stories from the supplemental data were manually coded using the same topic scheme.

A series of custom programs were written to collect about 1.4 million tweets with hyperlinks from 2007 through 2023. About 300,000 of these stories were removed from the data set for inactive links that could not be retrieved via Wayback Machine, for linking to an alternative news source or domain, or linking to multimedia without text content to analyze (e.g., Daily Wire frequently tweeted YouTube videos, which were removed from this analysis). Also removed were tweets linking to strictly self-promotional materials (e.g., @nytimes frequently tweets deals for its other subscription services, such as Wirecutter, the NYT Crossword, and NYT Cooking).

The novel data set analyzed in this chapter following screening procedures consists of about 1.1 million articles tweeted by various news media content creators (New York Times, Wall Street Journal, NPR, Atlanta Black Star, The Root, Daily Wire, Christianity Today and

Inside Climate News) from June 2007 through April 2023. These sources represent a more diverse set of news media content creators in terms of intended audience, which is appropriate for the analysis of niche news coverage of "fringe" issues like climate change. The data also include stories from La Opinion, a Spanish-language news source based in Los Angeles, CA. Stories tweeted by La Opinion were translated using Google Translate for English content in order to apply SVM and LDA models, which were trained on English-language content. Google Translate has been shown to be highly reliable for the purpose of translation prior to automated text classification and content analysis (Reber, 2019).

3.5 Methods

In this chapter, attention to climate change in American news media sources is measured as the proportion of tweeted news stories classified as being primarily about climate change divided by all tweeted stories about climate change, by news source. Stories were classified as climate change using the Support Vector Machine enriched with guided Latent Dirichlet Allocation features described in the previous chapter. I provide a brief review of these methods here.

Additionally, linear trends in attention to climate change were estimated using mixed effects models. Attention as SVM classification is compared against levels of attention measured as Boolean classification, Boolean 2+ hit classification to evaluate differences in trends across different classification methods. Below I also review the procedure to estimate confidence intervals around climate change attention estimates using Monte Carlo simulation.

3.5.1 Guided Latent Dirchlet Allocation

The topic distribution for each story was estimated using a guided Latent Dirichlet Allocation (LDA) topic model discussed in the previous chapter. The guided LDA model was trained on a random selection of about 230,000 observations from the database of New York Times front page stories and Twitter timelines data. A term-topic probability table was constructed using prior information about highly informative terms per minor topic the Comparative Agendas Project topic scheme. (See the term-topic table in Appendix B). The guided topic model uses prior information about terms' centrality to topics as initialization parameters (rather than random initialization parameters). This approach enables researchers to structure the topic model output, labeling latent topics prior to their formation rather than in an ad-hoc fashion after the model is fit (Jones *et al.*, 2021).

The average of 100 posterior topic distributions per story were used predicting out of sample data using an enriched Support Vector Machine classification model. The guided LDA modeled topics at the minor topic level; following Eshima *et al.* (2020), the model's topic features were summed according to their theoretical hierarchy in the CAP coding scheme (e.g., all minor topics under the *Macroeconomics* major topic were summed to estimate portions of stories that might be assigned to the *Macroeconomics* major topic). Topic distributions are used only for the purpose of classification in this chapter, as in Dogru & al. (2021), Inkpen & Razavi (2014) and Chen & Li (2016); however, topic distributions are examined more closely in Chapter 4 in discussions of topic associations and association diversity.

3.5.2 Enriched Support Vector Machine

The major topic was predicted for each story that was not already manually coded in the training data. The out-of-sample major topic classification accuracy overall was about 78 percent, similar to the predictive accuracy of other SVM models used to predict major topics using the CAP coding scheme (Hillard *et al.*, 2007, 2008; Purpura *et al.*, 2008; Purpura & Hillard, 2006). The distribution of major topic predictions is reviewed in the results section below. In Chapter 4, major topic distributions are reviewed more closely while taking into account the conditional probability of alternative major topic assignment.

Major topic predictions were also used as one-hot-representations in a two-stage enriched SVM to predict whether a story was primarily about climate change. The same pre-processing procedures that were applied to the training data were applied to the full 1.1 million-row database in order to estimate LDA topic features and apply the SVM model for climate change classification. As demonstrated in Chapter 2, the two-stage enriched SVM enriched with guided LDA features had the best out-of-sample performance when prediction climate change stories, with out-of-sample accuracy (F1) of about 87 percent. The processing and scaling procedures applied to the training data were applied to the hold out test set, as well as the full database examined in this chapter and the next.

3.5.3 Monte Carlo simulation

The enriched SVM classification model was high-performing but did not achieve perfect accuracy when predicting climate change stories on out-of-sample data. Confidence intervals around levels of climate change attention were estimated using Monte Carlo simulation to account for this classification error. This methodology proves beneficial for exploring climate change attention rates under different data assumptions while assessing the enriched SVM model's robustness and reliability.

The simulation takes into account the false positive rate (0.11) as well as false negative rates conditional on the presence of terms from the standard set of keywords used in Boolean classification ("climate change," "global warming," and "greenhouse gas"). False positive rates were 0.07 when one or more of the standard climate change key terms appears in the text two or more times, 0.03 when a key term appeared just once, and 1×10^{-6} when no climate change key terms appeared in the story. Two thousand Monte Carlo simulations were run for each story, taking these false positive and false negative rates into account.

For each of the simulation analyses below, a 90 percent highest posterior density interval (HPDI) was estimated from the simulated data. These simulations and their confidence intervals provide a more realistic look at climate change attention estimates while eliminating the false positive bias of Boolean classification and taking potential error into account.

3.5.4 Mixed effects model

Mixed effects models are performed to measure statistical differences in trends in attention to climate change by source and by method of classification. Mixed effects models include both fixed- and random-effects in the estimation of an outcome variable (Bates *et al.*, 2014). Mixed effects models are especially well suited to measure long-term trends in longitudinal data by treating time as a fixed effect. These models are also especially adept at handling repeated measures with potential autocorrelation, such as daily measures of attention to certain news topics per news source. For this analysis, source-specific intercepts and slopes are treated as random effects in order to analyze the effect of classification method on estimates of attention (Bartels, 2008; Bates *et al.*, 2014; Bell & Jones, 2015).

To test for statistical differences in climate change attention (*Attention*) under different classification methods, I compare three mixed effects models. I consider a baseline model with only one fixed effect, *Method* (one hit or more, two hits or more, and enriched SVM) as well as random intercepts by source, *Source*. Model 2 estimates the fixed effects of date (*Date*, centered and scaled) in addition to *Method*; finally, Model 3 includes interacting *Date* and *Method* fixed effects. Models 2 and 3 also allow for variance in *Date* slopes and intercepts by *Source*. The fixed effects for these models are presented in Table 3.1 in the Results section below. The specification for the interaction model is:

$$\text{Attention} = \beta_0 + \beta_1\text{Method} + \beta_2\text{Date} + \beta_3(\text{Method} \times \text{Date}) + u_{\text{ Date } \mid \text{ Source}} + \epsilon$$

To test for statistically meaningful differences in climate change attention across sources, I fit three additional mixed effects model evaluating variation of attention over time by source. To account for variation in how media sources respond to real-world events, all three models include random effects for sources' ideological *Lean* by year.[3] The first model considers only *Date* as a fixed effect to estimate the overall trend in climate change attention, allowing some

[3]"Left" for the New York Times front page, @nytimes and @NPR, "right" for @WSJ and @realDailyWire, and "other" for the other sources

year variation by ideological lean. The second model examines *Date* along with *Source*-specific intercepts. The third model fits an interaction between *Date* and *Source*, allowing overall long-term trends to differ by source. Below is the model specification for the full interaction model. Fixed effects for this model and the models described above are presented in Table 3.2 in the Results section below.

$$\text{Attention} = \beta_0 + \beta_1\text{Date} + \beta_2\text{Source} + \beta_3(\text{Date} \times \text{Source}) + u_{\text{Lean} \mid \text{Year}} + \epsilon$$

3.6 Results

This section reviews the output of SVM major topic classification and enriched SVM climate change classifications on the full 1.1 million-row database of stories tweeted between 2007 and 2023. First, is a comparison of the major topic agendas of the New York Times front page (1996 through 2023) compared to the @nytimes Twitter feed (2007 through 2023), followed by a comparison of New York Times front page coverage of climate change stories compared to @nytimes Twitter feed attention to climate change. I also compare overall major topic agendas by source and examine trends in climate change attention by source and classification method between 2007 and 2023.[4]

This section also provides an evaluation of climate change media storms (or "semi-storms" with relaxed storm criteria) and examines the prevalence of storms when weighting attention by public engagement on Twitter. I end this section with a review of the mixed effects models described above, testing the statistical significance of different climate change classification methods and differences in sources' attention to climate change.

[4]Note that major topic emphasis for each source includes data from the time the sources began using Twitter to distribute stories. Thus, time periods differ per source. However, source's major topic agendas are fairly constant over time on Twitter.

3.6.1 Major topic agendas

Figure 3.1 shows the monthly share of attention to major topics. The vertical dashed line at January 1, 2007, in Figure 3.1 indicates the starting point for supplemental New York Times front page and twitter data collection, prediction and content analysis. To make this figure legible, it includes only the seven major topics receiving the greatest amount attention from both the NYTFP and @nytimes combined during the 1996 to 2023 time period. All other topics were aggregated into a single "All other topics" category. Note that the major topics in the 1996 to 2006 data were manually coded according to primary topic. A sample of the data from the 2007-2023 period were manually coded to the same CAP coding scheme while the remainder were predicted using the highest-performing major topic classification model from Chapter 2 (SVM).

Figure 3.1 lends credence to the external validity of the SVM major topic classification model. There is not an immediate upset in the distribution of major topic attention from the end of manually coded data in 2006 to SVM predictions in 2007 and there are multiple identifiable trends that comport with prior expectations and lived experience to support its long-term efficacy. Of note are the spikes in attention to the *Health* topic corresponding with the outbreak of COVID-19 in early 2020 and subsequent spikes in attention to this topic that correspond with spikes in infection rates. There is also a swell in attention to *Health* in 2009 corresponding with debates and passage of the Affordable Care Act. The *Government Operations* major topic makes up a larger share of both front page and Twitter timeline attention following Donald Trump announcing his candidacy for office in the 2016 presidential election. There is also a slight increase in the size of the *Banking, Finance, and Domestic Commerce* major topic, which aligns with the fall of the housing market in fall of 2008 and the Great Recession lingering through 2009.

Figure 3.1: Major topic agendas (NYT front page and tweeted stories, 1996-2023)

Note: Figure 3.1 illustrates the number New York Times front page stories and @nytimes tweets containing valid links to nytimes.com stories aggregated by year. Valid links direct users to stories on nytimes.com that are active or accessible via Wayback Machine and were not identified as an advertisement, self-promotion, or multimedia.
Source: New York Times front page data set, 1996-2006 (Boydstun); data collection using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023).

While not all New York Times front page stories are tweeted and not all tweeted stories are published on the front page, the two domains share highly similar major topic agendas (cor = 0.8253). There are noticeable spikes in attention on the front page that are less apparent in the Twitter data. As described above, news media attention lurches from one topic to another, chasing breaking developments. This "lurching" behavior is more apparent in the New York Times front page data, where attention to these major topics varies widely from month to month; in contrast, the major topic attention distribution appears much more stable from month-to-month in the Twitter data.

Major topic agendas are quite similar overall but, as mentioned above, a high correlation of major topic agendas overall does not necessitate that *all* topics are highly correlated across domains. As is clearly illustrated in Figure 3.1, *Arts and Entertainment* and *Sports* account for a much higher percentage of the @nytimes Twitter timeline coverage than front page news coverage despite the high correlation value overall. As discussed above, topics compete for attention and there are greater constraints on how many topics can receive front page coverage relative to Twitter timeline coverage. Twitter timelines, like front page news, signal news sources' issue priorities with spatial constraints relaxed; we might expect to see more coverage of important but creeping issues like climate change on Twitter rather than the front page.

3.6.2 Climate change attention by domain

With physical spatial constraints removed, I hypothesized that niche issues like climate change will comprise a larger share of the New York Times policy agenda on Twitter compared to the New York Times front page. This hypothesis was not supported in the data. Climate change attention was aggregated by week for a detailed comparison of front page and Twitter timeline attention; Figure 3.3, however, illustrates the simulated *monthly* rate of climate change attention on the New York Times front page and New York Times Twitter feed to make trends interpretable. Monthly attention to climate change is also presented as the rolling three-month average to emphasize the similarities in quarterly trends between the New York Times front page news and @nytimes Twitter feed.

H3.1: Climate change attention is higher on twitter than front page news. **(Not supported)**

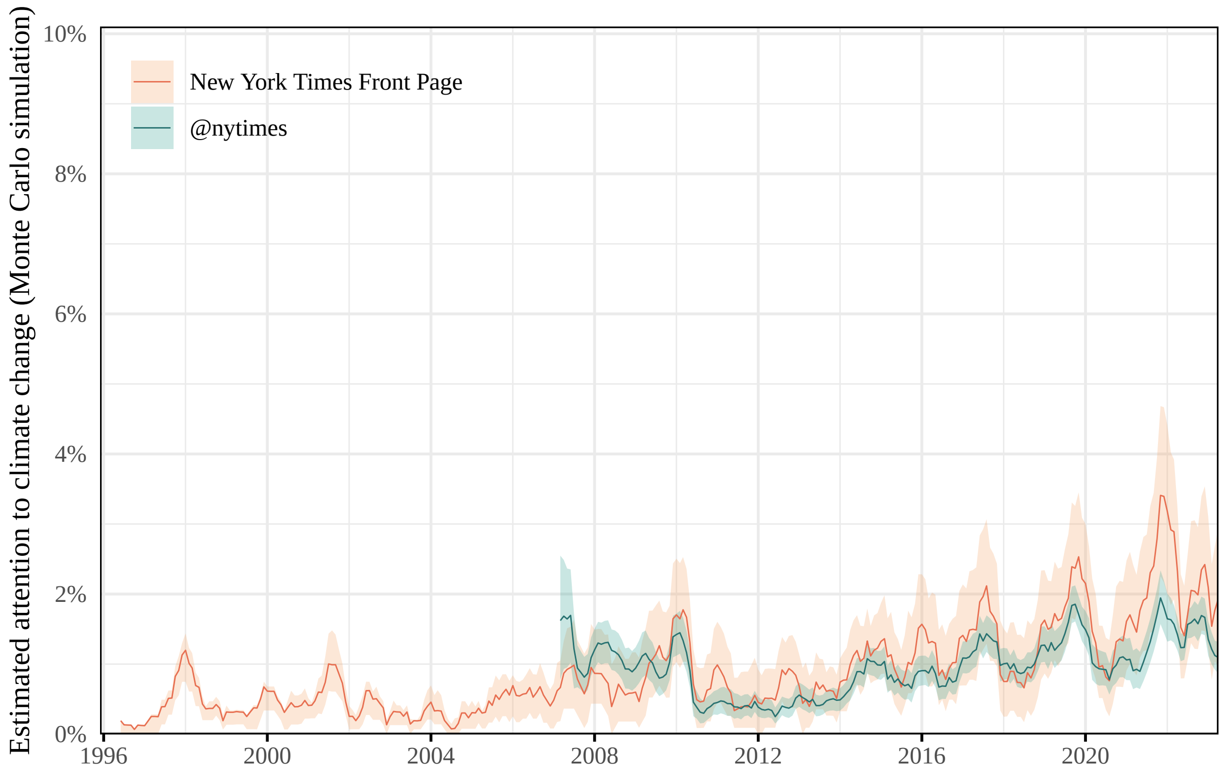

Figure 3.2: Attention to climate change in front page versus Twitter

Note: Figure 3.1 illustrates the number New York Times front page stories and @nytimes tweets containing valid links to nytimes.com stories aggregated by year. Valid links direct users to stories on nytimes.com that are active or accessible via Wayback Machine and were not identified as an advertisement, self-promotion, or multimedia.
Source: New York Times front page data set, 1996-2006 (Boydstun); data collection using various APIs and webcrawling with Python, Selenium, R, and Rvest (2007-2023).

A statistical model is not needed to illustrate null results here: there was just one case where there 90 percent HPDI of simulated weekly attention to climate change did not overlap when comparing front page news and the @nytimes timeline (the week of November 8, 2021). In that case, front page coverage of climate change was higher than Twitter timeline attention to climate change. Additionally, Twitter timeline attention to climate change was slightly lower than front page coverage of climate change for almost the full 2007 to 2023 period (although the estimate for each domain was within the 90 percent HPDI of the other).

Having compared major topic agendas and attention to climate change on the New York Times front page the @nytimes timeline, how do other news media sources compare in their overall major topic agendas? Figure 3.2 illustrates estimate topic agenda for each source from 2007 (or the earliest date the source began distributing stories via Twitter) through April 2023. The values in each cell represent the proportion of stories tweeted that were predicted to be about the topic in the y-axis (columns add to 100 percent).

A review of this figure also lends credence to the validity of the SVM model: the distribution of topic attention conforms with prior expectations about each source.

- *Environment* and *Energy* together account for 80 percent of @insideclimate's agenda.
- @CTMagazine has substantially higher attention to the *Church and Religion* topic; additionally, @CTMagazine tends to give greater attention to the *Civil Rights, Minority Issues, and Civil Liberties* topic, along with @realDailyWire, @TheRoot and @ATLBlackStar.
- @TheRoot and @ATLBlackStar also have great attention levels for the *Law, Crime, and Family Issues*, which includes coverage of police brutality.
- Finally, note that @WSJ give substantially greater attention to the *Banking, Finance, and Domestic Commerce* and *Space, Science, Technology, and Communications* topics than other news sources; @WSJ also had higher estimated attention to *Macroeconomics* relative to other sources.

Attention levels for the *Environment* major topic, which includes the climate change minor topic, are among the lowest values on this chart. As is made clear in this figure, perceived left-leaning sources (New York Times front page, @nytimes and @NPR) have higher levels of attention compared to perceived right-leaning sources (@WSJ, @realDailyWire). Interestingly, the Black- and Hispanic-owned sources have the lowest *Environment* attention overall. Do these trends differ for the issue of climate change, specifically?

	New York Times Front Page	@nytimes	@NPR	@insideclimate	@TheRoot	@ATLBlackStar	@LaOpinionLA	@WSJ	@realDailyWire	@CTmagazine
Macroeconomics	3%	2%	2%	0%	0%	0%	1%	6%	1%	0%
Civil Rights, Minority Issues, and Civil Liberties	3%	3%	4%	1%	19%	15%	3%	2%	13%	12%
Health	8%	7%	11%	2%	2%	2%	9%	8%	6%	4%
Agriculture	0%	0%	1%	1%	0%	0%	0%	1%	0%	0%
Labor and Employment	2%	1%	1%	0%	0%	0%	1%	2%	1%	0%
Education	3%	2%	2%	0%	2%	3%	1%	2%	2%	2%
Environment	2%	2%	2%	62%	0%	0%	1%	1%	1%	0%
Energy	1%	1%	1%	18%	0%	0%	0%	1%	1%	0%
Immigration	1%	1%	1%	0%	0%	0%	3%	1%	2%	1%
Transportation	2%	2%	2%	1%	0%	0%	3%	3%	1%	0%
Law, Crime, and Family Issues	8%	9%	9%	1%	25%	26%	11%	5%	15%	7%
Social Welfare	0%	0%	0%		0%	0%	0%	0%	0%	0%
Community Development and Housing Issues	1%	1%	1%	0%	0%	0%	1%	2%	0%	0%
Banking, Finance, and Domestic Commerce	5%	3%	2%	1%	1%	1%	2%	14%	1%	0%
Defense	9%	4%	4%	0%	0%	1%	1%	2%	3%	1%
Space, Science, Technology and Communications	2%	4%	4%	1%	1%	1%	4%	7%	2%	2%
Foreign Trade	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%
International Affairs and Foreign Aid	21%	19%	17%	3%	2%	7%	10%	16%	8%	12%
Government Operations	18%	12%	13%	3%	12%	6%	5%	9%	30%	3%
Public Lands and Water Management	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
Arts and Entertainment	4%	16%	14%	0%	27%	30%	30%	11%	10%	17%
State and Local Government Administration	2%	1%	0%	0%	0%	0%	0%	0%	1%	
Weather and Natural Disasters	1%	2%	3%	3%	0%	0%	2%	1%	0%	0%
Fires	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Sports and Recreation	3%	6%	3%	0%	4%	5%	11%	4%	2%	1%
Death Notices	0%	0%	0%		0%	0%	0%	0%	0%	1%
Churches and Religion	1%	0%	1%	0%	0%	0%	0%	0%	1%	35%
Other	0%	0%	0%		0%	0%	0%	0%	0%	0%

Figure 3.3: Estimated major topic attention, overall (2007-2023)

3.6.2.1 Climate change and semi-storms

Figure 3.4 shows monthly attention to climate change measured as the output of the enriched SVM model (dark blue) as well as secondary attention to climate change or climate change associations measured as rates of Boolean mentions of climate change keywords (two or more hits in green, one or more hits in yellow).[5] This figure illustrates the volume of stories that might have been classified as attention to climate change when using only Boolean classification.

Of note in this figure is the absence of media storms. Figure 3.4 illustrates climate change attention by month and smoothed for visual presentation; both of these steps reduce the severity of spikes in attention to climate change, potentially masking the presence of climate change "media storms" in the data. However, when examining weekly attention to climate change (standard for media storm analyses), there was not a single case in the 16-year study period where climate change attention met the criteria to be considered a "media storm": a 150 percent increase in attention and 20 percent of attention for 7 days (Boydstun *et al.*, 2014b).

As Boydstun *et al.* (2014b) argue, this operationalization of media storms is not rigidly fixed: while these three criteria above (size, explosiveness, duration) were established on sound theoretical considerations, they can be adjusted to accommodate the requirements of a given research task. That said, weekly climate change attention rarely accounted for or exceeded just 10 percent of a source's weekly stories. There were just four instances of these when measuring attention with enriched SVM classification (or 38 when using two or more hits, 179 when Boolean one or more hits).

[5]Note: @insideclimate was selected for analysis to increase set of articles labeled as "climate change" stories. However, this source is typically omitted from figures. Given that this source is intended to focus primarily on climate change, the level of attention from this source expands figure scales. This source was omitted to make trends in other sources legible.

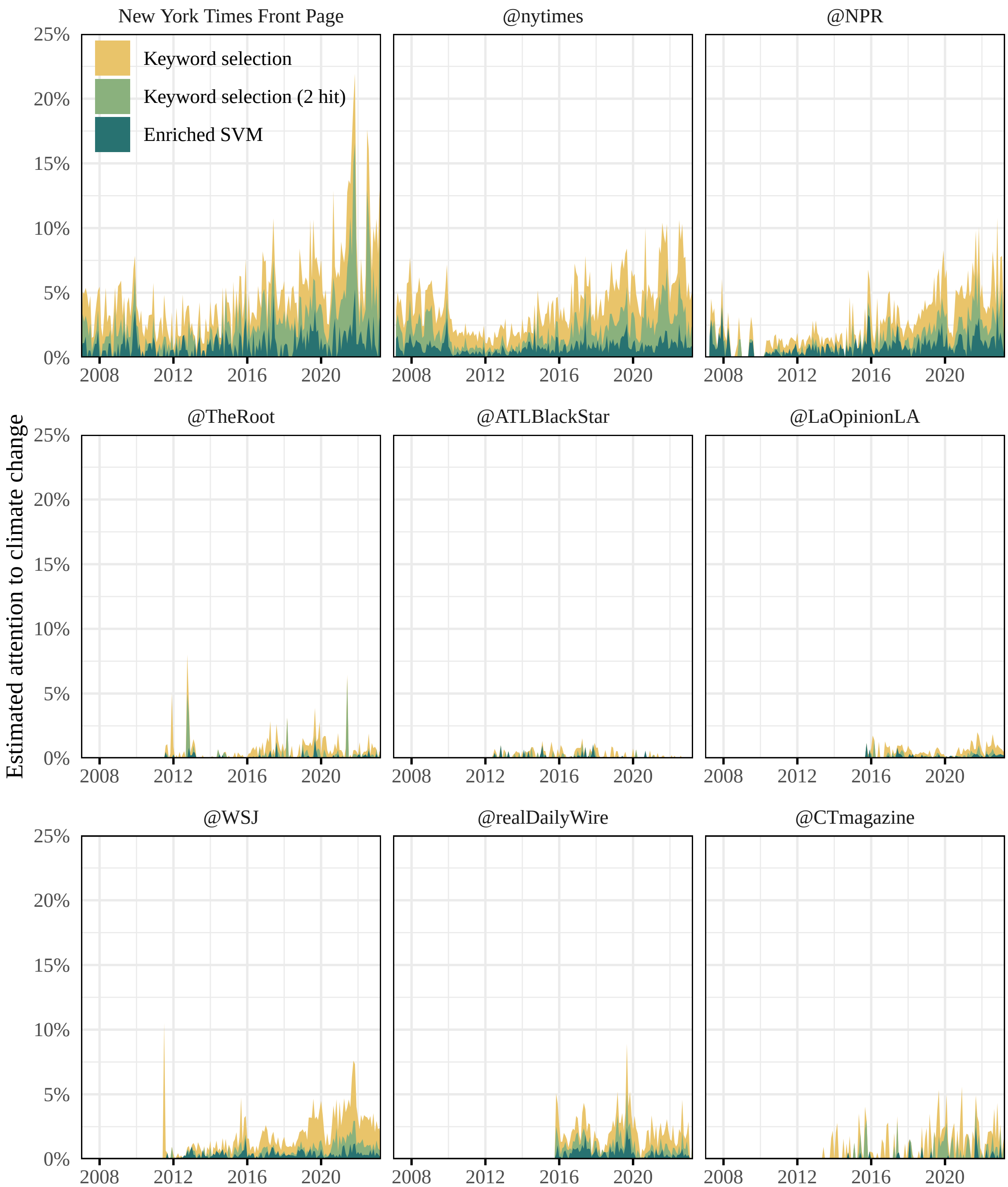

Figure 3.4: Attention to climate change and climate change associations

These "semi-storms" were driven primarily by international events, focusing on daily developments in United Nations Climate Change conferences. One of the observed "semi-storms" was from @NPR's coverage of the 2007 UN Climate Change Conference in Bali. The remainder were New York Times front page coverage of the December 2009 Copenhagen Summit, the November 2021 Conference of Parties (COP26) in Glasgow, and President Trump's announcement that the United States would back out of the 2015 Paris Agreement on climate change mitigation. Coverage of climate change during this last example was more diverse: whereas other instances of climate change semi-storms focused on daily developments of climate talks, the semi-storm surround Trump's decision covered Trump's past attempts to spread misinformation about climate change being a hoax, the United States's role as a major greenhouse gas emitter on the global stage, and how states and local governments are attempting to address the climate crisis in the absence of federal action.

With multiple news sources' Twitter timelines collected (as well as timeline meta data such as number of retweets) it is possible to examine a fourth criterion for media storms, "multi-media-ness." True media storms "register as such across multiple news outlets in a given media system" (Boydstun *et al.*, 2014b, p. 512). With this additional criterion, there would have been no instances of "semi-storms" in the 2007 to 2023 period as there were no instances where semi-storms occurred concurrently across sources or domains. Only one instance was a "near miss semi-storm," meeting the fourth criteria, that being coverage of the November 2021 Glasgow climate talks on the front page of the New York Times, which peaked at 11 percent attention, and @NPR, which peaked at 6 percent.

There were several more instances of semi-storms when measuring climate change attention with Boolean classification, particularly near the end of 2021, when single-hit classification reached 20 percent monthly New York Times front page attention. However, these stories were primarily focused on the Biden Administration's infrastructure bill, which included provisions related to climate change mitigation and adaptation. There was also a focus on the U.S. debt crisis, which was often associated with the Biden Administation's infrastructure bill and climate change provisions, as well as the unanticipated degree of Democratic opposition to the infrastructure bill from Senator Machine (D-WV) and Senator Sinema (D-AZ).[6]

Upon manual review of these two months where climate change hits peaked, the enriched SVM climate change classification model had a false negative rate of about 3 percent to 8 percent, similar to the false negative rate estimated for the enriched SVM model for out of sample data when training models in Chapter 2. Of the 63 stories mentioning climate change keywords that were not already predicted to be primarily about climate change by the enriched SVM, just two should have been classified as climate change:

— *Saving History With Sandbags: Climate Change Threatens the Smithsonian* (Flavelle, 2021)
—*With Methane and Forest Deals, Climate Summit Offers Hope After Gloomy Start* (Tankersley *et al.*, 2021)

Based on the CAP coding scheme, other countries attempting to address policy topics are categorized under the *International Relations and Foreign Aid* major topic. Given the importance of international cooperation in addressing this global issue, stories like the following three might be considered primarily about climate change in future research. However, this research treats stories like the following as primarily international, consistent with past CAP research.

— *An Electricity Crisis Complicates the Climate Crisis in Europe* (Eddy & Sengupta, 2021)
— *Skateboards, Climate Change and Freedom: Germany's Next-Generation Parliament* (Bennhold & Eddy, 2021)
— *On a Pacific Island, Russia Tests Its Battle Plan for Climate Change* (Troianovski, 2021)

The climate change attention rate when measured as Boolean hits inflated attention estimates by a factor of about three (2+ hits) to four (1 hit or more); these findings are consistent with the review of Boolean classification performance in Chapter 2.

[6]Note: These stories were predicted to be primarily about climate change when the stories focused primarily on the climate elements of the infrastructure bill but were not classified as climate change when these provisions were merely mentioned. This is an example of one of the main weaknesses of automated classification: all stories about the infrastructure bill should be in the same topic, as a development on the same story related to the same topic.

3.6.2.2 Public amplification and attenuation

Examining social media data offers the added benefit of transforming news on "the front page of the internet" into a participatory platform. Here, the public's engagement with tweeted stories contributes to measuring attention. In the analysis of media storms, we are searching for levels of media focus on events that "media consumers cannot help but know about an issue because it is so prominent in the news" (Boydstun *et al.*, 2014b, p. 511). By retweeting stories on their timelines, Twitter users can play a pivotal role in disseminating information and amplifying awareness of specific issues. While there were no climate change media-storms during the study period and only a handful of "semi-storms," we can test whether attention to climate change is more likely to meet media storm and semi-storm criteria when weighting tweeted stories by Twitter user retweets.

Retweets are taken into account here by treating each retweet as equal in weight to each New York Times tweet: public-weighted attention is measured as the number of tweeted stories about climate change plus the number of retweets on climate change stories, divided by the sum of all tweeted stories and the total number of retweets on all stories. Below I review public-weighted attention to climate change stories among sources that have a relatively high degree of public engagement on their tweets (average of 30 or more retweets per tweet) to prevent a small number of Twitter users from over-exaggerating the amplification or attenuation effects on issue attention.

When attention is measured with audience amplification, climate change is much more likely to meet more of the criteria for media storms and semi-storms. Figure 3.5 illustrates the weekly average of climate change attention (measured by enriched SVM) as the black vertical lines; green lines extending from these attention estimates indicate *attention amplification* by public engagement. Red lines indicate the level of attenuation, or the level that climate change attention falls when taking retweets on all stories tweeted by source and time period into account.

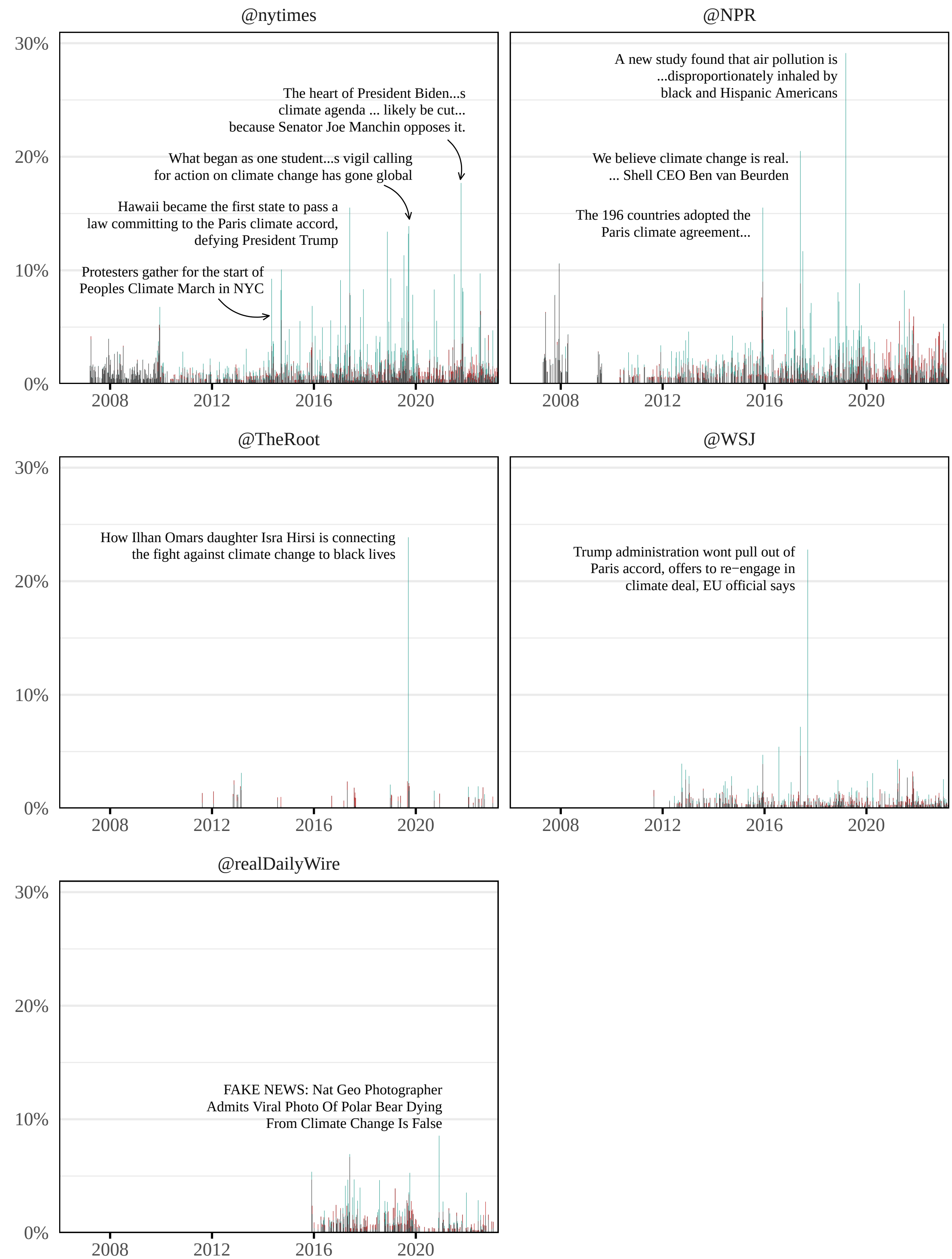

Figure 3.5: Attention to climate change weighted by retweets (2007-2023)

Public amplification of stories appears to be audience-dependent: the left-leaning @nytimes and @NPR climate change attention levels are frequently and highly amplified by their audience (relative to other stories) while there is infrequent public engagement with climate stories from the right-leaning @WSJ and @realDailyWire. When users do engage with climate change stories on Twitter, it appears they amplify messages that are associated with each source's issue priorities: @TheRoot and @NPR followers amplified climate change stories emphasizing racial justice; @realDailyWire followers amplified a story suggesting one a climate change narrative is based on misinformation; @nytimes followers tended to amplify stories about climate change protests and government action to address climate change.

This figure also illustrates the impact of issue congestion. At the start of 2020 with the outbreak of the COVID-19 pandemic, there was a noticeable shift away from both coverage and amplification of climate change stories. Not only did sources like @nytimes and @NPR tweet stories about climate change at a lower rate, but these tweets were more likely to be ignored relative to stories about other topics (like the global pandemic). Attention to climate change appears during this period appears to have been attenuated by the public in favor of other important issues.

3.6.3 Long term trends in climate change attention

The analyses above indicate a substantial amount of variability in weekly and monthly news media climate change attention, amplification and attenuation. However, does the long-term trend show an upward trajectory in attention, as past research suggests? Using generalized linear mixed effects models, I test three hypotheses related to long-term trends in climate change attention in American news media: (H3.2a) climate change attention has increased between 2007 and 2023; (H3.2b) Boolean classification significantly inflates estimated attention to climate change; and (H3.2c) Boolean classification inflates estimated rates of climate change attention growth.

As the motivation for this analysis is to examine differences in inferred rates of climate change attention and attention growth compared to past literature, I subset the analyzed data to traditional "prestige" sources (New York Times front page, @nytimes, @NPR, @WSJ). To emphasize long-term trends, climate change attention is aggregated by month.

Table 3.1: Mixed effects models estimating difference in attention by classification method

	News media attention to climate change		
	(1)	(2)	(3)
Boolean (1+ hits)	1.505***	1.444***	1.447***
	(0.079)	(0.078)	(0.078)
Boolean (2+ hits)	0.801***	0.729***	0.731***
	(0.079)	(0.078)	(0.078)
Date		0.375***	0.280***
		(0.072)	(0.085)
Date ×			
Boolean (1+ hits)			0.138*
			(0.073)
Boolean (2+ hits)			0.143**
			(0.072)
Constant	−4.923***	−4.979***	−4.983***
	(0.220)	(0.278)	(0.279)
N	2,166	2,166	2,166
Log Likelihood	7,442.100	7,504.619	7,507.073
AIC	−14,874.200	−14,993.240	−14,994.150
BIC	−14,845.800	−14,947.790	−14,937.340

*p < .1; **p < .05; ***p < .01

First, supporting Hypothesis 3.2a, the *Date* variable is positive and statistically significant in Models 2 and 3, indicating a general upward trend in climate change attention over time among these news sources.[7] The *Boolean (1+ hits)* and *Boolean (2+ hits)* coefficients are intercept shifts, both are positive and statistically significant in all three models. These coefficients are relative to the overall intercept (*Constant*), which refers to the intercept of the baseline category (enriched SVM). These intercept shifts support Hypothesis 3.2b: both Boolean classification methods inflate levels of climate change attention and the level of upward bias is statistically significant. Finally, the interaction terms between *Method* and *Date* are positive and statistically significant, supporting Hypothesis 3.2c, that the upward bias of Boolean classification has increased over time. Likelihood ratio tests comparing the models above suggest that Models 2 and 3 perform substantially better than Model 1. The interaction model (Model 3) fits the data slightly better than Model 2 (*Pr(>Chisq)* ≈ 0.086).

H3.2a: News media attention to climate change increased between 2007 and 2023. **(Supported)**
H3.2b: Boolean classification significantly inflates estimated attention to climate change. **(Supported)**
H3.2c: Boolean classification inflates estimated rates of climate change attention growth. **(Supported)**

The discussion above shows the external validity of the enriched SVM model to delineate stories about climate change from stories merely mentioning climate change and related key phrases. But it also illustrates the need to account for error and to model uncertainty around climate change attention estimates. The previous chapter showed that, on the training and test data, Boolean classification had high false positive rates; the results of the statistical model indicate that the false positive rate is substantial and significant.

[7] Note that the model was a generalized linear mixed effects model: the coefficients do not correspond directly to increases in attention but instead show the change in the log of the expected *Attention* value for a one-unit change in the predictor variable. The coefficients must be summed and exponentiated to retrieve estimates for *Attention* values on a (0,1) scale.

3.6.4 Differences in attention by source

A separate mixed effects model was fit to evaluate differences in climate change attention by *Source*, assuming fixed effects in the interaction between *Source* and *Date* (Model 3). To account for variation in how media sources respond to real-world events, this model and the models below, include random effects for the ideological *Lean* of source by year. I compare this model's performance to simpler models: first, a simple long-term fixed *Date* effect model with the random *Lean* | *Year* effects; and second, a long-term fixed *Date* effect with fixed *Source* intercepts and random *Lean* | *Year* intercepts.

With these models I test the following two hypotheses: (H3.3a) attention is higher among left-leaning news sources than right-leaning sources; and (H3.3b) attention is increasing more rapidly among Black- and Hispanic-owned news sources. Likelihood ratio tests comparing model performance suggest that the interaction model (Model 3) fits the data substantially better than models 1 and 2 (*Pr(>Chisq)* < 0.001). The fixed effects coefficients and intercepts for these models are presented in Table 3.2 below.

Prior to fitting the model, @nytimes was set as the base news source as it had the lowest levels of climate change attention among left-leaning sources, overall. Setting this group as the baseline sets a higher bar for identifying statistically significant differences between @nytimes and right-leaning sources with the highest levels of climate change attention. This also provides clarity regarding differences between coefficients for left-leaning, right-leaning and other sources.

The *Date* coefficient is positive and statistically significant in all three models, indicating a positive trend upward in attention overall. These results lend additional support for H3.2a, that attention has increased from 2007 to 2023, with an alternatively-specified model. Intercept shifts in Model 2 indicate that the New York Times front page, and @NPR coverage of climate change did not significantly differ from @nytimes. Right-leaning sources, however, had substantially lower and statistically significant intercepts in Model 2, supporting H3.3a.

Table 3.2: Mixed effects model estimating climate change attention by source

	(1)	(2)	(3)
Date	0.198***	0.207***	0.131***
	(0.075)	(0.071)	(0.003)
New York Times front page		0.191	0.241
		(0.230)	(0.171)
@NPR		0.025	0.057
		(0.233)	(0.172)
@TheRoot		−2.698***	−2.661***
		(0.319)	(0.003)
@ATLBlackStar		−2.801***	−2.851***
		(0.328)	(0.003)
@LaOpinionLA		−1.872***	−2.197***
		(0.374)	(0.003)
@WSJ		−0.984***	−0.972***
		(0.254)	(0.189)
@realDailyWire		−0.728**	−0.245
		(0.297)	(0.406)
@CTmagazine		−2.309***	−2.757***
		(0.275)	(0.228)
Date ×			
New York Times front page			0.134
			(0.137)
@NPR			0.079
			(0.137)
@TheRoot			−0.099***
			(0.003)
@ATLBlackStar			−1.918***
			(0.003)
@LaOpinionLA			0.386***
			(0.003)
@WSJ			0.215
			(0.243)
@realDailyWire			−0.728
			(0.506)
@CTmagazine			0.988***
			(0.320)
Constant	−4.697***	−4.712***	−4.740***
	(0.098)	(0.167)	(0.003)
N	1,304	1,304	1,304
Log Likelihood	8,999.858	9,037.568	9,053.918
AIC	−17,981.720	−18,041.130	−18,057.840
BIC	−17,935.160	−17,953.190	−17,928.510

*p < .1; **p < .05; ***p < .01

H3.3a: Climate change attention is higher among left-leaning sources than right-leaning sources. **(Supported)**

I anticipated that due to the rise of race-related focusing events in the late 2010s, such as incidents of police brutality and the emergence of the Black Lives Matter movement, there would be an increased likelihood of discussing other topics, like climate change, in the context of racial justice. Research about racial inequities in regarding climate change and pollution was published (e.g., Schlosberg & Collins, 2014) and distributed on Twitter during this period (e.g., Lambert, 2019). Academic research and investigative journalism emphasizing racial disparities in the effects of climate change should, in turn, drive up climate change attention among news sources whose intended audiences are primarily people of color and potentially issue publics for issues related to social justice and racial equity.

Model 3 results in Table 3.2 provides mixed results, howvever. Black- and Hispanic-owned news sources had substantially lower intercepts in Models 2 and 3. However, interaction coefficients with *Date* were negative and statistically significant for Black-owned news sources @TheRoot and @ATLBlackStar. The interaction coefficient for @LaOpinionLA is positive and statistically significant but this may be misleading without contextualizing the coefficient in terms of the model intercept, source intercept, and *Date* variable. According to Model 3, @LaOpinionLA climate change attention was estimated to increase from about 0.1 percent in the Fall of 2015 to about 0.2 percent in Spring of 2023. While the relative effect of *Date* for @LaOpinionLA appears substantial (average attention doubled in 8 years), the absolute effect was very small (an increase of one tenth of one percent).

H3.3b: Attention to climate change is increasing more recently and rapidly among Black- and Hispanic-owned news sources. **(Not supported)**

3.7 Discussion

Above I applied an enriched SVM classification model to estimate climate change attention across a vast amount of news data distributed via the New York Times front page and various news media content creators on Twitter (N = 1,085,260). The analyses above indicate that attention to climate change has increased over time but levels of attention are substantially lower than past research indicates when measuring attention to climate change as the share of stories mentioning climate change key terms.

These findings show that Boolean classification is not appropriate for measuring news media attention to climate change. The impact of climate change is frequently mentioned in weather reports; stories on natural disasters, and wildfires; news roundups summarizing the top stories of the day might mention climate change but these stories should not be classified as *about* climate change.

Critical to these analyses is the finding that Boolean classification not only inflates the level of estimated climate change attention, it inflates the apparent growth rate of climate change attention. Given the reliability of the enriched SVM to produce estimates of stories that are primarily about climate change, the difference in the estimated climate change attention growth rates between these two classification methods suggests that the issue of climate change is becoming associated with other topics at an increasing rate over time. While Boolean classification is an inappropriate method of classification for climate change stories, it becoming less appropriate over time.

Contrary to my expectations, there was no evidence to support the hypothesis that climate change attention is higher on the @nytimes Twitter timeline relative to the New York Times front page. I expected that by removing spatial constraints from the New York Times front page, this source would devote a greater share of its attention to looming and creeping issues. While the results above do support the general notion that topic agendas differ by topic across domains, the issue of climate change was not one of the topics that differed between domains. It is possible that upon removing spatial constraints of the front page paper, the

@nytimes Twitter timeline fills its agenda space with topics Twitter users might engage with (such as the arts and sports).

Consistent with past research, I found that left-leaning sources have higher levels of climate change attention relative to right-leaning sources. I expected levels of attention to increase faster for Black- and Hispanic-owned news sources over time. Instead, I found that @TheRoot, @ATLBlackStar and @LaOpinionLA have not engaged much with the issue. As shown in Figure 3.3, this is not due to these sources intending to be apolitical; to the contrary, these sources had some of the highest rates of attention to issues related to police brutality and civil rights. This may suggest that the issue of climate change as not been sufficiently expressed in terms that are impactful to people of color. This is reviewed more closely in the next chapter.

CHAPTER 4

EXPANDING ISSUE ASSOCIATIONS AND TOPIC DIFFUSION

4.1 Introduction

Previous chapters cautioned against employing keyword hits for climate change classification tasks; yet, when utilized in a different context, keyword hits can retrieve valuable information from text data. In particular, keywords here are used as the central points for a 255-topic guided Latent Dirichlet Allocation (LDA) topic model. As discussed below, these keywords form the starting points for an iterative process to infer latent topics in text data. Climate change keywords are also used here to examine the diffusion of this issue into other major topics. Specifically, we might treat the false positives in Boolean classification in Chapters 2 and 3 as climate change as relevance to other topics. In short, keyword hits provide incomplete information about media attention or agendas but, understood as issue diffusion, provide more valuable information about the breadth of the issue into the full space of political issues in American politics.

The latent topics inferred from text data mentioned above are used to examine trends in climate change issue associations in American news media over time. As discussed in previous chapters, the costs of climate change are estimated to be wide, far reaching, and severe; it is critical that public's information environment reflect the potential costs of inaction on the climate crisis. Content analysis using LDA features enables examination of how much and how often climate change stories discuss topics related to energy, the environment, public lands and water, as well as issues that historically have been more exotic to the issue, such as immigration, labor and defense. Analyses below indicate that the number of issues discussed in climate change stories (degree centrality score) has increased in prestige news media sources over time. Similarly, the number of unique topics that mention climate change as a relevant consideration have also increased over time.

News media is critical to the public's awareness and understanding of climate change, as news media is often the only exposure the public has to the issue (Bolsen & Shapiro, 2018; Brulle *et al.*, 2012). Climate change is what might be called a "non-obtrusive" issue in that, for many, it is not visible in daily life and the public must be told about it (Gene Zucker, 1978). Those who live in regions experiencing the impacts of climate change are more likely to believe it is a threat (Myers *et al.*, 2013). The way that climate change and environmentalism has historically been covered has led to additional perception barriers related to social identity.

Concern about climate change and environmentalism in general have suffered from being perceived as post-materialist issues indicative of privilege or elitism (Buttel & Flinn, 1978; Inglehart, 1984; Morrison & Dunlap, 1986; Taylor, 1997), as having a historically white bias (Taylor, 2002), or even being too effeminate for white males (Swim *et al.*, 2020). The motivating question of this chapter is whether news media convey the issue of climate change in ways that are meaningful to diverse audiences. As Haden *et al.* (2012) show, engaging with issue publics about climate change in terms where they have the greatest amount of efficacy may have the potential to overcome partisan bias and other psychological barriers to build public commitment to pro-environmental behaviors (e.g., farmers and the impact of climate change on crop yields, methods for mitigation). Additionally, recent research has shown the importance of engaging diverse audiences in addressing issues related to media isolation and echo chambers; Saveski et al., for example, find that tweets designed to be appealing to ideologically diverse audiences had higher engagement from ideologically diverse audiences (Saveski *et al.*, 2022). Diversity in climate change associations, then, may be critical to building public buy-in for climate policy reform and pro-environmental beliefs and behaviors.

In this research I evaluate the breadth of discussion on climate change stories (and climate change mentions on non-climate change stories) for a signal on whether the issues associated with climate change has become reflective of the varied expected impacts of climate change in academic literature and IPCC reports. This chapter illustrates that climate change issue associations have increased and strengthened over time. Climate change is associated with a greater number of topics now than it was in 2010 and substantially more topics than the 2000s and before. A wider variety of major topics make up the content of climate change stories and climate change mentions are becoming more prevalent in coverage of other issues.

Due to the experimental nature of the topic model used in this research, a substantial portion of this chapter is devoted to a description of the sampling process, measures of coherence and qualitative validity checks. Results overall suggest that CAP-based guided LDA may be a helpful tool to evaluate the breadth and depth of topic associations in the media's watchdog or patrol coverage of climate change and other topics. However, the validity of measured topic associations is conditional on prior knowledge about the source and its willingness to treat the issue of climate change seriously. As discussed below, topic associations for the New York Times appear to be a valid indicator of the depth of climate change coverage; for sources like the Daily Wire, measures of topic association appear to measure the source's creativity in denigrating climate science and the actors involved in climate action.

4.2 Past research

In addition to the attention issues receive, the way issues are communicated in news media matters. The way climate change is framed in news media may effectively encourage support for climate action (Rickard *et al.*, 2016). As noted above, it matters whether the scope of climate change is constrained to impacts on exotic wildlife halfway across the globe or if its effects are in our backyard. The portrayal of climate change as established in science or shrouded in uncertainty (or worse, a hoax) may impact public perception, policy reform and collective action. Content analysis of climate change media coverage has tended to be from manual review of climate change stories. "It matters whether the environment is discussed in terms of the spaceship-ness of the Earth, the greenhouse-ness of climate change, or the disease-ness of pollution" (Myerson & Rydin, s.d., p. 25). However, as argued in Chapter 1, manual content analysis is difficult to scale and the rate at which news media available for consumers is expanding, automation in content analysis is become more crucial than ever.

4.2.1 Manual content analysis

Chapter 1 of this report began with a brief overview of how much (and how) the New York Times covered the issue of climate change between 1996 and April 2023. There were a handful of distinct periods characterized by the how much climate change extent and breadth of coverage. Attention was rare in the 1990s and 2000s; few front page stories covered climate change were and the few that did were primarily related to energy efficiency improvements in consumer goods. The early 2010s were rife with attention to (without sponsorship of) climate denierism, although it was in this period that the New York Times began causally linking climate change with extreme weather events. In the mid 2010s, front page coverage of climate change expanded in ways that may be more meaningful to a more diverse audience: the issue was associated with the Christian concept of Environmental Stewardship, with immigration, racial justice, the economy, and more. Front page news about climate change has since focused primarily on political actors acting: international conferences, political positions, and congressional gridlock on the issue.

Most academic literature on content analysis of climate change news media is based on manual review. Feldman *et al.* (2017), for example, manually identify emerging frames related to climate change impacts across several policy issues, similar the research presented here. Feldman *et al.* (2017) find important differences in impact frames related to environment, health, and the economy when comparing New York Times coverage to Wall Street Journal. Similarly, Ford & King (2015) review climate change of leading print in the US and Canada (including NYT and WSJ) around the same time period and find increasingly prevalent usage of frames related to adaptation, rather than climate change impact or mitigation.

A review of right-leaning bloggers websites by Elsasser & Dunlap (2013) finds that climate change discussions were composed of four distinct ways of discussing the issue: misinformation (e.g., the Earth is cooling), followed by other potential causes for climate change (e.g., solar radiation), then downplaying the effects, followed by the costs of meaningful climate change mitigation efforts. Halfway to full automation, Jang & Hart (2015) use manually constructed frame dictionaries to evaluate the prevalence of frames in public discussions on Twitter and find that about two-third of online debates between 2012 and 2014 relate to the veracity of climate change and denierism.

4.2.2 Automated content analysis

In using automated content analysis, researchers are effectively trading rich contextual framing analysis for a broader picture of themes and associations that is scalable to massive amounts of text data. Framing analysis provides deeper insights into how an issue is being communicated but future research ought to couple qualitative review of frames with automated methods. As noted in Chapter 2, some climate change communications scholars have used LDA in their research and interpreted LDA topic distributions as frames (Boussalis *et al.*, 2016; Boussalis & Coan, 2016); however, LDA for framing analysis may not be appropriate given the "bag-of-words" nature of the model. Framing analysis depends heavily on semantics and local context of terms and phrases (i.e., the order of words). Bag-of-words models, such as LDA, are agnostic toward semantics and local context but excel when evaluating global context of texts and surpass Neural Networks in producing topic distributions that are easily interpretable by humans (Pan & Ding, 2019). LDA features are not treated as frames in this chapter but rather as issue associations.

Keller *et al.* (2020), for example, use LDA features summarized into four topics, finding an increasing emphasis on terms related to "impact" in climate change stories between 1997 and 2016. Vu *et al.* (2019) managed to produce more fine-tuned LDA features to report an increased emphasis on impacts as well, but specifically a shift from economic impacts to more "natural" impacts of climate change between 2011 and 2015. Hase *et al.* (2021) use a

Structural Topic Model, similar to LDA, to identify themes in domestic and international news media coverage of climate change including climate science, impacts (on the ecosystem, humans, economy), causes and solutions, politics, as well as awareness and education.

However, as discussed below, a critical component of the automated content analysis research pipeline is characterized by ad-hoc or post-hoc rationalizations or labeling of topic clusters according to the topics' most informative terms. The component of the research pipeline makes replication, corroboration, and collaboration difficult to impossible. Each LDA model is task- and corpus-specific. The guided LDA approach suggested here offers researchers greater control over initialization parameters that produce latent topics distributions (Jones *et al.*, 2021). Critiques of automated content analysis can then be directed at inferences regarding a transparent topic structure and assumed top words that are central to topics, rather than an unknowable process of feature review and labeling. Unguided LDA may be most appropriate for analyzing emerging themes in text documents for which the reader has no prior information. However, content analysis of climate change is sufficiently established to justify the use of the prior information in the development of LDA features.

4.2.3 Network analysis

This research provides a novel contribution to the analysis of LDA features in climate change communications research. It is the first research to use guided LDA topics in the CAP coding scheme as network nodes to analyze topic associations. Treating topics as nodes and content topic proportions as connections or issue associations, we can measure the centrality or interconnectedness of climate change in the major topic space of news media, as well as the diffusion of climate change in other topics.

As mentioned above, how the issue of climate change is conveyed by the media influence the public's information environment. Carley & Palmquist (1992) introduced the concept of mental models, the cognitive framework that people use to process and understand their information environment. Mental models are internal representations based on networks of concepts, as well as other factors. How people process and understand information can play a significant role in how they perceive choice sets, opportunities and how they interact with the world.

Network analysis has been applied in previous research on policy frames. Shim *et al.* (2015), for example, analyzed policy frames through semantic network analysis related to nuclear energy and connections to catastrophic events and other themes. Kim *et al.* (2007) in the evaluation of international conflicts and the need for international aid. Past semantic network analysis is often performed at the term or sentence level, as the examples above. Here, I evaluate topic associations in terms of documents' theta scores from the guided LDA model, i.e., the estimated *proportion* of a document related to a policy topic.

4.3 Theory

Research on the framing of climate change articles has centered around terms of threat (or costs), solutions (Feldman *et al.*, 2017), and efficacy (Haden *et al.*, 2012). The literature on framing in climate change communication finds mixed results and sometimes "boomerang effects" on public opinion (i.e., the way in which a message was framed produced the opposite effect from what was intended) due to the polarized nature of public opinion on climate change and the polarizing nature of the topic (Hart & Nisbet, 2012). Framing a climate change story in terms of technical solutions may increase readers' propensity to free-ride and await a solution beyond their personal efficacy; on the other hand, overemphasizing the costs of climate change may cause readers to express defeatism in the face of an insurmountable problem (Johns & Jacquet, 2018).

The perceived proximity and intensity of climate change costs may also spur readers into action (Haden *et al.*, 2012; Scannell & Gifford, 2013). Construal Theory suggests that the public's perceived importance of climate change is influenced by the temporal, spatial, social, and hypothetical proximity of the costs of climate change, as well as other dimensions such their intensity (Rickard *et al.*, 2016; Spence *et al.*, 2012). In Haden *et al.* (2012), the costs of climate change are framed in terms of agricultural impacts and the intended audience (or sample of respondents) consisted of farmers. The issue of climate change in this study was communicated in terms of costs to a secondary issue, one that most directly appealed to an issue public; in response, that issue public expressed support for pro-environmental practices in their occupation, where individual efficacy is feasibly much greater than individual efficacy in affecting state or federal environmental policy.

The findings in Haden *et al.* (2012) suggest that issue proximity may also increase commitment to pro-environmental policies and behaviors to curb the effects of climate change and that these effects may overcome partisanship. Yet the association of climate change with other issues has not been explored in climate change communication literature. Issue association content analysis of news media coverage is thus a crucial addition to climate change communications literature.

This chapter makes three novel contributions to climate change communication literature: first, the analyses in this chapter explore whether and how climate change is associated with other important policy topics, and whether there are notable differences in association trends across sources and time. Second, this chapter is the first of its kind to employ a guided LDA topic model to address many of the shortcomings of unsupervised LDA and other techniques in automated content analysis (discussed in the next section). Third, the keyword-assisted LDA uses pre-labeled data to construct a hierarchical topic structure mirroring the four-digit level of the Comparative Agendas Project (CAP) topic coding schema, arguably the most comprehensive topic structure constructed in government and media policy agendas and communications literature (Jones *et al.*, 2023). This approach corresponds well with qualitative content analysis and represents an improvement over unsupervised automated techniques in content analysis.

First, this chapter explores the diversity of in-depth engagement with various policy topics in climate change articles, as well as the diffusion or permeation of climate change into coverage of other topics. Once again, the costs of climate change are predicted to be pervasive, and as the costs of climate change become nearer and clearer, the pervasiveness of climate change in nearly every other topic should be measurable in news media through the increased prevalence of climate change as a secondary topic and the increased diversity of secondary topics in climate change articles.

Hypothesis 4.1: The number of topics discussed in climate change stories has increased over time.

Hypothesis 4.2: The diversity topics where climate change is discussed has increased over time.

Secondary topic diversity trends are expected to differ across sources depending on the source's perceived political ideological leanings and intended audience (as well as perceived impacts of climate change on the intended audience or policies prioritized by the audience). Conservative news media outlets are not expected to earnestly provide in-depth reporting of climate change, given that many conservatives in U.S. politics believe (and tweet) that climate change is a hoax (Fownes *et al.*, 2018). Rather, in right-leaning news media outlets climate change will primarily be discussed in terms of the topic itself, its veracity, or as attacks on political actors involved in climate change policy or protest (e.g., ad hominem attacks on Greta Thunberg during climate change protests).

Hypothesis 4.3: Climate change issue associations are less extensive and less diverse among right-leaning sources.

4.4 Data

The data analyzed in this chapter include the same set of news stories examined in the previous chapter: stories tweeted by various sources[1] between June 2007 and April 2023 as well as stories that appeared on the New York Times front page in the same timer period. New York Times front page data from 1996 to 2006 are examined separately, as these data do include only the first three paragraphs from each story and should not be compared to

[1]Sources include @nytimes, @NPR, @WSJ, @TheRoot, @ATLBlackStar, @LaOpinionLA, @insideclimate @realDailyWire and @CTmagazine.

full-content stories. Tweeted stories were removed from analysis for inactive links that could not be retrieved via Wayback Machine, for linking to an alternative news source or domain, or linking to multimedia without text content to analyze. Also removed were tweets linking to strictly self-promotional materials as well as — critically — news roundups.

Following screening and cleaning procedures, the final database analyzed here contains about 1.1 million stories (N = 1,085,260). These stories were classified according to their primary topic using the best performing major topic classification model in Chapter 2. This Support Vector Machine (SVM) model had about 78 percent out-of-sample major topic classification accuracy. Monte Carlo simulations were conducted to analyze trends in major topic classifications while accounting for classification error.

The distributions of major topics (according to the Comparative Agenda Project coding scheme) were estimated for each of the stories in the data. Major topic distributions were estimated using the guided LDA process described below.

4.5 Methods

This section reviews the methods used in this chapter to test the hypotheses above. Discussion of methods includes the guided Latent Dirichlet Allocation (LDA) model introduced in Chapter 2, the SVM and enriched SVM models from Chapter 2 were used for major topic and climate change classification, as well as measures of (and operationalization of) issue association, diversity and diffusion.

4.5.1 Guided Latent Dirichlet Allocation

Stories' topic distributions are estimated using a guided Latent Dirichlet Allocation (LDA) model introduced in Chapter 2 (Jones *et al.*, 2021). This section provides a very brief overview of the elements of the LDA sampling process improved upon with guided LDA. See Blei *et al.* (2003) for a more comprehensive review of the LDA procure.

LDA is a Bayesian statistical procedure that produces probabilistic estimates of latent topic distributions in text documents. This procedure works by iteratively assigning latent topic probabilities to terms and measuring the likelihood of observing the full corpus when simulating text data using term-topic probability assignments. The procedure begins by setting random initialization parameters: terms are assigned topics at random for an initial estimate of term-topic probabilities and topic distributions through the corpus, which are then refined through sampling and Bayesian updating.

For each word in each document, the word is temporarily excluded from the document and the topic distribution for that document is estimated. The probability of each topic conditional on the presence of the selected word is estimated, then these estimates are multiplied for a combined probability for each topic. The word is reassigned to a new topic based on the updated probabilities and the term-topic assignment counts are updated. Guided LDA improves on LDA and supervised LDA by structuring the initialization parameters using prior information about a domain: rather than start from terms' random assignment to topics, topic assignment is conditional on prior information about topics and their terms.

4.5.1.1 LDA limitations

There are several drawbacks to LDA that are overcome by using guided LDA. One drawback to LDA is that the output is sensitive to the K selected by the researcher: a K too small may result in distinct topics being combined while a K too large may produce incoherent topics (Boyd-Graber *et al.*, 2014; Roberts *et al.*, 2016). The number of topics in a corpus must be assumed to be known and user-specified, although it is an open research question best to optimize the K number of documents Roberts *et al.* (2016). Some have attempted to optimize K by optimizing metrics meant to measure model performance in terms of topic coherence and perplexity Zhao & Zou (2015); however, some research suggests that these measures negatively correlate with human judgment about topic coherence Chang *et al.* (2009). Others have bypassed this issue by simply running several LDA models setting K at different values and using the features from all of these models (e.g., Inkpen & Razavi, 2014). This suggests that, rather than optimizing K with statistical measures, the number of topics ought to be established in theory or with prior information.

The LDA procedure also suffers in terms of replicability, as topics must be labeled post-hoc by evaluating the terms with the highest topic probabilities (Bischof & Airoldi, 2012). After the process of optimizing the term-topic probabilities, researchers examine each latent topic to assign a topic label based on the terms with the highest probability for that latent topic. This process also necessarily requires a researcher's subjective judgement when labeling topics according to perceived themes among terms that are predictive of a latent topic, a procedure which cannot be replicated, although some have attempted to assign topic labels based on topic correlations with prelabeled data (e.g., Barberá *et al.*, 2019). Additionally, statistical measures of model fit tend to improve as K increases, meaning that the topic models that better fit the data also require more time and subjective judgments from the researcher.

Research relying on LDA and its variants is often task-specific and limited in the scope of its external applications: the output of the program will be highly curated to the research question, without much room for application outside of the corpus it was trained on and for the specific task it was trained to do. There are supervised versions of the LDA process that do take prior information into account.

Supervised LDA (sLDA) estimates latent topics that maximize the likelihood of observing pre-labeled classifications. However, the latent topics estimated by the model are not specified in advance, only the number of topics; following the sLDA, there still remains the task of reviewing and labeling topics based on topics' highest probability terms Lakshminarayanan & Raich (2011). While sLDA may fit the data to pre-labeled data, it does not avoid LDA's pitfalls related to transparency and replicability. Guided, or keyword-assisted, LDA offers a way forward.

4.5.1.2 Guided LDA improvements

The guided LDA approach improves over unsupervised and supervised LDA (sLDA) by granting the researcher additional control over the terms that are central to each topic. In regular LDA settings, terms all have a random prior probability of belonging to any given latent topic (Blei *et al.*, 2003); sLDA uses pre-labeled data and produces latent topic distributions by maximizes the joint likelihood of articles' classifications and content (Glenny & al., 2019). With guided LDA, however, the user can set the prior topic probabilities — *lexical priors* — for selected terms (Eshima *et al.*, 2020; Jones *et al.*, 2021).

The key feature of the guided LDA is a researcher-constructed term-topic dictionary of lexical priors that assists the LDA in optimizing the term-topic probabilities that best fit the data while imposing a somewhat flexible structure on the latent topics (Eshima *et al.*, 2020; Jagarlamudi *et al.*, 2012; Jones *et al.*, 2021; Meng & al., 2019). The iterative process described above, rather than starting from random initialization parameters, starts with weakly informative prior probabilities that guide the remainder of the process. Using lexical priors in automated content increases replicability of research: topic labels are defined prior to fitting the LDA model rather than after. Any researcher with database of news media documents can use lexical priors in Appendix B to produce a topic model with the same structure as the one presented here, with central terms that are likely similar to those presented here (depending on the prevalence of those terms in the corpus).

The issue of post-hoc rationalizations in topic labeling is abated by using lexical priors following the Comparative Agendas Project (CAP) coding scheme using terms frequently (and exclusively) associated with four-digit topic codes in the prelabled New York Times Front Page 1996-2006 data set. Not all subjectivity was removed from the process, however, as a review of the top terms per topic requires the subjective judgment of the researcher to determine whether the guided LDA produces valid topics while considering the topic the terms were meant to represent. While subjective, this process can be replicated, as it involves only a "valid" or "not valid" judgment, rather than a much more intricate process of labeling a topic based on its top terms.

The guided LDA is also extremely flexible: the model updates term-topic probabilities so that the initial terms selected as belonging to a topic are not deterministic. However, deterministic elements can be used in this approach and were employed sparingly here for proper separation of topics: a feature is available only through a guided LDA. Taking advantage of the fact that prior probabilities of 0 or 1 cannot be updated, certain words can be selected to never (or always) appear in a latent topic. Consider the *Energy* topic: a large proportion of *Energy* stories in the New York Times front page 1996-2006 data mention the costs of energy (price per barrel of oil) production when covering energy issues. Latent topics estimated by both LDA and sLDA models on these data considered terms like “price” and “cost” to be highly predictive of the “Energy” topic (or, more fitting, a *cost per barrel of oil* subtopic).[2] Additionally, the user can specify any number of unguided latent topics for the model to fit, so that latent topics that are not accounted for in the term-topic dictionary may still be predicted and evaluated.

Another benefit of the guided LDA, as discussed above, is the potential to directly connect new research to past research: because the lexical priors are based in the CAP coding scheme, inferences about story proportions can be directly tied to other research using CAP. The Comparative Agendas Project topic coding scheme is arguably the most comprehensive topic structure constructed in government and media policy agendas and communications literature and its usage here provides a link to many other analyses using the CAP schema. Other media communications researchers, when not coding communications data with this schema directly, have attempted to build crosswalk tables to link LDA output to the CAP topic structure (e.g., Barberá & al., 2019).

[2]Worse, terms that were associated with “cost” or “price” were guilty of *Energy* by association: if a story had discussed the price or costs of climate change, the model output would suggest a high association with *Energy* when it should have indicated high association with *Macroeconomics*; any appearance of “cost” or “price” in a story would cause the LDA output to be biased toward the *Energy* topic. By assigning deterministic prior probabilities to such terms, the model is better able to separate energy terms from macroeconomic terms and improves the validity of topic distributions (and thus improves the validity of topic associations). This functionality is not possible when using LDA without topics’ lexical priors.

That said, slight modifications were made to the topic structure to account for: 1) infrequent topics lacking predictive terms where a coherent topic could not be statistically constructed in the training data (e.g., agricultural marketing, youth employment, intergovernmental relations); 2) topics with conceptual and term overlap substantial enough that one or more of the topics produced incoherent sets of predictive words (e.g., "Pest Control" and "Food Safety" subtopics in the agriculture major topic); 3) topics and terms expected to arise in articles published after 2006, which do not appear in the New York Times Front Page data set (e.g., the rise of cryptocurrencies, streaming services, and COVID-19). Finally, major topics that refer to regions (International Affairs, State and Local Government) are comprised of minor topics for geographic subunits (e.g., Eastern Europe, the American West), the keywords for which consist of the names of countries, states and major cities, as well as region-specific political titles and positions (e.g., prime minister, governor).

In sum, the process of estimating topic proportions using CAP-structured lexical priors on a guided LDA requires fewer subjective judgments, increases research replicability, and introduces potential opportunities to link LDA inferences across studies. Next, I briefly review the process for creating the CAP-structured lexical prior term-topic probability table.

4.5.1.3 Lexical prior selection process

Lexical priors were selected from highly informative terms in pre-labeled data to create a solid foundation of prior information about news media and the CAP coding scheme. The New York Times front page 1996-2006 data (Boydstun, 2014) was pre-labeled using the CAP coding scheme at both the two-digit major topic and the four-digit minor topic levels. Each four-digit topic was individually examined for highly informative terms: the term frequency inverse document frequency (tf-idf) was calculated for each four-digit topic. The tf-idf score indicates the relative importance and exclusivity of a term to a given document or topic. Up to five terms were selected for each four-digit topic with slight modifications for expected shifts in terms and topics beyond the 2006 data (e.g., inclusion of terms "obama_administration," trump_administration," and "biden_administration" as lexical priors in the *Government Operations: Executive Branch* minor topic). There were 243 minor topics created in the CAP lexical prior table for 28 different topics, as well was 12 additional minor "unguided" topics that operate under random initialization procedures as in unguided LDA.

4.5.2 Guided LDA validity

As presented in Chapter 2, the guided LDA major topic distributions were highly informative when applied using a linear Support Vector Machine for major topic classification. The out-of-sample predictive accuracy of the LDA-SVM model was about 66 percent, nearly as reliable as the Naive Bayes model (72 percent). Simply using the largest LDA feature had about 38 percent out of sample accuracy. The OOS F1 score was about 83 percent for an SVM model predicting climate change stories using guided LDA features as predictors, only 2 percentage points behind the standard SVM model using terms as predictors and 6 percentage points behind the enriched SVM model using terms and guided LDA features as predictors. These are powerful indicators of the validity of the guided LDA features. Beyond classification accuracy, there are also statistical measures of topic coherence, such as probabilistic topic coherence.

4.5.2.1 Probabilistic coherence

Probabilistic coherence estimates the coherence of a topic by considering the differences in observing top topic words together in documents in a corpus (Jones *et al.*, 2021). Briefly, probabilistic coherence takes the average probability of top words appearing together in a corpus. For the general environment topic, for example, calculate the probability of observing *earth* in documents where *environment* was mentioned, $\mathrm{P}(\textit{earth}|\textit{environment})$, then subtract the probability of observing *earth* in all documents in the corpus, $\mathrm{P}(\textit{earth})$. Assuming these two terms are observed together frequently and exclusively, then their difference, $\mathrm{P}(\textit{earth}|\textit{environment})$ - $\mathrm{P}(\textit{earth})$, should be near zero (suggesting that coherence scores near zero indicate relatively stronger coherence). Repeating this process for the top five terms for each topic and averaging the differences produces the distribution of probabilistic coherence scores in Figure 4.1.

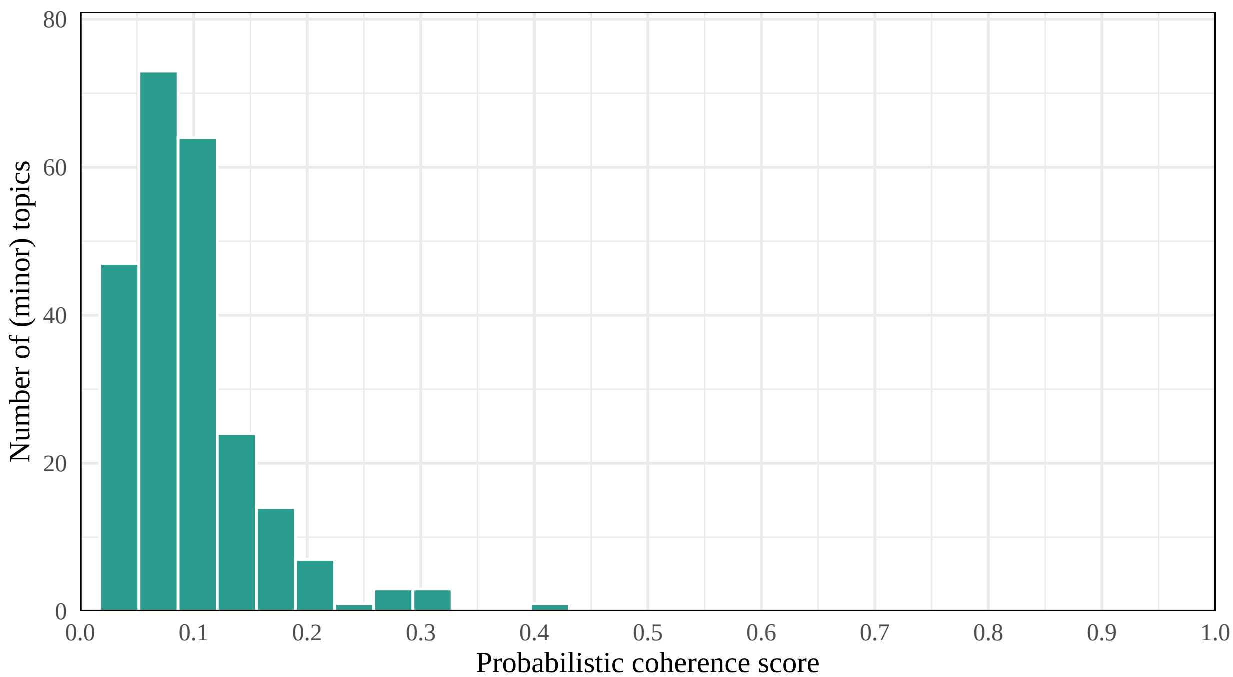

Figure 4.1: Estimated probabilistic topic coherence

Note that measures of topic coherence, including the probabilistic topic coherence presented here, are not *metrics*: there is no objective "good" or "bad" measure of coherence. While the distribution of probabilistic coherence illustrated above is promising, past research suggests that statistical scores of model fit may not conform with human judgment Chang *et al.* (2009). Therefore, there was an extensive review of the major topic distribution estimates in out-of-sample data to ensure validity in estimated topic distributions and inferences based on these estimates. Manual review of topic validity conforms to the coherence distribution above: most topics appear to be highly valid with only a small number of (minor) topics with noisier topic signals (but low prevalence in the data, indicating that these topics do not impact the results below). Two qualitative checks are provided below. One of which is a *New York Times* story. First is a review of the estimated topic distribution of this manuscript.

4.5.2.2 Qualitative validity check: Dissertation example

Having read to this point, reader, what would you say this research about? This report has covered several topics up to this point: a meta analysis of methods used to categorize news stories, technological advances enabling automated data collection and information retrieval, trends in news media, and treating social media as a news media distribution platform. "Climate change" has been used more than 400 times. Government operations, racial justice, energy production and other topics have come up in the discussion of attention to climate change. So how does this all translate to a major topic distribution according to the Comparative Agenda Project coding scheme? Figure 4.2 provides the estimated topic proportion of the first three chapters of this dissertation; it includes the top four major topics as well as the five influential terms used here that contributed to each topic.

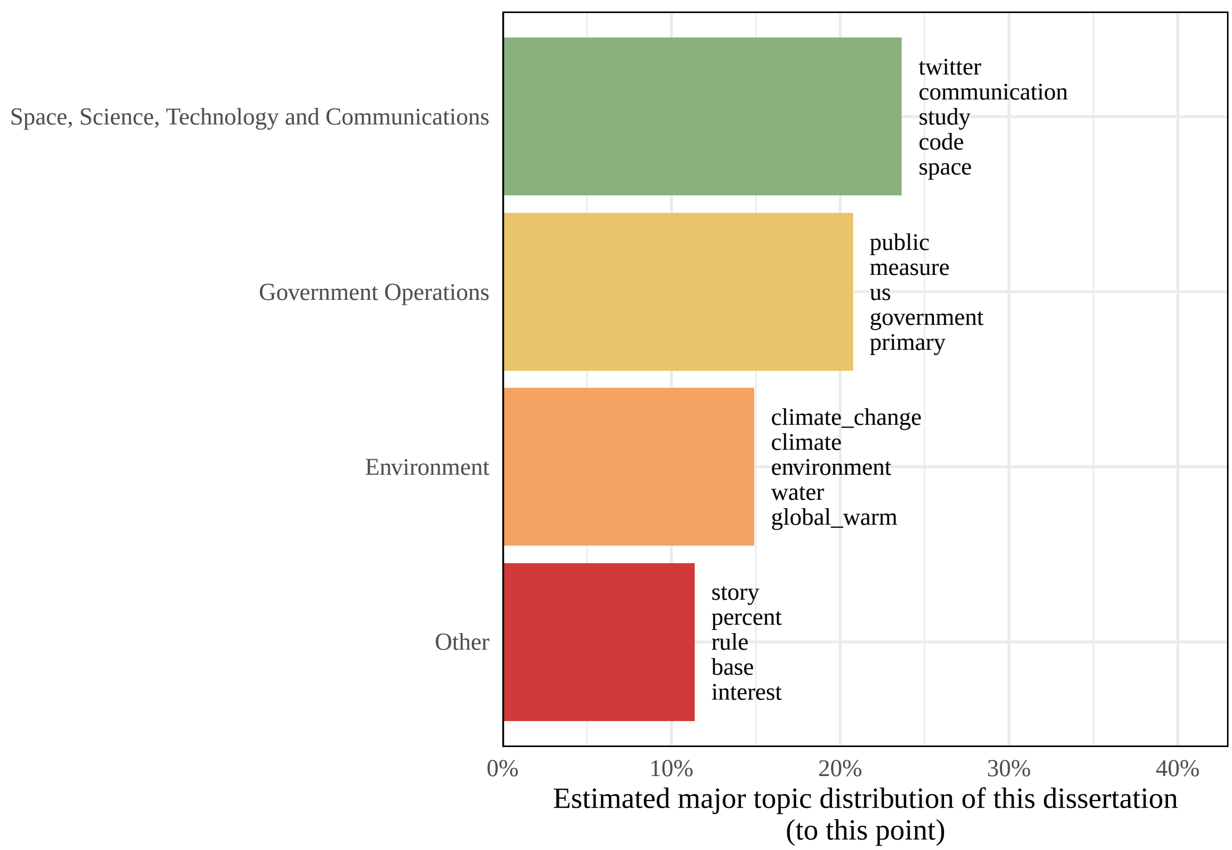

Figure 4.2: Estimated major topic distribution of this dissertation

Figure 4.2 is at once a testament to the validity of the guided LDA model and a cautionary tale. The figure illustrates the guided LDA estimates of the topic distribution for this dissertation up to Figure 4.2 (for topics account for more than ten percent of the total content). This figure shows the importance of domain in the context of automated content analysis: the term-topic probabilities were optimized using news media training data and the topic structure was designed to sort terms related to news coverage (e.g., "story") into an "other" category. LDA models generally suffer when applied to domains outside of the data they were trained on.

This figure also includes the terms (processed and stemmed) that were frequently used and highly relevant to major topics. According to the model, about 24 percent of the dissertation belongs to the *Space, Science, Technology and Communication* major topic. This was driven by discussions of Twitter and references to communication, such as in "climate change communications research." The model also estimates a high proportion for *Government Operations*, driven by discussion of the relationship between media attention, public opinion, and public policy (the *Government Operations* major topic includes a *public relations* minor topic, see Appendix B). The model also estimated a high proportion of the *Environment* major topic, which was driven almost entirely by the term "climate change," which had been used more than 500 times up to this point.

From extensive qualitative review, topic proportion (theta) scores of about 5 to 10 percent appear to be the aggregation of noisy and uninformative terms. At these theta values, topic proportion estimates are somewhat noisy and unreliable. Thus, analyses in the Results section are repeated under varying assumptions of sufficient signal clarity (or different theta thresholds).

4.5.2.3 Qualitative validity check: New York Times example

Figure 4.3 below illustrates the topic distribution as estimated by the guided LDA model, showing the predicted proportion of each major topic. The text used for this example comes from a front page news article from the New York Times with the headline "California taking big gamble, tries to curb greenhouse gases" (Barringer, 2006). As mentioned above, low topic scores are typically due to uninformative terms that appear frequently across all topics; topic scores around 5 percent and below are likely attributable to these noisy terms. Topic scores above 5 percent tend to be valid and meaningful representations of the text. Compare the topic distributions from Figure 4.3 to the first paragraph of the article:

> "In the Rocky Mountain States and the fast growing desert Southwest, more than 20 power plants, designed to burn coal that is plentiful and cheap, are on the drawing boards. Much of the power, their owners expected, would be destined for the people of California. But such plants would also be among the country's most potent producers of carbon dioxide, the king of gases linked to global warming. So California has just delivered a new message to these energy suppliers: If you cannot produce power with the lowest possible emissions of these greenhouse gases, we are not interested."

One of the major contributors to climate change is how we produce power and so many climate change articles like the one above are associated with energy production. Here the validity of the *Energy* topic score is clearly visible, with several energy-related terms appearing in the first paragraph (e.g., "power plants," "coal", "energy suppliers"). The same can be said for *State and Local Government* major topic (with terms like "Rocky Mountain States," "Southwest," "California," etc.) and Environment ("global warming," "emissions," and "greenhouse gases"). The topic scores for *Banking and Finance and Macroeconomics* may seem questionable compared to the first paragraph, but the content later in the story refers to the buying and selling of energy in the Southwest, as well as an optimistic take on California's ability to reduce emissions without "wrecking its economy."

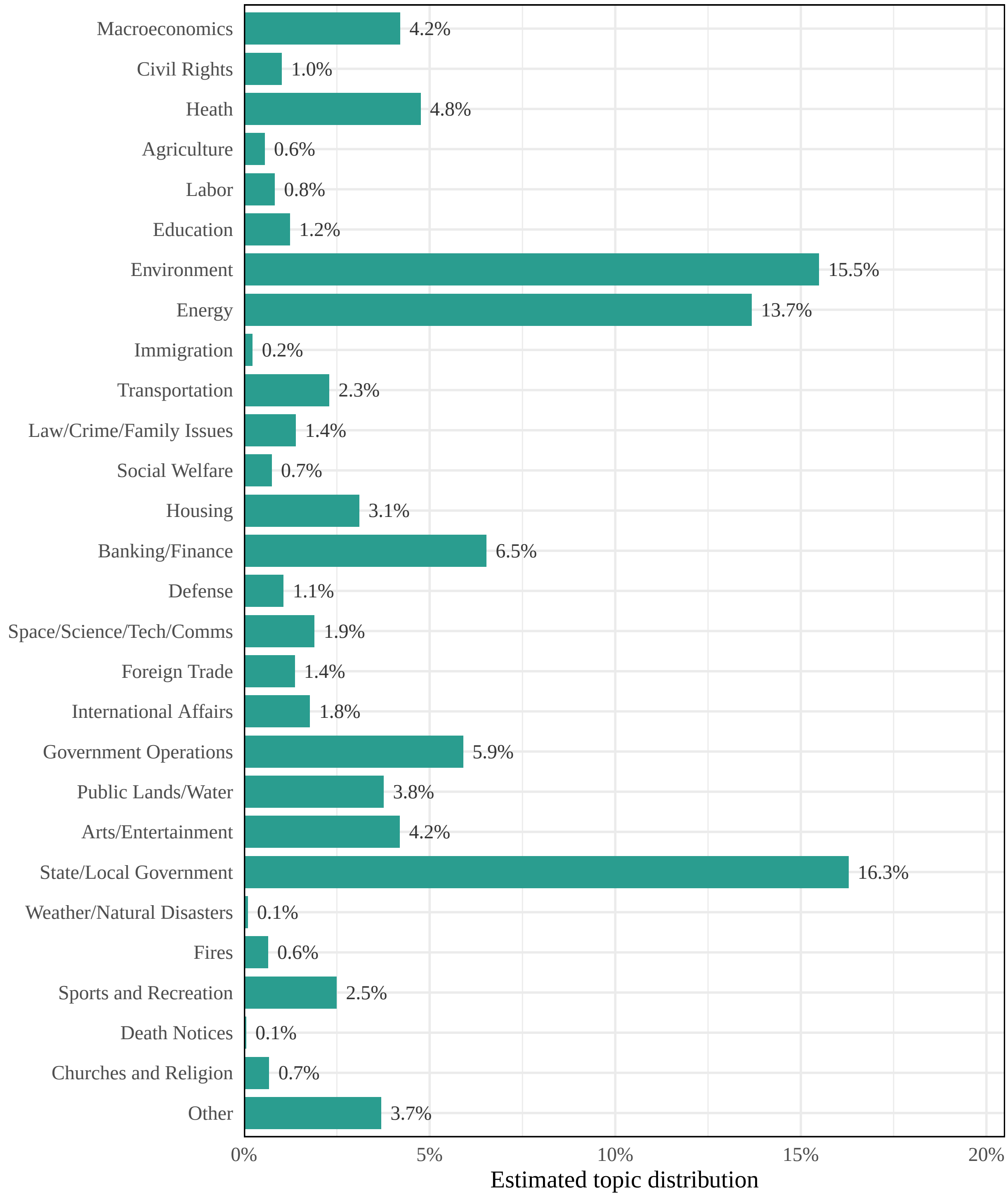

Figure 4.3: Guided LDA qualitative validity check

Note: Figure 4.3 illustrates the output of a guided LDA model applied to the a New York Times story's content following text pre-processing, lemmatization, and n-grammed lexical priors listed in Appendix B. The categories along the y-axis represent the 27 major topics in the Comparative Agendas Project coding schema, as well as an "Other" category for latent topics not otherwise classified in the schema. The x-axis represents the LDA theta score (or proportion of the document predicted to belong to a given topic).

The figure above provides an example of how a secondary issue like energy production can be associated with climate change. The analyses later in this chapter will summarize trends in climate change topic associations like those above across the one-million row database. Below I examine these trends across time and by source.

4.5.3 Degree centrality

In graph theory and semantic networks, node importance is often measured using *degree centrality*: the number of connections that a node has (Borge-Holthoefer & Arenas, 2010). Treating topics as nodes in a network of news media issue association, degree centrality can be measured in terms of the number of topics that are discussed together in a single document. Climate change degree centrality is reported below as the number of major topics that are estimated to make up at least 15 percent of story content, as estimated by the guided LDA model. The diffusion of climate change into other topics is measured as the number number of unique major topics that mention the standard set of climate change keywords two or more times in article content. Climate change degree centrality and diffusion was measured using Monte Carlo simulations to produce a confidence interval around the centrality estimate.

4.5.4 Issue association diversity (entropy)

As argued above, diversity in issue associations may be useful in building support for climate reform from a more diverse audience. Here, I measure the entropy of climate change discussion as an indicator for the potential to engage with diverse audiences. As discussed above, *entropy* measures the how varied the elements in a system are. Entropy is measured here as Shannon's entropy (Boydstun *et al.*, 2014a):

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Maximum entropy scores (the log of the number of topics) would indicate an equal distribution of discussion to all other topics in a body of text while lower scores indicate that discussion is dominated by fewer topics. Below I examine entropy using varying minimum thresholds of story proportions estimated to be about the CAP major topics. Higher entropy scores indicate greater diversity of attention and the potential to diversify the audience exposed to the story.

4.6 Results

This section reviews the changing issue association landscape for climate change over time, starting with the number of unique topics that are frequently discussed in-depth in climate change coverage by year. Next, I review the diversity of major topic-climate change associations by source and over time and report that issue associations are diversifying over time. Additionally, I review the topic associations that appear to be strengthening over time. Finally, I repeat these analyses with non-climate change articles, examining the number of topics that mention climate-change related keywords to analyze the prevalence and breadth of climate change as a secondary consideration in various topics over time.

4.6.1 Increasing topic associations: degree centrality

As defined above, *degree centrality* refers to the number of connections a node has. Treating topics as nodes and content topic proportions as connections or issue associations, we can measure the centrality or interconnectedness of climate change in the major topic space of news media. Here I examine the proportion of topics in climate change stories, examining the connections directed to the climate change node; later in this chapter I consider co-occurrence of climate change and and review the topic distribution of the semantic network from the other direction (non-climate change stories mentioning climate change).

As argued above, the costs of climate change are estimated to be wide and far reaching and so it is important that public's information environment reflect the potential costs of inaction on the climate crisis. The centrality score for climate change has increased in news media over time. Figure 4.4 illustrates the degree centrality measured as the number of unique major topics in the CAP coding scheme with a minimum theta threshold, or a minimum proportion of story content belonging to a given topic.[3]

While there is ample evidence for the face validity and external validity of the guided LDA measure, degree centrality is measured at varying levels of estimated topic proportions, assuming that higher estimates are less prone to error. While apparently valid, the guided LDA features are experimental and varying the standards for issue associations enables a greater degree of confidence in observed results that are consistent across different theta thresholds. Similar trends in degree centrality are observed over time and by source, and higher standards for topic associations appear to be associated with varying intercepts in degree centrality rather than changes degree centrality slopes over time.

Higher estimates of topic proportions are also assumed to indicate serious treatment of the topic and its associations (whether positive or negative). However, the reader should be skeptical of topic associations for sources that do not take climate change seriously. The Daily Wire, for example, very rarely engaged with the issue of climate change seriously. When the Daily Wire engaged with the issue of climate change, it misrepresented research on the causes of climate change on a geological scale as these causes were responsible for the Earth's rapid warming in the past century.[4] High climate change centrality (and entropy) scores for the Daily Wire were due its clever ways to denigrate climate figures, such as Greta Thunberg traveling to the U.S. by yacht[5], which produced a high *Transportation* score, or by mocking Bill Nye's statement on MSNBC that climate change would require Canada to engage in agriculture, which produced a high *Agriculture* score.[6]

[3]Note that the theta estimate for the climate change minor topic was subtracted from the *Environment* major topic theta aggregation for this analysis and following analyses.

[4]See the following link for the full story: https://www.dailywire.com/news/scientists-we-know-what-really-causes-climate-james-barrett?

[5]https://web.archive.org/web/20190829172838/https://www.dailywire.com/news/51183/greta-thunberg-sailed-new-york-avoid-contributing-ashe-schow

[6]https://www.dailywire.com/news/watch-bill-nye-science-guy-panics-climate-change-emily-zanotti?%3Futm_source=twitter&utm_medium=social&utm_campaign=dwtwitter

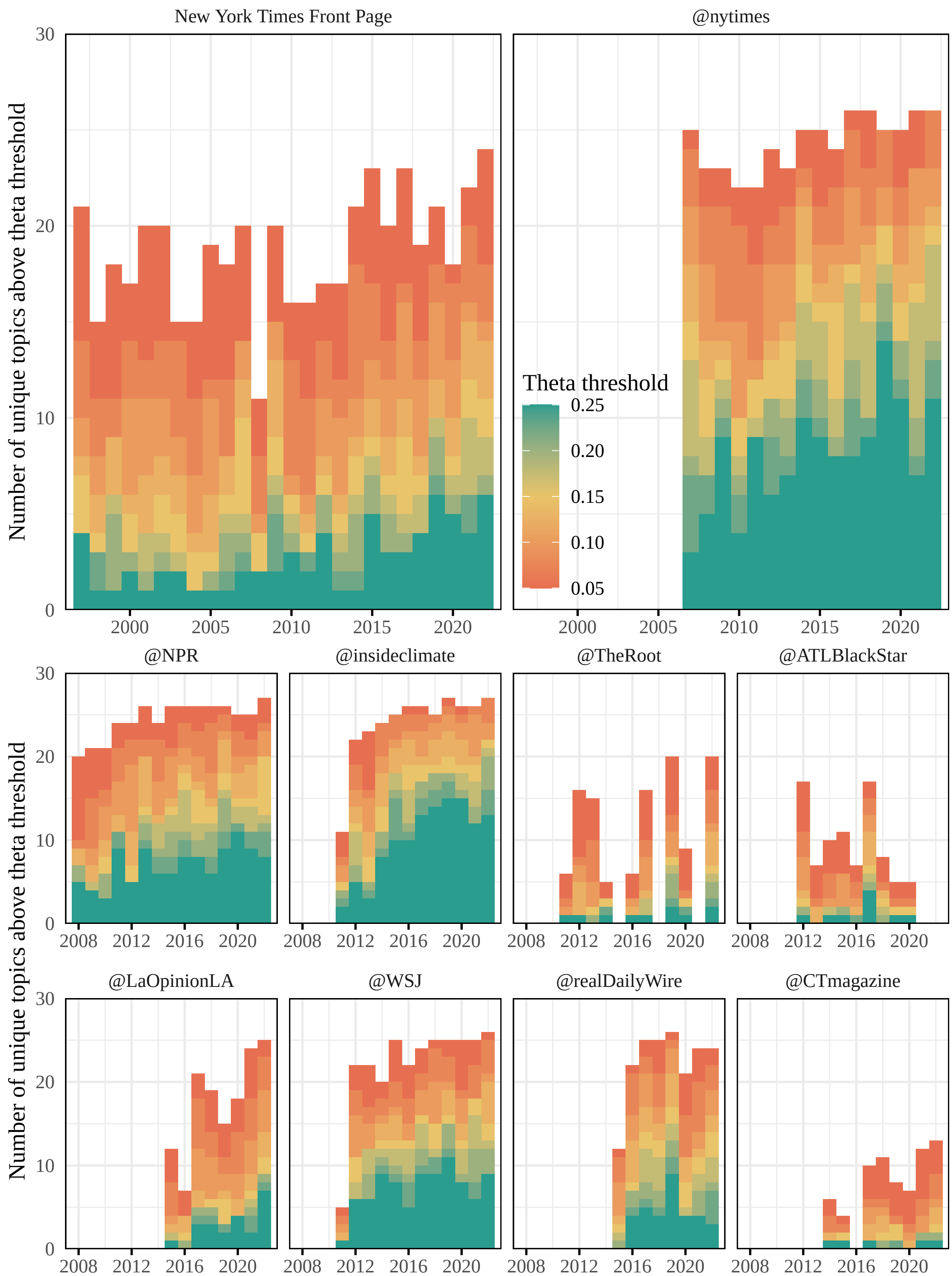

Figure 4.4: Climate change centrality in major topic space (varying theta thresholds)

H4.1: The number of topics discussed in climate change stories has increased over time.**(Partial support).**

4.6.2 Increasing diversity: entropy

As discussed above, *entropy* measures the how varied the elements in a system are. While the number of topics associated with climate change has increased over time, it is likely that only a few topics dominate climate change discussions while most other topics are infrequently mentioned. In other words, the number of topic associations may be growing but they may also be skewed toward a smaller number of topics.

As above, the guided LDA model produces valid estimates of topic proportions but the level at which a story can be said to have made a meaningful association between climate change and another topic is unknown. Thus, entropy analyses were repeated at varying levels of topic association thresholds (theta values from 0.05 to 0.25), indicated by the stacked lines in Figure 4.5. Interestingly, the trends in the entropy-over-time lines in Figure 4.5 suggest that left-leaning sources like @nytimes and @NPR appear to be more willing to engage more deeply into non-climate issues on climate change stories over time. There is a much wider gap between the lowest and highest theta thresholds for these sources early in the study periods and a much narrower gap in later years. In other words, raising the threshold for topic engagement has less of an impact on @nytimes and @NPR over time compared to other sources (and domains).

The data show an increase in estimated entropy over time for all "prestige" sources and a noisy or constant trend for "new" media sources. The values plotted in Figure 4.5 are far below the maximum entropy score, the log of the number of topics ($\log(27) \approx 3.296$), indicating that topic associations are far from evenly distributed across all major topics in the CAP coding scheme. Instead, as discussed below, climate change associations topic associations are typically dominated by a small number of topics, such as the *Energy* topic and the *Environment* topic (modified to omit theta estimates for the *climate change* subtopic).

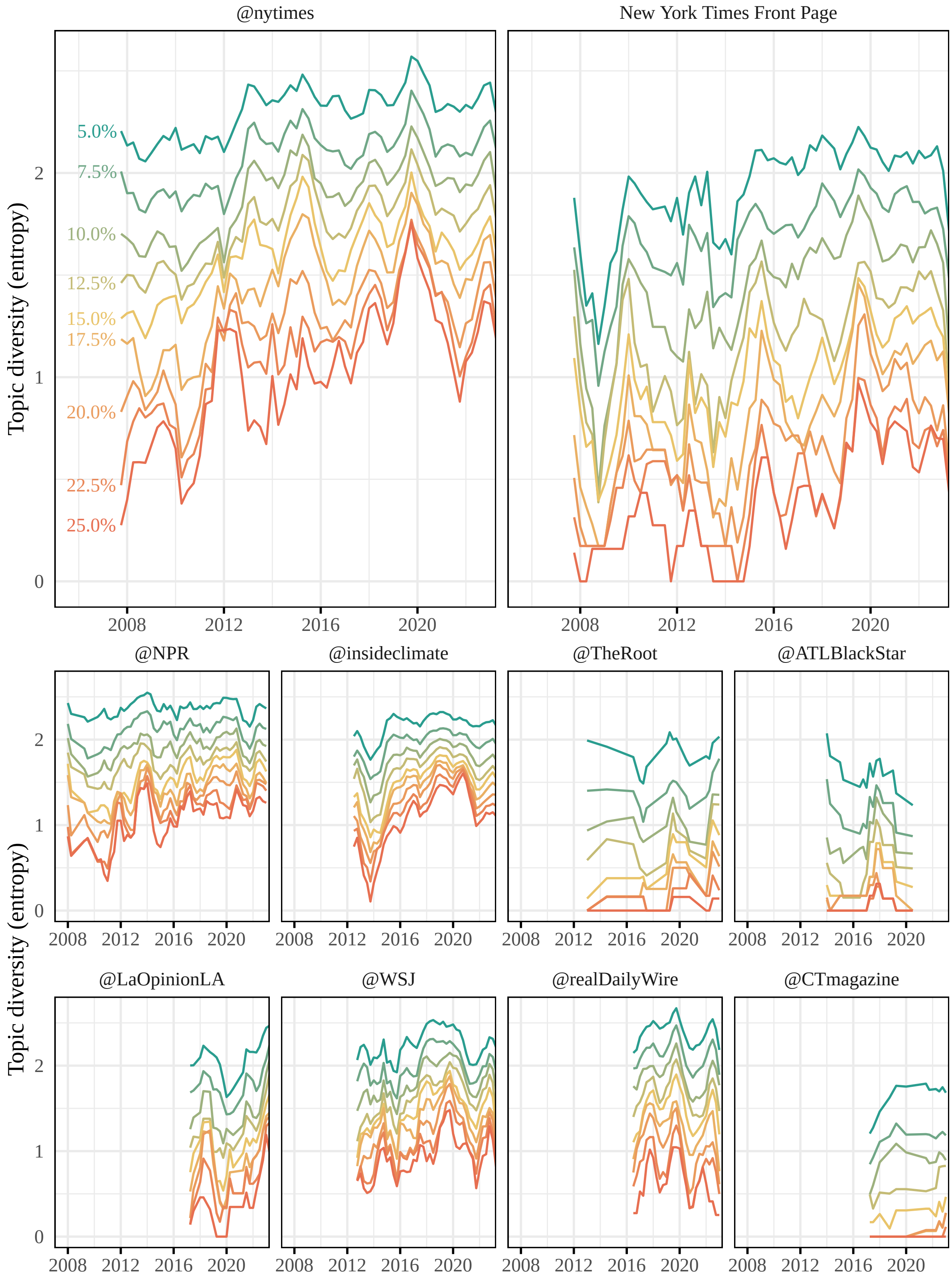

Figure 4.5: Diversity of climate change topic associations (entropy) by theta thresholds

H4.2: The diversity topics where climate change is is discussed has increased over time. **(Supported)**

As the costs of climate change becoming nearer and clearer, I expected climate change issue associations to become more diverse in news media coverage of climate-related events. The data support this hypothesis for prestige media, as the number of topics receiving in-depth attention on climate change has been increasing and diversifying over the past twenty years.[7]

The hypothesis was not supported for all media organizations however. A general upward trend in centrality and diversity for prestige media, including New York Times, National Public Radio, and Wall Street Journal (although this trend appeared weaker for Wall Street Journal in terms of entropy). Topic centrality and diversity are essentially constant over time for "new" media sources that do not typically cover climate change stories (e.g., @TheRoot).[8]

The New York Times had substantially lower centrality and diversity scores compared to @nytimes, suggesting the importance of watchdog or patrol journalism for climate change coverage. When events occur that warrant front page coverage of climate change stories, discussion of these events skews toward relatively fewer topics compared to coverage by @nytimes. As argued above, news stories distributed via Twitter may provide an alternative "front page of the internet" to examine stories that are emphasized by a source without the spatial constraints of a physical newspaper. The entropy scores illustrate that New York Times' non-front-page coverage of climate change stories is much more diverse in its associations with other major issues compared to front-page coverage.

[7]As a robustness check, the same analysis was performed for *Weather and Natural Disasters* coverage, which had F1 scores similar to *climate change* F1 scores according to predictive models reviewed in Chapter 2 as well as similar levels of news media attention in Chapter 3. Centrality and entropy scores varied over time but showed no clear upward trajectory over time, as is visible for climate change coverage.

[8]It is possible that "new" media sources choose to retweet external sources with coverage of climate change with high discussion of topics that are high priority to the sources' intended audiences. Source retweets were not collected for this analysis, nor were tweets linking to external domain names. However, when climate change is expressed in terms appealing to an issue public, we might expect the sources appealing to those issue publics to be more likely to pick up that story and reemphasize the elements important to their audience.

4.6.3 Emerging topic associations

The guided LDA gives the opportunity to examine the topic associations that have emerged from text data. Centrality and entropy are good indicators of trends overall but we can also examine which topics were contributing to these measures. The figures below include the annual aggregate of theta scores for each topic by source. These scores were scaled on the log scale for legibility. Topic associations were removed if the estimated theta score was below a 0.15 threshold for a clearer signal of serious engagement with other major topics.

The relative lack of diversity in New York Times front page coverage of climate change stories compared to @nytimes stories distributed via Twitter is apparent in Figure 4.6. Although recall that the 1996-2006 New York Times front page data included only the first three paragraphs of stories, making observation of serious topic associations unlikely in these data. That said, climate change topic associations in the 2007-2023 New York Times front page coverage of climate change are scarce compared to topic associations in @nytimes coverage during this time period. In-depth engagement with topics in front page stories were primarily related to *Environment* and *Energy* throughout the time period, although other topics on climate change stories began to expand after 2007.

The results closely mirror the review of trends in front page coverage of climate change stories at the start of this chapter: most coverage was related to efficiency improvements and minor policy updates; the first instances of linkage with weather were in 2009; 2010 marked the rise of denierism, which corresponds well to *Government Operations* (which includes a "misinformation" minor topic). The figure also features an uptick in *Science* in 2010, which includes language regarding scientific studies in general. The highlighted cells in *Law, Crime, and Family Issues* correspond well with climate related protests and protests in the United States and around the globe, particularly the September 2019 climate strikes.

Figure 4.6: Major topic prelalence in climate change stories

There were a great deal more topic associations on climate stories distributed via Twitter during the 2007-2023 period. This is somewhat expected given the greater number of climate change stories shared on Twitter relative to the number of stories making front page news.[9] From this figure we can see how New York Times patrols the issue of climate change, with most attention going toward the *Energy* and *Environment* major topics, followed by *Weather and Natural Disasters* and *Government Operations*.

While centrality is much higher for @nytimes climate change coverage, there are several topics that do not appear to have received in-depth engagement on climate change stories. *Macroeconomics*, *Civil Rights*, *Immigration*, *Social Welfare* and *Housing* are all likely to be impacted by the effects of climate change and yet receive very little discussion in climate change stories shared by @nytimes. Although an absence of these associations may be due to: 1) LDA topics failing to fully capture terms related to these topics; 2) overly restrictive criteria for theta estimates that may tend to be small due to topics consisting of a small number of highly informative terms (the climate change LDA topic is one example).

4.6.4 Emerging associations by source

Inside Climate News (@insideclimate) is a niche news organization dedicated to coverage of climate change. We might treat this source as a watchdog indicating what topics *might have been* associated with climate change over the study period. While there was much more discussion of climate change from @inside climate, its issue associations are distributed similarly to @nytimes (with an emphasis on energy, environment, and natural disasters). However, @insideclimate has much higher centrality than other sources: @inside climate associated climate change with nearly every other major topic over the study period, with the exception of *Labor*, *Housing*, and *Death Notices*.

[9] Although, again, similar results were not observed for the *Weather and Natural Disasters* topic.

Figure 4.7: Topic prelalence in tweeted climate change stories

Note that issue associations tend to correspond to sources' intended audience: in addition to strong associations with *Energy* and *Environment*, @WSJ has stronger associations with *Banking, Finance, and Domestic Commerce* while @LaOpinionLA has stronger associations with *International Affairs and Foreign Aid*. The qualitative review of retweet-weighted attention revealed that climate change attention was higher when @TheRoot tweeted stories associating climate change and racial equity; however, such stories did not register in the analysis of in-depth topic associations with the established criteria. Failure to register may share similar explanations above: these topic associations may not register because the association criteria are too strict.

4.6.5 Increasing diffusion: secondary attention to climate change

While Boolean classification using keyword hits inflates estimates of climate change attention, as illustrated in Chapters 2 and 3, keyword hits can still be informative regarding the diffusion of climate change as a secondary consideration in other topics. Examining centrality of secondary attention may help elucidate where major topics and climate change keywords co-occur in order to examine the diffusion or permeation of climate change into other important social, political and policy considerations.

Below are degree centrality estimates for climate change as a secondary topic. Centrality is estimated using the unique number of topics co-occurring with two or more hits of any of the standard climate change key terms ['climate change', 'global warm', 'greenhouse gas']. As above, the analyses below are cautious regarding classification error and so calculate centrality measures at varying co-occurrence instance thresholds (1-5) per year. For example, if a topic co-occurred with "climate change" keywords just once in a one-year period, it would be dropped in the analysis where the instance threshold was two or higher.

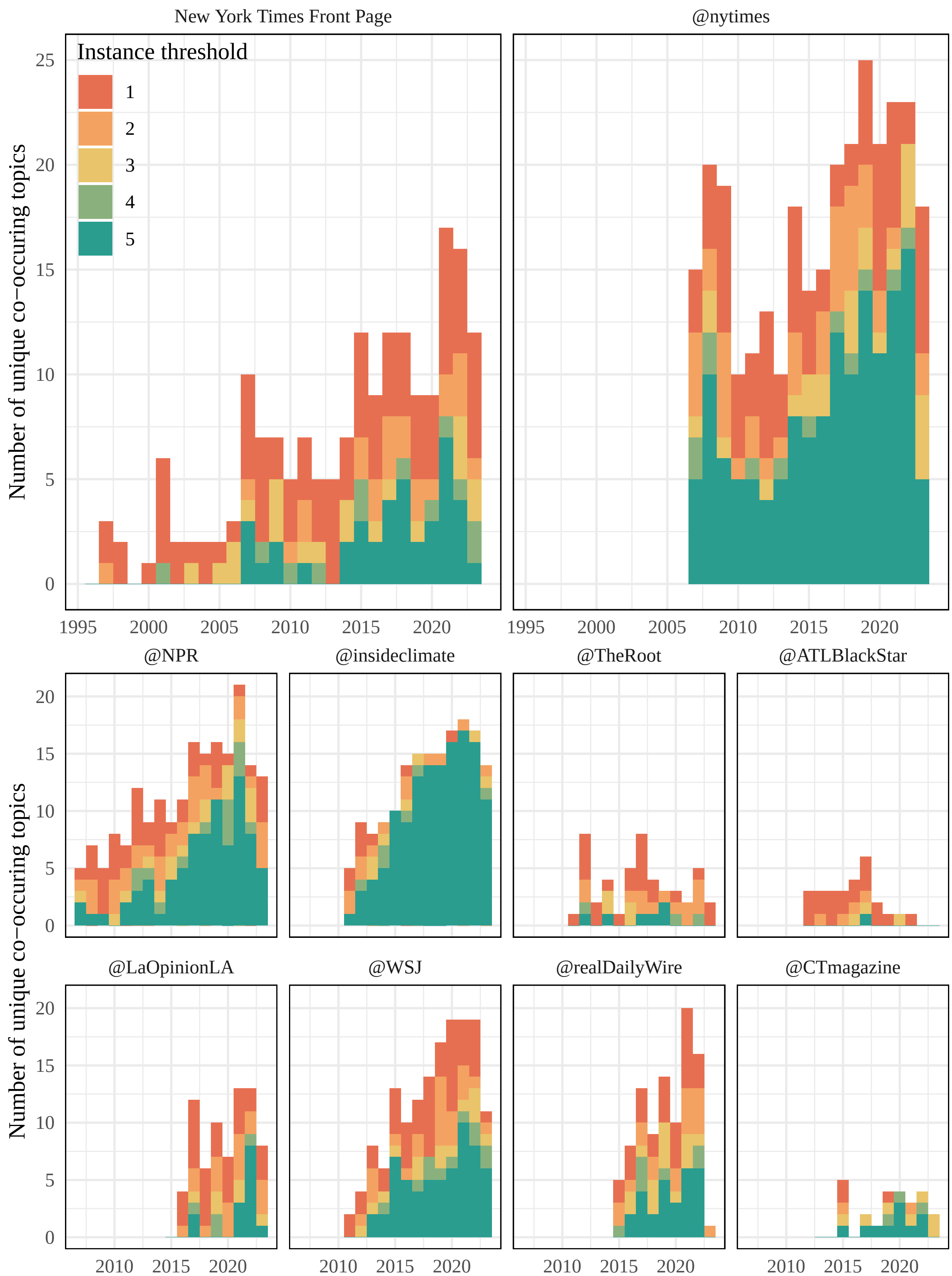

Figure 4.8: Topics associated with climate change (Monte Carlo simulations)

Similar to degree centrality, the topic-climate change co-occurrence rate appears to be increasing over time among prestige sources. Other sources appear to be using climate change keywords in other topics as well, though the signal is less clear. The expansion of secondary attention diffusion over time is clear and illustrates again the importance of proper climate change classification: what was previously treated as climate change attention through Boolean classification is secondary attention or diffusion of the issue into other topics.

Figure 4.8 illustrates that climate change is becoming a relevant consideration in the coverage of a greater number of topics over time. Longitudinal trends appear to be insensitive to the criteria set for climate change diffusion: increasing the threshold for the number of instances of co-occurrence between climate change and another topic does not often impact the trends over time (although there is a downward shift in the number of unique topics mentioning climate change in story content, the longitudinal trends are not sensitive to shifts in instance thresholds).

Hypothesis 3.3: Topic co-occurrence with climate change keywords is lower for right-leaning sources. **(Partially supported)**

4.6.6 Emerging major topic associations

Climate change is mentioned in an increasing number of major topics in U.S. news media over time. For some topics, climate change is mentioned more frequently as a proportion of all stories covering that topic over time. Figure 4.9 illustrates the major topics where climate change was mentioned at least twice in story content in at least three stories per year. Cell fills indicate the proportion of all coverage on that topic for that year that mention climate change at least twice.

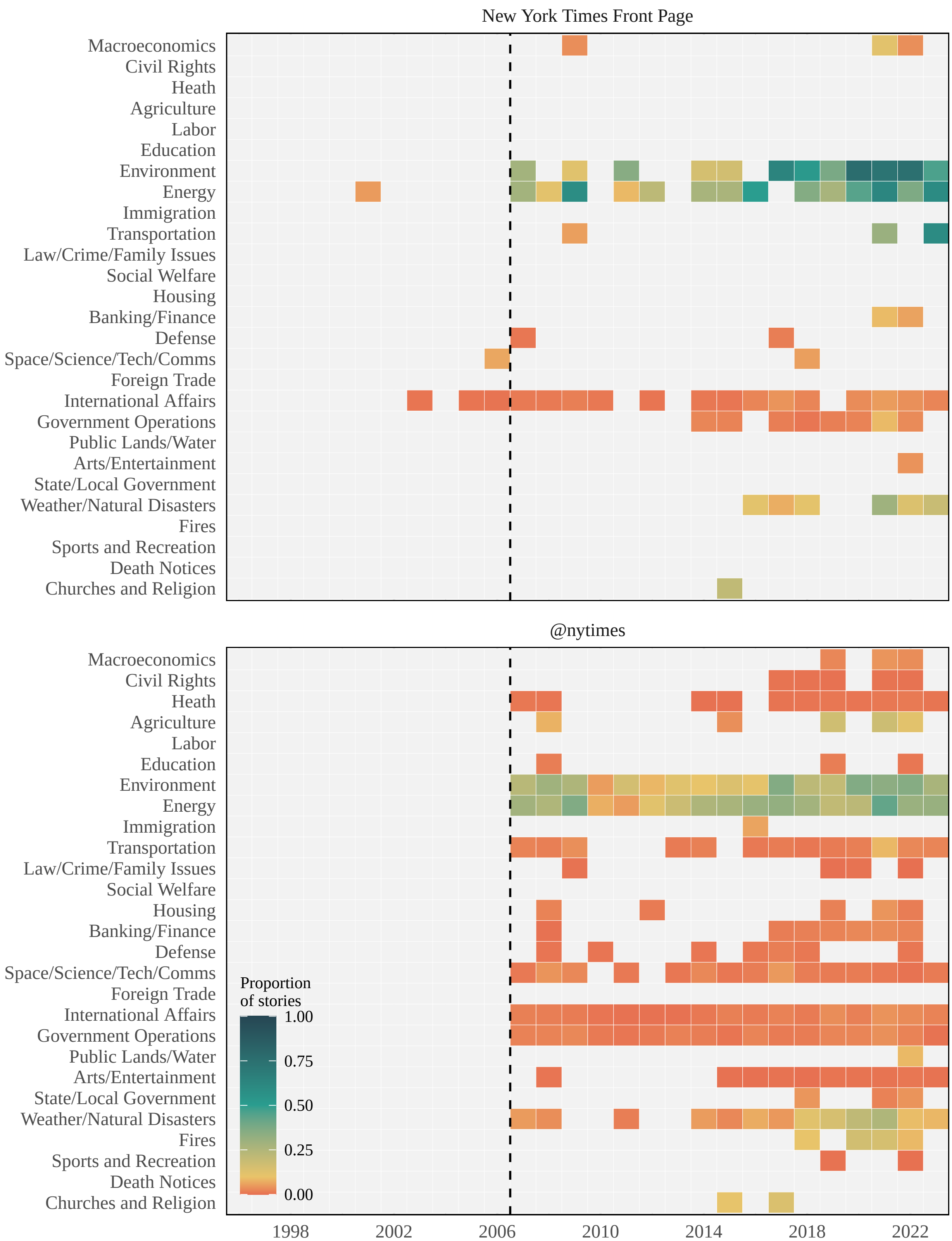

Figure 4.9: Topic associations over time

While climate change has become a relevant consideration in more major topics over time, these instances are quite rare for most topics. For example, @nytimes stories mentioned climate change in less than 10 percent of its *Transportation* stories for the full 2007-2023 period (save 2019). Interestingly, front page coverage of *Environment* and *Energy* issues were more likely to have mentioned climate change in the 2019-2022 period; assuming that the stories receiving front page coverage were attention-grabbing events, the higher rate of co-occurrence on the front page suggests that front page coverage of energy and the environment is becoming more relevant to climate change, even if it is not *primarily about* climate change.

There were extensive quality assurance checks to ensure that the observations reported here are valid. There were very few instances where major topic classifications were incorrect, which may have inflated the centrality measure. However, as discussed in the previous chapter, climate change classifications continued to perform well and, in most cases, the predicted major topic classification appeared to be correct. Consider the following excerpts:

> *Scientists Boost Crop Performance by Engineering a Better Leaf.* "Now, researchers say that by using genetic modifications to increase the efficiency of photosynthesis, they significantly increased yields in a food crop, soybeans, providing a glimmer of potential that such methods could someday put more food on tables as climate change and other threats make it harder for vulnerable populations across the globe to feed their families." @nytimes on *Agriculture*, August 18, 2022.

> *The Ticking Clock for Miami's Condo Empire.* "[...] Because we're learning so much about sea-level rise and climate change and we're realizing that a lot of our old measures are outdated." [...] He told me that he accepted the reality of climate change — he'd seen with his own eyes that the sea levels around his private dock were climbing. And he was as wary as anyone about the pace of development in Miami Beach, where, he stressed, the towers rise so high that some

residents rarely catch a glimpse of the sun. @nytimes on *Community Development and Housing Issues*, January 30, 2022.

See Appendix C for a list of New York Times examples from the year in which stories on nearly every major topic in the CAP coding scheme mentioned climate change (2022).

Similar results can be observed for other prestige news media: both @NPR and @WSJ show increasing climate change diffusion through other major topics, shown by the increasing number of filled cells in Figure 4.10 , as well as the cells indicating higher percentages over time (for certain topics). Climate change diffusion is still quite low, with several topics not mentioning climate change at all (meeting the above criteria). Aside from @insideclimate, climate change is less diffuse in non-prestige media sources, although diffusion does appear to be increasing for @LaOpinionLA.

Figure 4.10: Diversity of climate change topic associations (entropy) by theta thresholds

4.7 Discussion

In sum, climate change has become more central in discussions about a greater number of important political issues over time. Climate change is expected to impact nearly every facet of human existence and it is both alarming and reassuring that climate change coverage is diversifying in terms the major topics discussed. Additionally, many important topics are not linked to climate change in meaningful ways, particularly issues that right-leaning individuals tend to prioritize, such as immigration, labor, and economics.

CHAPTER 5

FINAL REMARKS

This research has provided new pathways to collect and analyze massive amounts of news data to evaluate climate change attention and diversity in the themes that dominate discussion of climate change. There is general consensus among climate change communications scholars that the issue must be communicated to the public in terms that are engaging to diverse audiences and in ways that resonate with personal efficacy (Bolsen & Shapiro, 2018). The research presented here suggests that U.S. news media is beginning to accomplish at least the former: attention to climate change is increasing and becoming more diverse in terms of the major topics it addresses in its coverage of climate change stories.

However, the issues that were most central to the climate change issue association network appear to be priorities for the ideological left: namely, the environment and energy. Associations with apolitical issues such as extreme weather and natural disasters are growing but issue priorities for the ideological right remain low in association centrality. Building a broad coalition of a pro-environment public will require the media to engage further in the anticipated risks of climate change as reported by the IPCC, especially the risks to issues prioritized by the right. From a strategic communications perspective, it may be beneficial to focus on issue associations that may cross pressure conservatives in terms of interests: for example, a pro-environmental agricultural lobby no longer aligning with the merchants of doubt supporting the coal, gas, and oil industries.

The automated nature of the data collection process and content analysis allows for the continual expansion of the research pipeline to include additional news sources and historical data. Future research might expand the scope of the target issue to include green energy solutions. One component that was noticeably missing from content analysis was the jobs impact debate: discussions relating to the transition from coal to green energy has tended to focus on the trade-off between protecting jobs and the protecting the environment. Media coverage of green energy developments were coded under the *Energy* topic in accordance with the CAP coding scheme. But given that green energy transitions are one of the leading ways reduce greenhouse gas emissions, it may be worthwhile to expand the scope of issue from climate change alone to climate change and its proposed solutions.

One motivation for this project was to provide a powerful and accessible tool in classification and content analysis to researchers in political communications, not merely to test the strength of these tools "out of the box" or without human intervention. As such, next steps on this project will include diagnoses of model errors and corrections to human judgement, if any. Next steps include using open source data in order to share models and increase collaboration news media analysis. Open source data sources include newsgroup (Cai *et al.*, 2009) and Common Crawl (Dodge *et al.*, 2021). Additionally, improved classification accuracy may require more in-depth and detailed tagging. For example, it may be worthwhile to tag each sentence for its topic (if any) or its frame. Using the full store of 1.1 million stories to train custom word embeddings may improve neural network model accuracy and open additional research pathways relating to interactions between topic associations and sentiment analysis.

References

Althaus S.L. & Tewksbury D. (2002). Agenda setting and the “new” news: Patterns of issue importance among readers of the paper and online versions of the new york times. Communication Research 29 (2).

Baker P. (2022). Biden’s agenda hangs in the balance if republicans take congress. New York Times. https://www.nytimes.com/2022/10/29/us/politics/bidens-agenda-midterms-congress.html.

Barberá P. & al. et (2019). Automated text classification of news articles: A practical guide. Political Analysis 1–24.

Barberá P., Casas A., Nagler J., Egan P.J., Bonneau R., Jost J.T. & Tucker J.A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. American Political Science Review 113 (4): 883–901.

Barringer F. (2006). California, taking big gamble, tries to curb greenhouse gases. New York Times. https://www.nytimes.com/2006/09/15/us/california-taking-big-gamble-tries-to-curb-greenhouse-gases.html.

Bartels B. (2008). Beyond "fixed versus random effects": A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data. The Society for Political Methodology 9: 1–43.

Bates D., Mächler M., Bolker B. & Walker S. (2014). Fitting linear mixed-effects models using lme4. arXiv 1406 (5823).

Baumgartner F.R., Boef S.L.D. & Boydstun A.E. (2008). The decline of the death penalty and the discovery of innocence. Cambridge University Press, 2008.

Bayes Robin & Druckman J.N. (2021). Motivated reasoning and climate change. Current Opinion in Behavioral Sciences 42.

Bearak M. (2023). Inside the global race to turn water into fuel. New York Times. https://www.nytimes.com/2023/03/11/climate/green-hydrogen-energy.html.

Belfer E., Ford J.D. & Maillet M. (2017). Representation of indigenous peoples in climate change reporting. Climatic Change 145 (1): 57–70.

Bell Andrew & Jones K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. Political Science Research and Methods 3 (1): 133–153.

Bell Shannon Elizabeth & York R. (2010). Community economic identity: The coal industry and ideology construction in west virginia. Rural Sociology 75 (1): 111–143.

Bennhold K. & Eddy M. (2021). Skateboards, climate change and freedom: Germany's next-generation parliament. New York Times. https://www.nytimes.com/2021/10/03/world/europe/germany-elections-young-voters-greens.html.

Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S. & Matsuo A. (2018). Quanteda: An r package for the quantitative analysis of textual data. Journal of Open Source Software 3 (30): 774.

Bischof J. & Airoldi E.M. (2012). Summarizing topical content with word frequency and exclusivity. Proceedings of the 29th International Conference on Machine Learning (ICML-12).

Blei D.M., Ng A.Y. & Jordan M.I. (2003). Latent dirichlet allocation. Journal of machine Learning research 3: 993–1022.

Bohr J. (2020). Reporting on climate change: A computational analysis of US newspapers and sources of bias, 1997–2017. Global Environmental Change 61 (102038).

Bolsen T. & Shapiro M.A. (2018). The US news media, polarization on climate change, and pathways to effective communication. Environmental Communication 12 (2): 149–163.

Borge-Holthoefer J. & Arenas A. (2010). Semantic networks: Structure and dynamics. Entropy 12 (5): 1264–1302.

Boussalis C. & Coan T.G. (2013). Signals of doubt: Text-mining climate skepticism. Workshop: London School of Economics 27.

Boussalis C. & Coan T.G. (2016). Text-mining the signals of climate change doubt. Global environmental change 36: 89–100.

Boussalis C., Coan T.G. & Holman M.R. (2018). Climate change communication from cities in the USA. Climatic Change 149 (2): 173–187.

Boussalis C., Coan T.G. & Poberezhskaya M. (2016). Measuring and modeling russian newspaper coverage of climate change. Global environmental change 41: 99–110.

Boyd-Graber J., Mimno D. & Newman D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. Handbook of mixed membership models and their applications.

Boydstun A.E. (2008). How policy issues become front-page news.

Boydstun A.E. (2013). Making the news: Politics, the media, and agenda setting. University of Chicago Press.

Boydstun A.E. (2014). New york times front page dataset.

Boydstun A.E., Bevan S. & III. H.F.T. (2014a). The importance of attention diversity and how to measure it. Policy Studies Journal 42 (2): 173–196.

Boydstun A.E., Hardy A. & Walgrave S. (2014b). Two faces of media attention: Media storm versus non-storm coverage. Political Communication 31 (4): 509–531.

Boydstun A.E. & Russell A. (2016). From crisis to stasis: Media dynamics and issue attention in the news. Oxford Research Encyclopedia of Politics.

Boykoff M.T. (2011). Who speaks for the climate?: Making sense of media reporting on climate change. Cambridge University Press.

Branigan A. (2018). On what happens when folks are denied their fried chicken. The Root. https://www.theroot.com/on-what-happens-when-folks-are-denied-their-fried-chick-1823157435.

Broad W. (1997). Another possible climate culprit: The sun. The New York Times September 23, 1997.

Broder J. (2009). With something for everyone, climate bill passed. New York Times. https://www.nytimes.com/2009/07/01/us/politics/01climate.html.

Brüggemann M. & Engesser S. (2017). Beyond false balance: How interpretive journalism shapes media coverage of climate change. Global Environmental Change 42: 58–67.

Brulle R.J. (2019). Networks of opposition: A structural analysis of US climate change countermovement coalitions 1989–2015. Sociological Inquiry.

Brulle R.J., Carmichael J. & Jenkins J.C. (2012). Shifting public opinion on climate change: An empirical assessment of factors influencing concern over climate change in the US, 2002–2010. Climatic change 114 (2): 169–188.

Budak C., Goel S. & Rao J.M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. Public Opinion Quarterly 80 (1): 250–271.

Budge I. (2002). Mapping policy preferences: 21 years of the comparative manifestos project. European Political Science 1 (3): 60–68.

Buttel Frederick H. & Flinn W.L. (1978). Social class and mass environmental beliefs: A reconsideration. Environment and Behavior 10 (3): 433–450.

Cacciatore M.A., Scheufele D.A. & Iyengar S. (2016). The end of framing as we know it… and the future of media effects. Mass communication and society 19 (1): 7–23.

Cai D., Wang X. & He X. (2009). Probabilistic dyadic data analysis with local and global consistency. Proceedings of the 26th Annual International Conference on Machine Learning 105–112.

Calvin K., Dasgupta D., Krinner G., Mukherji A., Thorne P.W., Trisos C. & Romero J. (2023). Climate change 2023: Synthesis report. Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report of the Intergovernmental Panel on Climate Change.

Carley K. & Palmquist M. (1992). Extracting, representing, and analyzing mental models. Social forces 70 (3): 601–636.

Chang J., Gerrish S., Wang C., Boyd-Graber J. & Blei D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems 22: 288–296.

Chen Y.H. & Li S.F. (2016). Using latent dirichlet allocation to improve text classification performance of support vector machine. IEEE Congress on Evolutionary Computation.

Chinn S., Hart P.S. & Soroka S. (2020). Politicization and polarization in climate change news content, 1985-2017. Science Communication 42 (1).

Clarke B., Otto F., Stuart-Smith R. & Harrington L. (2022). Extreme weather impacts of climate change: An attribution perspective. Environmental Research: Climate 1 (1).

Climate Change I.P. on (2022). Climate change 2022: Impacts, adaptation, and vulnerability. Working Group II Contribution to the Sixth Assessment Report, IPCC.

Cohen B.C. (2015). Press and foreign policy. Vol. 2321. Princeton university press.

Cortes C. & Vapnik V. (1995). Support-vector networks. Machine learning 20 (273–297).

Davenport C. (2014). Industry awakens to threat of climate change. The New York Times January 23, 2014.

Dodge J., Sap M., Marasović A., Agnew W., Ilharco G., Groeneveld D., Mitchell M. & Gardner M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv 2104.08758.

Dogru H.B. & al. et (2021). Deep learning-based classification of news texts using doc2vec model. 1st international conference on artificial intelligence and data analytics. IEEE.

Downs A. (1972). Up and down with ecology: The "issue-attention" cycle. Public Interest 28: 38–50.

Druckman J.N. (2014). Pathologies of studying public opinion, political communication, and democratic responsiveness. Political Communication 31 (3): 467–492.

Dunlap R.E. (2013). Climate change skepticism and denial: An introduction. American behavioral scientist 57 (6): 691–698.

Dunlap R.E. & Jacques P.J. (2013). Climate change denial books and conservative think tanks: Exploring the connection. American Behavioral Scientist 57 (6): 699–731.

Dunlap R.E. & Liere K.D. (1984). Commitment to the dominant social paradigm and concern for environmental quality. Social science quarterly 65 (4).

Dunlap R.E., Xiao C. & McCright A.M. (2001). Politics and environment in america: Partisan and ideological cleavages in public support for environmentalism. Environmental politics 10 (4): 23–48.

Eddy M. & Sengupta S. (2021). An electricity crisis complicates the climate crisis in europe. New York Times. https://www.nytimes.com/2021/10/29/climate/europe-energy-crisis-cop.html.

Egan P.J. & Mullin M. (2017). Climate change: US public opinion. Annual Review of Political Science 20: 209–227.

Elsasser S.W. & Dunlap R.E. (2013). Leading voices in the denier choir: Conservative columnists' dismissal of global warming and denigration of climate science. American Behavioral Scientist 57 (6): 754–776.

Entman R. (1993). Framing: Toward clarification of a fractured paradigm. Journal of communication 43 (4): 51–58.

Eshima S., Imai K. & Sasaki T. (2020). Keyword assisted topic models.

Fan R.E., Chang K.W., Hsieh C.J., Wang X.R. & Lin C.J. (2008). LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9: 1871–1874.

Feber R.E., Raebel E.M., D'cruze N., Macdonald D.W. & Baker S.E. (2017). Some animals are more equal than others: Wild animal welfare in the media. Bioscience 67 (1): 62–72.

Feldman L., Hart P.S. & Milosevic T. (2017). Polarizing news? Representations of threat and efficacy in leading US newspapers' coverage of climate change. Public Understanding of Science 26 (4): 481–497.

Feldman T.A.M. Lauren & Leiserowitz A. (2014). The mutual reinforcement of media selectivity and effects: Testing the reinforcing spirals framework in the context of global warming. Journal of Communication 64 (4): 590–611.

Flavelle C. (2021). Saving history with sandbags: Climate change threatens the smithsonian. New York Times. December 5, 2023.

Ford J.D. & King D. (2015). Coverage and framing of climate change adaptation in the media:

A review of influential north american newspapers during 1993–2013. Environmental Science & Policy 48: 137–146.

Fownes J.R., Yu C. & Margolin D.B. (2018). Twitter and climate change. Sociology Compass 12 (6).

Gene Zucker H. (1978). The variable nature of news media influence. Annals of the International Communication Association 2 (1): 225–240.

Gillis J. (2010). In weather chaos, a case for global warming. The New York Times August 14, 2010.

Gillis J. (2014). U.s. Climate has already changed, study finds, citing heat and floods. The New York Times May 7, 2014.

Gillis J. (2016). Flooding of coast, caused by global warming, has already begun. The New York Times September 3, 2016.

Gillis J. & Schwartz J. (2015). Deeper ties to corporate cash for doubtful climate researcher. The New York Times February 22, 2015.

Glenny V. & al. et (2019). A framework for streamlined statistical prediction using topic models. arXiv 1904 (6941).

Haden V.R., Niles M.T., Lubell M., Perlman J. & Jackson L.E. (2012). Haden, van r., et al. "Global and local concerns: What attitudes and beliefs motivate farmers to mitigate and adapt to climate change?." PloS one 7.12 (2012): e52882. PloS one 7 (12).

Hajer M.A. (2012). A media storm in the world risk society: Enacting scientific authority in the IPCC controversy (2009–10). Critical policy studies 6 (4): 452–464.

Hart P.S. & Nisbet E.C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. Communication research 39 (6): 701–723.

Hase V., Mahl D., Schäfer M.S. & Keller T.R. (2021). Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). Global Environmental Change 70 (102353).

Heim K. (2013). Framing the 2008 iowa democratic caucuses: Political blogs and second-level intermedia agenda setting. Journalism & Mass Communication Quarterly 90 (3): 500–519.

Helleputte T., Gramme P. & Paul J. (2022). LiblineaR: Linear predictive models based on the LIBLINEAR c/c++ library. R package version 2.10-22.

Hillard D., Purpura S. & Wilkerson J. (2007). An active learning framework for classifying political text. Annual Meeting of the Midwest Political Science Association.

Hillard D., Purpura S. & Wilkerson J. (2008). Computer-assisted topic classification for mixed-methods social science research. Journal of Information Technology & Politics 4 (4): 31–46.

Hollander B.A. (2008). Tuning out or tuning elsewhere? Partisanship, polarization, and media migration from 1998 to 2006. Journalism & Mass Communication Quarterly 85 (1): 23–40.

Holt D. & Barkemeyer R. (2012). Media coverage of sustainable development issues–attention cycles or punctuated equilibrium? Sustainable development 20 (1): 1–17.

Hoyle G. A. (2021). Is automated topic model evaluation broken? The incoherence of coherence. Advances in neural information processing systems 34: 2018–2033.

Hu M. & Liu B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Huggel C., Bouwer L., Juhola S., Mechler R., Muccione V., Orlove B. & Wallimann-Helmer I. (2022). The existential risk space of climate change. Climatic Change 174.1-2 (8).

Hvitfeldt E. & Silge J. (2021). Supervised machine learning for text analysis in r. CRC Press.

Inglehart R. (1984). Generational change and the future of the atlantic alliance. Political Science & Politics 17 (3): 525–535.

Inkpen D. & Razavi A.H. (2014). Topic classification using latent dirichlet allocation at multiple levels. International Journal of Computational Linguistics 5 (1): 43–55.

Iyengar S. (1987). Television news and citizens' explanations of national affairs. American Political Science Review 81 (3).

Iyengar S. & Simon A. (1993). News coverage of the gulf crisis and public opinion: A study of agenda-setting, priming, and framing. Communication research 20 (3).

Jacques P.J. (2012). A general theory of climate denial. Global Environmental Politics 12 (2): 9–17.

Jagarlamudi J., III H.D. & Udupa R. (2012). Incorporating lexical priors into topic models. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.

Jang S.M. & Hart P.S. (2015). Polarized frames on “climate change” and “global warming” across countries and states: Evidence from twitter big data. Global environmental change 32.

Java A., Song X., Finin T. & Tseng B. (2007). Why we twitter: An analysis of a microblogging community. International Workshop on Social Network Mining and Analysis 118–138.

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features. European conference on machine learning. Springer, Berlin, Heidelberg.

Johns L.N. & Jacquet J. (2018). Doom and gloom versus optimism: An assessment of ocean-related US science journalism (2001-2015). Global Environmental Change 50: 142–148.

Jones B.D. & Baumgartner F.R. (2005a). The politics of attention: How government prioritizes problems. University of Chicago Press.

Jones B.D. & Baumgartner F.R. (2005b). The polittics of attention: How government prioritizes problems. University of Chicago Press.

Jones B.D., Baumgartner F.R., Theriault S.M., Epp D.A., Lee C., Sullivan M.E. & Cassella C. (2023). Policy agendas project: codebook.

Jones B.D., Larsen-Price H. & Wilkerson J. (2009). Representation and american governing institutions. The Journal of Politics 1: 277–290.

Jones T., Knapp B. & Park B. (2021). Tidylda: Latent dirichlet allocation using 'tidyverse' conventions.

Joyce C. (2016). Scientists report the planet was hotter than ever in the first half of 2016. National Public Radio. https://www.npr.org/sections/thetwo-way/2016/07/19/486655867/scientists-report-the-planet-was-hotter-than-ever-in-the-first-half-of-2016.

Keller T.R., Hase V., Thaker J., Mahl D. & Schäfer M.S. (2020). News media coverage of climate change in india 1997–2016: Using automated content analysis to assess themes and topics. Environmental Communication 14 (2): 219–235.

Kim J.H., Su T.-Y. & Hong J. (2007). The influence of geopolitics and foreign policy on the US and canadian media: An analysis of newspaper coverage of sudan's darfur conflict. Harvard International Journal of Press/Politics 12 (3): 87–95.

King G., Lam P. & Roberts M.E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science 61 (4): 971–988.

Kompas T., Pham V.H. & Che T.N. (2018). The effects of climate change on GDP by country and the global economic gains from complying with the paris climate accord. Earth's Future 6 (8): 1153–1173.

Krauss C., Myers S. & Revkin A. (2005). As polar ice turns to water, dreams of treasure abound. The New York Times October 10, 2005.

Krayenhoff E.S., Moustaoui M., Broadbent A.M., Gupta V. & Georgescu M. (2018). Diurnal interaction between urban expansion, climate change and adaptation in US cities. Nature Climate Change 8 (12): 1097–1103.

Kuhn M. & Johnson K. (2013). Applied predictive modeling. Vol. 26. New York: Springer.

Kwak H., Lee C., Park H. & Moon S. (2010). What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web 591–600.

Lakshminarayanan Balaji & Raich R. (2011). Inference in supervised latent dirichlet allocation. International Workshop on Machine Learning for Signal Processing.

Lambert J. (2019). Study finds racial gap between who causes air pollution and who breathes it. National Public Radio. https://www.npr.org/sections/health-shots/2019/03/11/702348935/study-finds-racial-gap-between-who-causes-air-pollution-and-who-breathes-it.

Lawrence R.G. & Boydstun A.E. (2017). Celebrities as political actors and entertainment as political media. How political actors use the media: A functional analysis of the media's role in politics 39–61.

Leiserowitz A.A., Maibach E.W., Roser-Renouf C., Smith N. & Dawson E. (2013). Climategate, public opinion, and the loss of trust. American behavioral scientist 57 (6): 818–837.

Leiserowitz A., Roser-Renouf C., Marlon J. & Maibach E. (2021). Global warming's six americas: A review and recommendations for climate change communication. Current Opinion in Behavioral Sciences 42: 97–103.

Liu X., Lindquist E. & Vedlitz A. (2011). Explaining media and congressional attention to global climate change, 1969-2005: An empirical test of agenda-setting theory. Political Research Quarterly 64 (2): 405–419.

Luo M. & Goodstein L. (2007). Emphasis shifts for new breed of evangelicals. New York Times. https://www.nytimes.com/2007/05/21/us/21evangelical.html.

Manning C., Raghavan P. & Schutze H. (2009). An introduction to information retrieval. In: Cambridge University Press. Cambridge University Press.

Maowen W., Dong Z.C., Weiyao L. & Qiang W.Q. (2012). Text topic mining based on LDA and co-occurrence theory. 7th International Conference on Computer Science & Education (ICCSE) (525–528).

Mcauliffe J. & Blei D. (2007). Supervised topic models. Advances in neural information processing systems 20: 121–128.

McComas Katherine & Shanahan J. (1999). Telling stories about global climate change: Measuring the impact of narratives on issue cycles. Communication research 26 (1): 30–57.

McCombs Maxwell & Valenzuela S. (2020). John wiley & sons. Setting the agenda: Mass media; public opinion.

McCombs M.E. & Shaw. D.L. (1972). The agenda-setting function of mass media. Public opinion quarterly 36 (2): 176–187.

McCombs M. & Zhu J.-H. (1995). Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. Public Opinion Quarterly 59 (4): 495–525.

McCright A.M. & Dunlap R.E. (2003). Defeating kyoto: The conservative movement's impact on US climate change policy. Social problems 50 (3): 348–373.

McCright A.M. & Dunlap R.E. (2011). The politicization of climate change and polarization in the american public's views of global warming, 2001–2010. The Sociological Quarterly 52 (2).

McNeil D. (2018). Tick and mosquito infections spreading rapidly, c.d.c. finds. New York Times. https://www.nytimes.com/2018/05/01/health/ticks-mosquitoes-diseases.html.

Meng Y. & al. et (2019). Weakly-supervised hierarchical text classification. Vol. 33. Proceedings of the AAAI conference on artificial intelligence.

Meraz S. (2011). Using time series analysis to measure intermedia agenda-setting influence in traditional media and political blog networks. Journalism & mass communication quarterly 88 (1): 176–194.

Mitchel A. (1997). G.o.p. Hopes climate fight echoes health care outcome. New York Times. https://www.nytimes.com/1997/12/13/world/gop-hopes-climate-fight-echoes-health-care-outcome.html.

Mitchell D., Heaviside C., Vardoulakis S., Huntingford C., Masato G., Guillod B.P. & Allen M. (2016). Attributing human mortality during extreme heat waves to anthropogenic climate change. Environmental Research Letters 11 (7).

Miyato T., Dai A.M. & Goodfellow I. (2016). Adversarial training methods for semi-supervised text classification. arXiv 1605 (7725).

Morrison D.E. & Dunlap R.E. (1986). Environmentalism and elitism: A conceptual and empirical analysis. Environmental management 10 (5): 581–589.

Myers S., Revkin A. & Romero S. (2005). Old ways of life are fading as the arctic thaws. New York Times. https://www.nytimes.com/2005/10/20/science/earth/old-ways-of-life-are-fading-as-the-arctic-thaws.html.

Myers T.A., Maibach E.W., Roser-Renouf C., Akerlof K. & Leiserowitz A.A. (2013). The relationship between personal experience and belief in the reality of global warming. Nature climate change 3 (4): 343–347.

Myerson G. & Rydin Y. (s.d.). Sustainable development: The implications of the global debate for land use planning. Environmental planning and Sustainability 1996: 19–34.

Naik G. (2013). Rising sea level tied to faster melt. Wall Street Journal. https://www.wsj.com/articles/SB10001424127887324682204578517312171788462.

Nassif R. & Fahkr M.W. (2019). Supervised topic modeling using word embedding with machine learning techniques. International conference on advances in the emerging computing technologies. IEEE.

Oreskes N. & Conway E.M. (2010). Defeating the merchants of doubt. Nature 465 (7299): 686–687.

Ortiz-Bobea A., Knippenberg E. & Chambers R.G. (2018). Growing climatic sensitivity of US agriculture linked to technological change and regional specialization. Science advances 4 (12).

Pan S. & Ding T. (2019). Social media-based user embedding: A literature review. arXiv 1907 (725).

Peluso N.L. & Harwell E. (2001). Territory, custom, and the cultural politics. Violent Environ 83.

Pennington J., Socher R. & Manning C.D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (1532–1543).

Popovich N. & Migliozzi B. (2018). Where are america's winters warming the most? In cold places. New York Times. https://www.nytimes.com/interactive/2018/03/16/climate/us-winter-warming.html.

Purpura S. & Hillard D. (2006). Automated classification of congressional legislation.

Purpura S., Wilkerson J. & Hillard D. (2008). The US policy agenda legislation corpus volume 1-a language resource from 1947-1998. LREC.

Reber U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. Communication methods and measures 13 (2): 102–125.

Revkin A. (s.d.). Warming's effects to be widespread. The New York Times June 12, 2000.

Rickard L.N., Yang Z.J. & Schuldt J.P. (2016). Here and now, there and then: How "departure dates" influence climate change engagement. Global Environmental Change 38: 97–107.

Roberts M.E., Stewart B.M. & Tingley D. (2016). Navigating the local modes of big data. Computational social science 51.

Sandhu M., Vinson C.D., Mago V.K. & Giabbanelli P.J. (2019). From associations to sarcasm: Mining the shift of opinions regarding the supreme court on twitter. Online Social Networks and Media 14 (100054).

Sanger D. & Bumiller E. (2004). Bush dismisses idea that kerry lied on vietnam. New York Times. https://www.nytimes.com/2004/08/27/us/the-2004-campaign-the-president-bush-dismisses-idea-that-kerry-lied-on-vietnam.html.

Saunders C., Grasso M.T. & Hedges C. (2018). Attention to climate change in british newspapers in three attention cycles (1997–2017). Geoforum 94: 94–102.

Saveski M., Beeferman D., McClure D. & Roy D. (2022). Engaging politically diverse audiences on social media. Proceedings of the International AAAI Conference on Web and Social Media 16: 873–884.

Scannell L. & Gifford R. (2013). Personally relevant climate change: The role of place attachment and local versus global message framing in engagement. Environment and Behavior 45 (1): 60–85.

Schäfer M.S., Ivanova A. & Schmidt A. (2014). What drives media attention for climate change? Explaining issue attention in australian, german and indian print media from 1996 to 2010. International Communication Gazette 76 (2): 152–176.

Scharkow M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using german online news. Quality and Quantity 47 (2): 761–773.

Schattschneider E.E. (1960). The semisovereign people: A realist's view of democracy in america. Holt, Rinehart; Winston.

Schlosberg D. & Collins L.B. (2014). From environmental to climate justice: Climate change and the discourse of environmental justice. Wiley Interdisciplinary Reviews: Climate Change 5 (3): 359–374.

Shanahan J. (1996). Green but unseen: Marginalizing the environment on television. Contributions in Political Science 367: 176–193.

Shim J., Park C. & Wilding M. (2015). Identifying policy frames through semantic network analysis: An examination of nuclear energy policy across six countries. Policy Sciences 48: 51–83.

Silla C.N. & Freitas A.A. (2011). A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery 22 (1–2): 31–72.

Spence A., W. Poortinga W. & Pidgeon N. (2012). The psychological distance of climate change. Risk Analysis: An International Journal 32 (6): 957–972.

Stern S., Livan G. & Smith R.E. (2020). A network perspective on intermedia agenda-setting. Applied Network Science 5: 1–22.

Stevens W. (1997). Doubts on cost are bedeviling climate policy. The New York Times October 6, 1997.

Stone P.J. & Hunt E.B. (1963). A computer approach to content analysis: Studies using the general inquirer system. Proceedings of the May 21-23, 1963, spring joint computer conference.

Stroud N.J. (2011). Niche news: The politics of news choice. Oxford University Press.

Sweetser G.J.G. Kaye D. & Wanta W. (2008). Intermedia agenda setting in television, advertising, and blogs during the 2004 election. Mass Communication & Society 11 (2): 197–216.

Swim J.K., Gillis A.J. & Hamaty K.J. (2020). Gender bending and gender conformity: The social consequences of engaging in feminine and masculine pro-environmental behaviors. Sex Roles 82 (5): 363–385.

Takeshita T. (2006). Current critical problems in agenda-setting research. International Journal of Public Opinion Research 18 (3): 275–296.

Tankersley J., Rogers K. & Friedman L. (2021). With methane and forest deals, climate summit offers hope after gloomy start. New York Times. https://www.nytimes.com/2021/11/02/world/europe/climate-summit-methane-forests.html.

Tankersley J., Sanger D. & Fassihi F. (2022). Biden condemns russia as threat to the world in u.n. speech. New York Times. https://www.nytimes.com/2022/09/21/world/europe/biden-russia-united-nations.html.

Taylor D.E. (1997). American environmentalism: The role of race, class and gender in shaping activism 1820-1995. Race, Gender & Class 16–62.

Taylor D.E. (2002). Race, class, gender, and american environmentalism. US Department of Agriculture, Forest Service, Pacific Northwest Research Station 534.

Troianovski A. (2021). On a pacific island, russia tests its battle plan for climate change. New York Times Anton Troianovski. https://www.nytimes.com/2021/10/19/world/europe/russia-climate-change.html.

Turner J.M. (2009). The specter of environmentalism: Wilderness, environmental politics, and the evolution of the new right. Journal of American History 96 (1): 123–148.

Tversky A. & Kahneman D. (1981). The framing of decisions and the psychology of choice. Science 211 (4481): 453–458.

Van Aelst P. & Walgrave S. (2016). Political agenda setting by the mass media: Ten years of research, 2005–2015. Handbook of public policy agenda setting.

Vargo C.J. & Guo L. (2017). Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online US news. Journalism and Mass Communication Quarterly 94 (4): 1031–1055.

Verberne S., D'hondt E., Bosch A.V. den & Marx M. (2014). Automatic thematic classification of election manifestos. Information Processing and Management 50 (4): 554–567.

Vu H.T., Liu Y. & Tran D.V. (2019). Nationalizing a global phenomenon: A study of how the press in 45 countries and territories portrays climate change. Global Environmental Change 58 (101942).

Wagner P. & Payne D. (2017). Trends, frames and discourse networks: Analysing the coverage of climate change in irish newspapers. Irish Journal of Sociology 25 (1): 5–28.

Walgrave A.E.B. Stefaan & Hardy A. (2017). The nonlinear effect of information on political attention: Media storms and US congressional hearings. Political Communication 34 (4): 548–570.

Walgrave S. & Vliegenthart R. (2010). Why are policy agendas punctuated? Friction and cascading in parliament and mass media in belgium. Journal of European Public Policy 17 (8): 1147–1170.

Weaver D., McCombs M. & Shaw D.L. (2004). Agenda-setting research: Issues, attributes, and influences. Handbook of political communication research 257: 257–80.

Yan Y. & Zheng K. (2020). Text classification model based on multi-level topic feature extraction. International conference on computer and communications. IEEE.

Zaller J. (1992). The nature and origins of mass opinion. Cambridge university press.

Zanocco C., Flora J. & Boudet H. (2022). Disparities in self-reported extreme weather impacts by race, ethnicity, and income in the united states. PLOS Climate 1 (6).

Zhao J.J.C. Weizhong & Zou W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. BMC bioinformatics 16 (1–10).

APPENDIX A

PACKAGES AND SCRIPTS

A.1 R Packages

Methods used in this report and supporting appendices (and their associated R packages) include:

- Data wrangling (rvest, polite, httr, selenium): There are several free packages in R and Python to help researchers collect publicly available data on the internet. Below
- Data management and storage (RSQLite, sqlite3): For research pipelines requiring continual data collection and analysis, managing data storage through a SQL database is highly recommended.
- Naive Bayes (quanteda): Typically included as a base-line model for its simplicity, speed and moderate predictive performance. This approach scores the most likely topic according to the joint probability of terms in a corpus appearing in a text of a given topic.
- Support Vector Machines (LiblineaR): SVMs find hyperplanes among predictors (such as terms or representations of terms) that best separate classes.
- Latent Dirichlet Allocation (tidylda): An unsupervised machine learning topic modeling technique that models a user-specified number of latent topics in a collection of texts. The model works by finding the term-topic probabilities that best fit the co-occurrence of terms in documents in the corpus. It begins with "flat" priors, meaning that all term-topic probabilities start at equal values.
- Seeded (or keyword-assisted) LDA (tidylda): Enables researchers to incorporate prior knowledge about the terms and topics in a corpus. Assigning term-topic priors imposes a topic structure on the data, thereby improving model fit and efficiency, easing topic interpretability, and obviating the need to assign post-hoc labels to topics.

- Global Vector Embedding (GloVe): Numerical representations for semantic meanings of features and phrases, pre-trained from Wikipedia articles using 300 dimensions. Term embeddings provide additional semantic information about terms given the local context of those terms; this method also helps to alleviate issues with synonyms and homonyms in natural language processing.
- Long Short-Term Memory Recurrent Neural Network (keras): Identifies relevant information in a text when predicting the target variable and uses information "gates" to consider relevant information and "forget" irrelevant information in the prediction task.
- Convolutional Neural Network (keras): Simplifies small blocks of information (e.g., a sequence of terms in a text, a block of pixels in an image) and identifies patterns among those simplifications that are relevant in a given task, such as text classification.

A.2 Scripts

This report utilized a novel database compiled as part of a larger project to continuously scrape and analyze news media content and metadata. This appendix details the programs constructed for the compilation of current and historic data from Twitter and various news sources.

Data collection procedures for this project consisted of three programs built in Python: 1) a program built to retrieve recent tweets using the Twitter API, 2) an extension of the first program that follows external links tweeted and retrieves linked content, and 3) a fail safe program that searches for missing stories using Wayback Machine, a nonprofit organization that maintains archives of internet postings long after they have been archived, relocated, or deleted by the original authors.

A.3 Usage of Twitter API

Most stories retrieved from Twitter during the January 2019 through April 2023 data collection period made use of Twitter's open source developer API through the *tweepy* module in Python. For much of this period, the Twitter developer API was open source with reasonable restrictions on the number of tweets that could be retrieved in one request (2,000 most recent tweets from a specified handle). Following the October 2022 purchase of Twitter, more restrictions were put into place, such as a cap on the total number of tweets that could be retrieved through the API in a 24 hour period. In 2023, Twitter began charging for retrieval of tweets through its API, reducing the platform's potential for academic research.

Should readers desire to use Twitter's paid API service, it is simple to integrate *tweepy* with SQL databases for frequent or periodic data supplementation.

Tweet retrieval program

```
# Name of Twitter API app
appname = "appname"

## API key
key = "example_key"

## API secret
secret = "example_secret"

## API access token and access secret
access_token = "example_token"
access_secret = "example_secret"

## Import relevant packages
import tweepy
import json
import pandas as pd
```

```python
import time
import re
import sqlite3
from dateutil import parser
from datetime import datetime
import pyarrow.feather as feather

## OAuth process, using the keys and tokens
auth = tweepy.OAuthHandler(key, secret)
auth.set_access_token(access_token, access_secret)

## Creation of the actual interface, using authentication
api = tweepy.API(auth)

## Connect to SQL database
## Assuming a SQL database has been created and saved in
## 'path/to/SQLdatabase.db'
con = sqlite3.connect("path/to/SQLdatabase.db")
cur = con.cursor()

## Subset status IDs
## Assuming the SQLdatabase.db contains at least columns 'source'
## (for news source) and 'status_id' for unique tweet identification
ids = pd.read_sql_query("SELECT source, status_id from example_database",
con)

## Create list of sources to search
```

```python
users = ['@nytimes']

## Call API for timeline of recent tweets, by user
for user in users:
    timeline = tweepy.Cursor(
    api.user_timeline,
    since_id = ids['status_id'][ids['source']==user].max(),
    screen_name = user,
    include_rts = False, # Remove retweets
    exclude_replies = True, # Remove replies
    tweet_mode="extended").items(1200) # Get 1,200 most recent tweets

## For each timeline, retrieve relevant data and meta data for storage in SQL
    for t in timeline:

        ## Convert to string
        json_str = json.dumps(t._json)

        ## Deserialise string into python object
        parsed = json.loads(json_str)

        ## If status_id not already retrieved, process for appending to SQL
        if parsed['id'] not in ids.status_id.values:

            tempuser = user
            tempid = parsed['id']
            tempdate = ' '.join([re.split(' [0-9]{2}:',
```

```
parsed['created_at'])[0],
parsed['created_at'].split(' ')[-1]])
tempdate = time.mktime(
datetime.strptime(
tempdate, '%a %b %d %Y').timetuple())
tempfavorite = parsed['favorite_count']
tempretweet = parsed['retweet_count']
temptweet = parsed['full_text']
try:
    templink = parsed['entities']['urls'][0]['expanded_url']
except:
    templink = ''
    pass
tempreplyuser = parsed['in_reply_to_user_id']
tempreplystatus = parsed['in_reply_to_status_id']
try:
    temphash = ";;".join(parsed['entities']['hashtags'])
except:
    temphash = str(parsed['entities']['hashtags'])
try:
    tempmentions = ";;".join(parsed['entities']['user_mentions'])
except:
    tempmentions = str(parsed['entities']['user_mentions'])

cells = (tempuser, tempid, tempdate, tempfavorite,
tempretweet, temptweet, templink, tempreplyuser,
tempreplystatus, temphash, tempmentions)
```

```
            cur.execute("INSERT INTO twitterspace (source, status_id,
            created_at, favorite_count, retweet_count, tweet, templink,
            in_reply_to_user_id, in_reply_to_status_id, hashtags,
            user_mentions) VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)",
            cells)

# Save to SQL
con.commit()
```

A.4 Historical Twitter Data

Twitter's open source API capped the number of tweets that could be retrieved per call at 2,000. As some sources tweeted more than 2,000 times per month, it was necessary to build a custom data retrieval program to collect older tweets. However, the 2022 Musk purchase also came with restrictions on the number of tweets that users can view in a 24 hour period. At the time of this report, that value is set at one thousand tweets for unverified users and ten thousand for verified users paying a monthly subscription fee.[1]

Program logic and sample code

Endless scrolling presents an obstacle for webscraping. During the period of data collection, Twitter operated using endless scrolling: users may continue to scroll down to load additional tweets. To prevent browsers from becoming burdensome, older tweets are offloaded at the top of the page as newer tweets are loaded at the bottom. This presents a scraping challenge as web elements become visible and invisible.

The endless scrolling problem was addressed with the following logic: scrape all visible tweets, scroll down the page by a specified number of pixels to load older tweets, and then scrape all visible tweets. This process is repeated until the scroll height is unchanged (i.e., until the screen cannot scroll down any further). This process necessarily produces duplicated rows, which are removed as a final step.

The logic for scraping the tweets from each user per day is as follows:

```
Scrape tweet data (tweet id, content, meta data, and external link)
Scroll px pixels down where px = height of screen
Scrape and append tweets to data frame
Repeat until scroll height unchanged
Remove duplicates
```

[1] It should be noted that waiting features were built into these programs to slow the web crawl. The scraped websites are under a constant barrage of bots without built-in procedures to slow their requests; this barrage of search requests or scrolling puts an undue stress on website servers. This was mitigated to the extent possible.

Dates and users

The process described above was repeated for every date since the source became a Twitter use until present. For each source (user) and for each day in the study period, the scraping program built a custom search phrase that consisted of the user, the search date, and the search date plus one day. In this way, the program conducted a search for tweets from each user one day at a time, from 2007 to the present.

Twitterscrape program logic

The logic of the Twitter scraping program is as follows:

```
For users (u) ∈ U:
    For date (x) ∈ X:
    Search twitter user from x to x + 1:
    Endless scrolling program (detailed above)
```

The second phase of the program used tweeted links to scrape sources' news media. For each valid link tweeted, the second program followed the link, collected visible data (URL, headline, body content, author) as well as meta data (keywords, article descriptions).

```
For each valid link in k ∈ K:
    Follow link
    Collect:
        Headline
        URL
        Content
        Metadata
```

Finally, for each link that was inactive or could not be found, the third program searched for the tweeted link on Wayback Machine. In most cases, Wayback Machine had archived the original link and hosted the content and metadata, which was available for scraping.

```
def newsearch(x):
    global inputElement
    inputElement = driver.find_element_by_css_selector('input')
    inputElement.send_keys(Keys.COMMAND,"a")
    inputElement.send_keys(Keys.DELETE)
    inputElement.send_keys(x)
    inputElement.send_keys(Keys.ENTER)


def wait(x):
    time.sleep(x)
    sys.stdout.flush()


def newtweets():
    tweets = driver.find_elements(by=By.XPATH, value='.//*[@role="article"]')
    cols = ['tweet','date','hour','tweetlink','templink','likes','rts','comments']
    tweetrow = pd.DataFrame(columns=cols, index=range(0,len(tweets)))
    for i in range(0,len(tweets)):
        try:
            tweetrow.loc[i, 'tweet'] = tweets[i].find_element(by=By.XPATH,
            value='.//div/div/div/div[2]/div[2]/div[2]/div').text
        except:
            print("tweetrow exception")
            pass
        tweetlinks = tweets[i].find_elements(by=By.XPATH, value='.//a')
        for tweetlink in tweetlinks:
            if '/status/' in tweetlink.get_attribute('href'):
```

```
                tweetrow.loc[i, 'tweetlink'] = tweetlink.get_attribute('href')
            if 't.co' in tweetlink.get_attribute('href'):
                tweetrow.loc[i, 'templink'] = tweetlink.get_attribute('href')
        tweetrow.loc[i, 'date'] = tweets[i].find_element(by=By.CSS_SELECTOR,
        value='time').text
        tweetrow.loc[i, 'hour'] = tweets[i].find_element(by=By.CSS_SELECTOR,
          value='time').get_attribute('datetime')
        tweetrow.loc[i, 'comments'] = re.sub('[^\d]', '',
        tweets[i].find_element(by=By.XPATH,
        value='.//*[@data-testid="reply"]').get_attribute('aria-label'))
        tweetrow.loc[i, 'rts'] = re.sub('[^\d]', '',
        tweets[i].find_element(by=By.XPATH,
        value='.//*[@data-testid="retweet"]').get_attribute('aria-label'))
        tweetrow.loc[i, 'likes'] = re.sub('[^\d]', '',
        tweets[i].find_element(by=By.XPATH,
        value='.//*[@data-testid="like"]').get_attribute('aria-label'))
    global tweetset
    tweetset = tweetset.append(tweetrow).drop_duplicates()

def twitterscrape(source,entry,exit):
    global tweetset
    cols = ['tweet','date','hour','tweetlink','templink',
    'likes','rts','comments']
    tweetset = pd.DataFrame(columns = cols)
    x = dates[dates.since.str.contains(entry)].index[0]
    print(x)
    y = dates[dates.until.str.contains(exit)].index[0]
```

```
print(y)
global z
for i in range(x,y):
    z = i
    df_length = len(tweetset)
    try:
        searchterm = "from:" + source + " since:" +
        dates.since[i] + " until:" + dates.until[i]
        wait(2)
        newsearch(searchterm)
        newtweets()
        last_height = 0
        q = True
        while q:
            newtweets()
            driver.execute_script("window.scrollBy(0, 3200);")
            wait(2)
            newtweets()
            new_height = driver.execute_script('return window.pageYOffset;')
            if new_height == last_height:
                q = False
            last_height = new_height
    except:
        if len(tweetset) == df_length:
            saveterm = source + "_" + entry + "_" + exit + ".xlsx"
            pd.DataFrame(tweetset).to_excel(saveterm, index = False)
            print('Part of ' + saveterm + ' saved: stopped at ' + dates.until[i])
```

```
                print(dates.until[i])
                break
            else:
                pass
    saveterm = source + "_" + entry + "_" + dates.until[i] + ".xlsx"
    pd.DataFrame(tweetset).to_excel(saveterm, index = False)
    print(saveterm + ' completed and saved')


twitterscrape(source="nytimes", entry="2007-06-04", exit="2023-04-04")
```

The program to collect historical tweets was built to control a web browser and collect its data. After collecting historical data, data collection efforts were migrated to Twitter's previously open source API. Most of the data collected for these analyses lay outside the reach of open source API or developer tools.

As of the date of this manuscript, the Twitter developer API is no longer open source. The number of tweets that can be viewed in a day is also capped as of late 2023, meaning that Twitter is no longer an viable source of open data for academic research.

APPENDIX B

GUIDED LDA COMPARATIVE AGENDAS PROJECT TOPIC STRUCTURE

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors

Major Topic	Minor Topic	Term	Beta	Prior
Macroeconomics	General	money	0.138	1
		fund	0.119	1
		price	0.106	1
		stock	0.050	1
		profit	0.038	1
	Inflation	inflation	0.056	1
		central bank	0.031	1
		short term	0.026	1
		interest rate	0.026	1
		federal reserve	0.016	1
	Unemployment	job	0.248	1
		unemployment	0.025	1
		unemployment rate	0.014	1
		work force	0.009	1
		labor market	0.009	1
	Treasury	treasury	0.026	1
		treasury secretary	0.012	1
		treasury department	0.012	1
		federal reserve	0.008	1
		monetary policy	0.006	1
	Budget	budget	0.110	1
		amy	0.018	
		surplus	0.008	1
		budget deficit	0.006	1
		congressional budget	0.005	1
	Tax Reform	tax	0.155	1
		tax cut	0.024	1
		tax credit	0.012	1
		tax rate	0.011	1
		tax break	0.009	1
	Industrial Policy	industry	0.176	1
		product	0.108	1
		growth	0.081	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		case	0.002	
		need	0.002	
	Theory	communist	0.020	1
		socialist	0.016	1
		capitalism	0.009	1
		jump	0.009	
		communism	0.005	1
Civil Rights, Minority Issues, and Civil Liberties	General	poster	0.015	
		amendment right	0.006	1
		civil liberty	0.006	1
		salvage	0.006	
		need	0.003	
	Ethnic Minority and Racial Group Discrimination	african american	0.070	1
		racist	0.052	1
		racism	0.042	1
		civil right	0.035	1
		slavery	0.021	1
	Gender and Sexual Orientation Discrimination	female	0.064	1
		gay	0.042	1
		gender	0.041	1
		feminist	0.007	1
		lesbian	0.006	1
	Age Discrimination	duo	0.006	
		reign	0.004	
		mistreatment	0.003	1
		case	0.003	
		need	0.003	
	Handicap or Disease Discrimination	disability	0.024	1
		mental	0.019	1
		disable	0.018	1
		wheelchair	0.007	1
		need	0.002	
	Voting Rights and Issues	vote right act	0.005	1
		right vote	0.005	1
		redraw	0.004	1
		disenfranchise	0.004	1
		redistricting	0.004	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Freedom of Speech and Religion	freedom	0.074	1
		religious	0.015	1
		religious freedom	0.014	1
		free speech	0.013	1
		religion	0.011	1
	Right to Privacy and Access to Government Information	abortion	0.091	1
		v wade	0.006	1
		roe v wade	0.006	1
		fetus	0.005	1
		ban abortion	0.004	1
	Death with Dignity	move	0.210	
		limit	0.092	
		fear	0.089	
		generation	0.048	
		martin	0.030	
	Surveillance	privacy	0.041	1
		surveillance	0.040	1
		warrant	0.024	1
		nsa	0.011	1
		security agency	0.007	1
	Eminent Domain	expert	0.128	
		producer	0.048	
		provision	0.031	
		house republican	0.016	
		contribute	0.013	
	Human Rights and DetainRights of Detainees on U.S. Soil	torture	0.025	1
		detainee	0.024	1
		interrogation	0.013	1
		geneva	0.009	1
		guantanamo	0.004	1
Health	General	patient	0.103	1
		health	0.084	1
		doctor	0.081	1
		dr	0.074	1
		medical	0.068	1
	Comprehensive Health Care Reform	medicaid	0.027	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		health insurance	0.025	1
		affordable care act	0.021	1
		obamacare	0.015	1
		health care system	0.009	1
	Insurance Reform, Availability, and Cost	medicare	0.040	1
		premium	0.029	1
		medicaid	0.003	1
		need	0.002	
		case	0.002	
	Regulation of Drug Industry, Medical Devisces, and Clinical Labs	fda	0.031	1
		murphy	0.018	
		drug administration	0.013	1
		drug company	0.007	1
		need	0.003	
	Facilities Construction, Regulation, and Payments	hospital	0.185	1
		clinic	0.036	1
		medical center	0.015	1
		emergency room	0.010	1
		need	0.002	
	Provider and Insurer Payment Regulation	insurance	0.039	1
		insurance company	0.013	1
		traitor	0.004	
		need	0.003	
		case	0.002	
	Medical Liability, Fraud, and Abuse	service company	0.005	
		situate	0.003	
		need	0.003	
		case	0.002	
		back	0.002	
	Health Manpower and Training	nurse	0.045	1
		reinvigorate	0.003	
		case	0.003	
		need	0.003	
		back	0.002	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Prevention, Communicable Diseases, and Health Promotion	covid	0.052	1
		virus	0.041	1
		disease	0.031	1
		cancer	0.027	1
		vaccine	0.022	1
	Infants and Children	child	0.246	1
		baby	0.051	1
		mother	0.035	1
		birth	0.028	1
		mom	0.028	1
	Mental Health and Mental Retardation	brain	0.058	1
		behavior	0.040	1
		mental health	0.029	1
		depression	0.019	1
		psychiatric	0.008	1
	Long-Term Care, Home Health, Terminally Ill, and Rehabilitation Services	propose	0.100	
		dementia	0.007	1
		hospice	0.003	1
		need	0.002	
		case	0.002	
	Prescription Drug Coverage and Costs	magic	0.009	
		prescription drug	0.009	1
		robber	0.003	
		case	0.003	
		need	0.002	
	Tobacco Abuse, Treatment, and Education	tobacco	0.024	1
		addiction	0.020	1
		nicotine	0.005	1
		case	0.003	
		need	0.002	
	Alcohol/Controlled and Illegal Drug Abuse, Treatment, and Education	alcohol	0.032	1
		adrian	0.004	
		need	0.003	
		case	0.003	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		kidnapper	0.002	
	Research and Development	research	0.179	1
		experiment	0.029	1
		dna	0.021	1
		gene	0.021	1
		clinical trial	0.008	1
Agriculture	General	farm	0.069	1
		garden	0.033	1
		beef	0.016	1
		agricultural	0.015	1
		grain	0.011	1
	Agricultural Trade	farmer	0.064	1
		crop	0.032	1
		agriculture	0.027	1
		harvest	0.016	1
		grower	0.007	1
	Government Subsidies to Farmers and Ranchers, Agricultural Disaster Insurance	bank	0.133	
		student	0.070	
		newly	0.030	
		ignore	0.014	
		rick	0.014	
	Food Inspection and Safety (Including Seafood)	food	0.170	1
		label	0.042	1
		poison	0.014	1
		bacterium	0.009	1
		nutrition	0.007	1
	Fisheries and Fishing	fish	0.058	1
		trap	0.031	1
		salmon	0.007	1
		fishery	0.005	1
		oyster	0.004	1
	Agricultural Research and Development	plant	0.109	1
		animal	0.064	1
		biologist	0.007	1
		acid	0.006	1
		romance	0.005	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
Labor and Employment	General	work	0.381	1
		worker	0.109	1
		employee	0.066	1
		career	0.046	1
		job	0.032	1
	Worker Safety and Protection, Occupational and Safety Health Administration (OSHA)	safety	0.103	1
		injury	0.052	1
		health	0.017	1
		miner	0.013	1
		occupation	0.006	1
	Employment Training and Workforce Development	talent	0.033	1
		workforce	0.013	1
		recruitment	0.007	1
		need	0.003	
		case	0.003	
	Employee Benefits	benefit	0.137	1
		retirement	0.032	1
		employment	0.029	1
		employee	0.026	1
		pension	0.020	1
	Employee Relations and Labor Unions	union	0.104	1
		labor	0.058	1
		labor union	0.005	1
		union member	0.004	1
		walkout	0.003	1
	Fair Labor Standards	income	0.108	1
		wage	0.055	1
		salary	0.029	1
		minimum wage	0.013	1
		overtime	0.008	1
	Parental Leave and Child Care	child care	0.014	1
		dodd	0.006	1
		care child	0.004	1
		need	0.003	
		case	0.002	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Migrant and Seasonal Workers, Farm Labor Issues	authority	0.151	
		band	0.037	
		assessment	0.029	
		transgender	0.019	
		tall	0.017	
Education	General	education	0.090	1
		school	0.069	1
		school district	0.021	1
		school system	0.008	1
		israeli palestinian	0.006	
	Higher Education	college	0.116	1
		course	0.100	1
		professor	0.077	1
		degree	0.049	1
		admission	0.015	1
	Elementary and Secondary Education	school	0.202	1
		student	0.128	1
		class	0.065	1
		teacher	0.059	1
		education	0.018	1
	School Shootings	endless	0.012	
		sandy hook	0.006	1
		school shoot	0.005	1
		need	0.003	
		back	0.002	
	Special Education	behavior	0.035	1
		mental illness	0.011	
		autism	0.008	1
		accommodation	0.007	1
		deaf	0.007	1
	Educational Excellence	math	0.025	1
		exam	0.016	1
		literacy	0.005	1
		test score	0.005	1
		need	0.003	
Environment	General	environmental	0.081	1
		environment	0.055	1
		earth	0.052	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		protection agency	0.011	1
		environmental protection agency	0.010	1
	Drinking Water Safety	water	0.169	1
		tap	0.028	1
		gallon	0.012	1
		drink water	0.008	1
		epa	0.005	1
	Waste Disposal	dump	0.018	1
		trash	0.017	1
		recycle	0.012	1
		garbage	0.011	1
		disposal	0.006	1
	Hazardous Waste and Toxic Chemical Regulation, Treatment, and Disposal	waste	0.041	1
		chemical	0.040	1
		epa	0.033	1
		pollution	0.027	1
		radiation	0.005	1
	Air Pollution, Global Warming, and Noise Pollution	climate	0.107	1
		climate change	0.106	1
		global warm	0.028	1
		fossil fuel	0.028	1
		carbon dioxide	0.013	1
	Indoor Environmental Hazards	toxic	0.022	1
		lung	0.021	1
		indoor	0.019	1
		hazard	0.007	1
		decrease	0.004	
	Species and Forest Protection	bear	0.135	1
		bird	0.028	1
		nature	0.021	1
		wolf	0.013	1
		whale	0.009	1
	Coastal Water Pollution and Conservation	ocean	0.046	1
		coast	0.036	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		shore	0.021	1
		wildlife	0.018	1
		lake	0.017	1
	Land and Water Conservation	land	0.115	1
		conservation	0.019	1
		public land	0.004	1
		conserve	0.003	1
		irrigation	0.003	1
Energy	General	energy	0.161	1
		energy agency	0.004	1
		back	0.002	
		case	0.002	
		need	0.002	
	Nuclear Energy and Nuclear Regulator Commission Issues	power plant	0.026	1
		reactor	0.010	1
		nuclear power	0.010	1
		radioactive	0.005	1
		nuclear plant	0.004	1
	Electricity and Hydroelectricity	power	0.261	1
		utility	0.035	1
		grid	0.015	1
		electrical	0.009	1
		megawatt	0.004	1
	Natural Gas and Oil (Including Offshore Oil and Gas)	oil	0.105	1
		gas	0.048	1
		pipeline	0.047	1
		drill	0.032	1
		natural gas	0.025	1
	Coal	coal	0.062	1
		carbon	0.050	1
		need	0.003	
		change	0.002	
		back	0.002	
	Alternative and Renewable Energy	solar	0.028	1
		clean energy	0.016	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		renewable energy	0.016	1
		emission	0.004	
		project	0.004	
	Energy Conservation	nfl	0.040	
		backyard	0.012	
		official unite	0.005	
		case	0.003	
		mileage	0.002	1
Immigration	Immigration	border	0.098	
		immigration	0.067	
		migrant	0.045	
		visa	0.021	
		asylum	0.015	
Transportation	General	car	0.164	1
		travel	0.051	1
		truck	0.046	1
		infrastructure	0.041	1
		highway	0.024	1
	Mass Transportation and Safety	station	0.070	1
		bridge	0.037	1
		subway	0.030	1
		transport	0.023	1
		fare	0.020	1
	Highway Construction, Maintenance, and Safety	street	0.136	1
		road	0.086	1
		traffic	0.054	1
		transportation	0.030	1
		lane	0.017	1
	Airports, Airlines, Air Traffic Control, and Safety	flight	0.071	1
		airport	0.049	1
		airline	0.047	1
		plane	0.033	1
		jet	0.031	1
	Railroad Transportation and Safety	train	0.179	1
		rail	0.022	1
		railroad	0.013	1
		derail	0.009	1
		commuter	0.008	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Auto Industry	ford	0.039	1
		gm	0.024	1
		tesla	0.024	1
		toyota	0.011	1
		chrysler	0.010	1
	Truck and Automobile Transportation and Safety	celebrate	0.065	
		crash	0.058	1
		accident	0.040	1
		kennedys	0.005	
		need	0.002	
	Martime Issues	ship	0.077	1
		boat	0.034	1
		port	0.033	1
		vessel	0.016	1
		cruise	0.015	1
Law, Crime, and Family Issues	General	claim	0.112	1
		investigation	0.102	1
		evidence	0.085	1
		crime	0.056	1
		fine	0.038	1
	Executive Branch Agencies Dealing with Law and Crime	justice department	0.065	1
		fbi	0.038	1
		resignation	0.023	1
		resign	0.015	1
		wrongdoing	0.011	1
	White Collar Crime and Organized Crime	fraud	0.048	1
		corruption	0.046	1
		hacker	0.018	1
		sec	0.011	1
		security exchange commission	0.007	1
	Illegal Drug Production, Trafficking, and Control	legal	0.145	1
		marijuana	0.032	1
		smuggle	0.011	1
		cocaine	0.010	1
		cartel	0.008	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Prisons	prison	0.097	1
		jail	0.056	1
		inmate	0.033	1
		tout	0.010	
		parole	0.009	1
	Juvenile Crime and the Juvenile Justice System	juvenile	0.011	1
		justice system	0.011	1
		adolescent	0.009	1
		boo	0.008	
		case	0.002	
	Child Abuse and Child Pornography	abuse	0.096	1
		sexual abuse	0.017	1
		need	0.002	
		case	0.002	
		back	0.002	
	Family Issues	parent	0.168	1
		adopt	0.046	1
		custody	0.024	1
		divorce	0.024	1
		social worker	0.005	1
	Police, Fire, and Weapons Control	gun	0.119	1
		firearm	0.021	1
		rifle	0.016	1
		gun control	0.015	1
		ammunition	0.011	1
	Police Issues	police	0.267	1
		police department	0.037	1
		law enforcement	0.031	1
		policeman	0.005	1
		need	0.002	
	Police Brutality	officer	0.011	
		charge	0.007	
		george floyd	0.006	1
		case	0.006	
		police brutality	0.006	1
	Criminal and Civil Code	execution	0.022	1
		death penalty	0.017	1
		rehabilitation	0.007	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		prison sentence	0.007	1
		death sentence	0.005	1
	Riots and Crime Prevention	riot	0.014	1
		bidens	0.006	
		capitol	0.005	
		tribune	0.005	
		democratic	0.004	
	Crimes	kill	0.210	1
		murder	0.059	1
		violence	0.054	1
		bomb	0.015	1
		hostage	0.013	1
	Criminal Justice Reform	arrest	0.130	1
		suspect	0.089	1
		defendant	0.029	1
		bail	0.016	1
		defense lawyer	0.009	1
	Sex Crimes	rape	0.050	1
		sexual assault	0.025	1
		sexual harassment	0.018	1
		student	0.014	
		sexual misconduct	0.010	1
Social Welfare	General	care	0.067	1
		volunteer	0.044	1
		nonprofit	0.040	1
		donation	0.032	1
		donate	0.025	1
	Food Stamps, Food Assistance, and Nutrition Monitoring Programs	result	0.208	
		care	0.076	1
		assist	0.021	1
		aid	0.009	1
		need	0.002	
	Poverty and Assistance for Low-Income Families	help	0.091	1
		poor	0.082	1
		aid	0.073	1
		poverty	0.035	1
		hunger	0.010	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Elderly Issues and Elderly Assistance Programs (Including Social Security Administration)	help	0.353	1
		social security	0.014	1
		payroll tax	0.003	1
		need	0.002	
		change	0.002	
	Social Services and Volunteer Associations	social service	0.006	1
		maya	0.006	
		case	0.003	
		back	0.002	
		need	0.002	
Community Development and Housing Issues	General	home	0.249	1
		house	0.184	1
		resident	0.085	1
		real estate	0.026	1
		mortgage	0.024	1
	Housing and Community Development	build	0.254	1
		design	0.083	1
		suburb	0.023	1
		suburban	0.014	1
		tenant	0.012	1
	Urban Economic Development and General Urban Issues	apartment	0.076	1
		urban	0.030	1
		rent	0.019	1
		rental	0.016	1
		landlord	0.013	1
	Housing Assistance for Homeless and Homeless Issues	shelter	0.052	1
		homeless	0.027	1
		homelessness	0.008	1
		need	0.002	
		case	0.002	
Banking, Finance, and Domestic Commerce	General	company	0.346	1
		financial	0.081	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		dollar	0.042	1
		asset	0.031	1
		retail	0.020	1
	U.S. Banking System and Financial Institution Regulation	bank	0.071	1
		banker	0.022	1
		citigroup	0.010	1
		bank america	0.010	1
		need	0.002	
	Securities and Commodities Regulation	sell	0.135	1
		buy	0.125	1
		board	0.094	1
		chairman	0.051	1
		investor	0.044	1
	Consumer Finance, Mortgages, and Credit Cards	credit	0.075	1
		loan	0.072	1
		debt	0.071	1
		payment	0.070	1
		borrow	0.022	1
	Insurance Regulation	insurance	0.042	1
		aig	0.005	1
		bowman	0.004	
		case	0.003	
		need	0.003	
	Bankruptcy	bankruptcy	0.029	1
		lender	0.026	1
		creditor	0.009	1
		file bankruptcy	0.006	1
		case	0.003	
	Corporate Mergers, Antitrust Regulation, and Corporate Management Issues	reply	0.018	
		pandemic	0.010	
		antitrust	0.006	1
		takeover	0.006	1
		market	0.006	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Small Business Issues and the Small Business Administration	business	0.314	1
		small business	0.017	1
		need	0.002	
		back	0.002	
		case	0.002	
	Copyrights and Patents	copyright	0.008	1
		intellectual property	0.007	1
		need	0.003	
		piracy	0.003	1
		case	0.003	
	Domestic Disaster Relief	disaster	0.052	1
		fema	0.009	1
		state emergency	0.008	1
		emergency management	0.003	1
		need	0.002	
	Tourism	tourist	0.035	1
		absolute	0.012	
		samsung	0.011	
		congressman	0.010	
		tourism	0.008	1
	Consumer Safety and Consumer Fraud	paris	0.060	
		middle	0.039	
		uncover	0.014	
		worthy	0.010	
		proxy	0.009	
	Sports and Gambling Regulation	casino	0.021	1
		dope	0.012	1
		lottery	0.009	1
		steroid	0.007	1
		drug test	0.005	1
	Cryptocurrencies	currency	0.034	1
		bitcoin	0.013	1
		cryptocurrency	0.007	1
		need	0.003	
		case	0.002	
Defense	Defense	military	0.143	1
		security	0.131	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		national security	0.029	1
		navy	0.024	1
		air force	0.019	1
	U.S. and Other Defense Alliances, U.S. Security Assistance	alliance	0.047	1
		nato	0.043	1
		close ally	0.005	1
		peacekeeping	0.003	1
		back	0.002	
	Military Intelligence, CIA, Espionage	intelligence	0.063	1
		cia	0.038	1
		spy	0.030	1
		intelligence agency	0.020	1
		intelligence official	0.014	1
	Military Readiness	military	0.052	1
		deploy	0.019	1
		homeland security	0.016	1
		counterterrorism	0.015	1
		department homeland security	0.010	1
	Arms Control and Nuclear Nonproliferation	nuclear	0.079	1
		nuclear weapon	0.019	1
		security council	0.018	1
		uranium	0.011	1
		nuclear deal	0.007	1
		nuclear deal	0.007	1
	Military Aid and Weapons Sales to Other Countries	weapon	0.096	1
		missile	0.045	1
		trait	0.007	
		proportion	0.003	
		need	0.002	
	Manpower, Military Personnel and Dependents, Military Courts	soldier	0.082	1
		sergeant	0.019	1
		troop	0.017	1
		military official	0.013	1
		sgt	0.010	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	VA Issues	veteran	0.077	1
		veteran affair	0.005	1
		ptsd	0.004	1
		need	0.002	
		case	0.002	
	Military Procurement and Weapons Systems Acquisitions and Evaluation	pentagon	0.042	1
		embark	0.009	
		missile defense	0.005	1
		case	0.003	
		need	0.002	
	Military Installations, Construction and Land Transfers	air base	0.003	1
		need	0.003	
		case	0.003	
		back	0.002	
		swagger	0.002	
	National Guard and Reserve Affairs	deploy	0.018	1
		national guard	0.014	1
		need	0.003	
		back	0.003	
		case	0.002	
	Civil Defense	domestic	0.058	1
		bomb	0.056	1
		militia	0.019	1
		anthrax	0.004	1
		case	0.002	
	Terrorism	terrorist	0.059	1
		terrorist	0.059	1
		taliban	0.056	1
		islamic state	0.046	1
		terrorism	0.037	1
		fbi	0.033	1
	National Guard	military base	0.008	1
		force base	0.004	1
		viability	0.004	
		need	0.003	
		case	0.003	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Military Contractors	aircraft	0.036	1
		beneficiary	0.011	
		deposition	0.007	
		pentagon	0.007	1
		baylor	0.005	
	War	war	0.140	1
		battle	0.070	1
		army	0.065	1
		conflict	0.042	1
		violence	0.035	1
	Relief of Claims Against U.S. Military	war crime	0.009	1
		renewal	0.007	
		drone strike	0.005	1
		mundane	0.004	
		need	0.002	
Space, Science, Technology and Communications	General	dr	0.119	1
		technology	0.096	1
		scientist	0.082	1
		science	0.068	1
		tech	0.038	1
	NASA, U.S. Government Use of Space, Space Exploration Agreements	space	0.115	1
		planet	0.033	1
		nasa	0.016	1
		astronaut	0.009	1
		shuttle	0.009	1
	Commercial Use of Space, Satellites	satellite	0.031	1
		orbit	0.011	1
		need	0.003	
		change	0.002	
		back	0.002	
	Telephone and Telecommunication Regulation	phone	0.122	1
		device	0.058	1
		communication	0.052	1
		telephone	0.031	1
		wireless	0.010	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Broadcast Industry Regulation	television	0.070	1
		tv	0.068	1
		audience	0.056	1
		cnn	0.036	1
		broadcast	0.030	1
	Computer Industry and Computer Security	apple	0.079	1
		computer	0.065	1
		code	0.041	1
		microsoft	0.025	1
		smartphone	0.013	1
	Research and Development	study	0.260	1
		scholar	0.021	1
		artifact	0.006	1
		archaeologist	0.004	1
		fossil	0.004	1
	Internet	twitter	0.085	1
		facebook	0.079	1
		internet	0.068	1
		social medium	0.065	1
		google	0.051	1
Foreign Trade	General	trade	0.194	1
		import	0.028	1
		trade commission	0.005	1
		ftc	0.004	1
		need	0.002	
	Trade Negotiations, Disputes, and Agreements	trade deal	0.010	1
		trade war	0.009	1
		nafta	0.005	1
		free trade	0.004	1
		free trade	0.004	1
		need	0.003	
	International Private Business Investments, Overseas Private Investment Corporation (OPIC)	acquire	0.034	1
		merger	0.015	1
		acquisition	0.014	1
		foreign investment	0.004	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		case	0.002	
	Productivity and Competitiveness of U.S. Business, U.S. Balance of Payments	export	0.045	1
		deficit	0.032	1
		import	0.006	1
		trespass	0.004	
		aretha	0.002	
	Tariff and Import Restrictions, Import Regulation	protection	0.083	1
		tariff	0.036	1
		wrench	0.006	
		free trade	0.004	1
		case	0.002	
	Exchange Rates and Related Issues	instagram	0.015	
		euro	0.005	1
		black	0.005	
		really	0.003	
		thats	0.003	
International Affairs and Foreign Aid	General	foreign	0.082	1
		king	0.064	1
		prime minister	0.064	1
		queen	0.040	1
		parliament	0.032	1
	Region: Middle East	iraqi	0.051	1
		afghan	0.047	1
		turkey	0.045	1
		pakistan	0.038	1
		syrian	0.036	1
	Region: Asia	china	0.194	1
		chinese	0.094	1
		japan	0.033	1
		indian	0.030	1
		north korea	0.028	1
	Region: Eastern Europe	poland	0.017	1
		kosovo	0.008	1
		hungary	0.008	1
		implicate	0.006	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		serbia	0.005	1
	Region: Latin America	mexican	0.041	1
		venezuela	0.017	1
		haiti	0.016	1
		brazil	0.007	1
		mexico	0.007	1
	Region: North Africa	african	0.055	1
		prior	0.031	
		egypt	0.030	1
		libya	0.023	1
		libyan	0.011	1
	Region: North America	canada	0.051	1
		canadian	0.038	1
		toronto	0.012	1
		greenland	0.006	1
		vancouver	0.005	1
	Region: Oceania	australia	0.036	1
		australian	0.024	1
		zealand	0.012	1
		need	0.002	
		back	0.002	
	Region: Post Soviet Countries	russian	0.132	1
		russia	0.121	1
		georgia	0.046	1
		moscow	0.033	1
		kremlin	0.015	1
	Region: Subsaharan Africa	africa	0.049	1
		south africa	0.017	1
		sudan	0.010	1
		ethiopia	0.007	1
		guinea	0.007	1
	Region: Western Europe	europe	0.081	1
		france	0.051	1
		france	0.051	1
		german	0.049	1
		britain	0.040	1
		italy	0.025	1
	Foreign Aid	relief	0.057	1
		civil war	0.026	1
		cuban	0.022	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		conflict	0.020	1
		aid	0.017	1
	Developing Countries	without border	0.004	1
		need	0.003	
		case	0.002	
		back	0.002	
		change	0.002	
	International Finance and Economic Development	executive vice president	0.008	
		imf	0.007	1
		international monetary fund	0.007	1
		need	0.003	
		case	0.003	
	Panama Canal	canal	0.010	1
		panama	0.007	1
		waterway	0.006	1
		need	0.003	
		case	0.002	
	United Nations, UNESCO, International Red Cross	un	0.042	1
		diplomat	0.033	1
		de	0.025	
		que	0.014	
		en	0.013	
Government Operations	General	us	0.237	1
		american	0.138	1
		government	0.137	1
		unite state	0.096	1
		republican	0.086	1
	Government Efficiency and Bureaucratic Oversight	oversight	0.031	1
		watchdog	0.012	1
		filibuster	0.010	1
		pork	0.010	1
		government shutdown	0.008	1
	Postal Service Issues	mail	0.057	1
		delivery	0.036	1
		postal service	0.006	1
		senate intelligence committee	0.004	
		postal	0.003	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
	Government Employee Benefits, Civil Service Issues	veto	0.020	
		pension fund	0.008	1
		environmentalist	0.007	
		environmental impact	0.005	
		need	0.003	
	Nominations and Appointments	nomination	0.047	1
		nominee	0.038	1
		nominate	0.030	1
		confirmation	0.024	1
		appointment	0.023	1
	Government Procurement, Procurement Fraud, and Contractor Management	contract	0.090	1
		bid	0.050	1
		contractor	0.031	1
		government spend	0.007	1
		procurement	0.003	1
	IRS Administration	audit	0.020	1
		irs	0.018	1
		income tax	0.016	1
		tax code	0.005	1
		internal revenue service	0.005	1
	Presidential Impeachment and Scandal	scandal	0.039	1
		resign	0.032	1
		impeachment	0.032	1
		impeach	0.009	1
		conflict interest	0.009	1
	Regulation of Political Campaigns	campaign finance	0.013	1
		starbucks	0.012	
		pac	0.007	1
		campaign contribution	0.004	1
		acronym	0.004	
	Lobbyists and Interest Groups	lobby	0.055	1
		lobbyist	0.027	1
		grassroots	0.005	1
		case	0.002	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		need	0.002	
	Political Campaigns	election	0.130	1
		vote	0.122	1
		voter	0.088	1
		candidate	0.084	1
		primary	0.039	1
	Bureaucratic Actions and Affairs	federal	0.143	1
		agency	0.135	1
		federal government	0.041	1
		bureau	0.029	1
		federal agency	0.009	1
	Executive Branch	trump	0.201	1
		president	0.188	1
		white house	0.065	1
		obama	0.060	1
		clinton	0.038	1
	Legislative Branch	law	0.153	1
		bill	0.113	1
		measure	0.085	1
		senate	0.063	1
		ban	0.056	1
	Judicial Branch	court	0.169	1
		justice	0.093	1
		supreme court	0.065	1
		judge	0.036	1
		overturn	0.016	1
	Census	census	0.024	1
		census bureau	0.008	1
		demographics	0.006	1
		redistricting	0.003	1
		case	0.002	
	District of Columbia Affairs	washington	0.241	1
		district columbia	0.007	1
		house president	0.006	
		care provider	0.005	
		need	0.002	
	White House Press Office and Other Government Public Relations	public	0.201	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		protest	0.095	1
		speech	0.071	1
		survey	0.050	1
		poll	0.038	1
	Ideology in Politics	conservative	0.151	1
		liberal	0.061	1
		governor	0.018	
		ideology	0.014	1
		need	0.002	
	Misinformation in Politics	conspiracy	0.019	1
		conspiracy theory	0.009	1
		misinformation	0.009	1
		app	0.005	
		user	0.005	
Public Lands and Water Management	General	island	0.101	1
		land	0.048	1
		interior	0.016	1
		interior department	0.005	1
		need	0.002	
	National Parks, Memorials, Historic Sites, and Recreation	park	0.131	1
		museum	0.057	1
		memorial	0.018	1
		monument	0.017	1
		smithsonian	0.004	1
	Native American Affairs	native	0.037	1
		tribe	0.028	1
		console	0.005	
		sioux	0.004	1
		tribal area	0.004	1
	Natural Resources, Public Lands, and Forest Management	mountain	0.052	1
		forest	0.027	1
		desert	0.023	1
		national park	0.011	1
		marsh	0.003	1
	Water Resources Development and Research	river	0.077	1
		water	0.038	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		lake	0.020	1
		dam	0.013	1
		reservoir	0.007	1
Arts and Entertainment	General	culture	0.093	1
		christmas	0.041	1
		holiday	0.039	1
		creative	0.023	1
		festival	0.022	1
	Fine Art	art	0.110	1
		artist	0.073	1
		exhibit	0.014	1
		museum	0.006	1
		graffito	0.005	1
	Books	book	0.186	1
		publish	0.120	1
		writer	0.013	1
		podcast	0.011	1
		fiction	0.011	1
	Cooking	cook	0.054	1
		dinner	0.043	1
		kitchen	0.028	1
		lunch	0.028	1
		recipe	0.017	1
	Mobile and Video Games	video game	0.013	1
		xbox	0.003	1
		need	0.003	
		case	0.003	
		back	0.002	
	Fashion	style	0.071	1
		mostly	0.067	
		fashion	0.051	1
		beauty	0.024	1
		salon	0.009	1
	Human Interest	family	0.229	1
		love	0.120	1
		wife	0.068	1
		brother	0.053	1
		sister	0.032	1
	Journalism	story	0.077	1
		paper	0.075	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		article	0.058	1
		writer	0.043	1
		newspaper	0.022	1
	Film, TV Series, and Streaming Services	director	0.130	1
		film	0.094	1
		actor	0.040	1
		hollywood	0.027	1
		comedy	0.020	1
	Music	music	0.109	1
		rock	0.058	1
		radio	0.041	1
		pop	0.041	1
		concert	0.024	1
	Night Life	restaurant	0.092	1
		bar	0.083	1
		wine	0.032	1
		beer	0.022	1
		dance	0.013	1
	Pets	dog	0.069	1
		master	0.036	
		cat	0.025	1
		pet	0.018	1
		need	0.002	
	Theatre	play	0.232	1
		perform	0.052	1
		theater	0.033	1
		hall	0.032	1
		studio	0.021	1
	Zoos	mood	0.023	
		dance	0.014	
		zoo	0.012	1
		uncommon	0.008	
		need	0.002	
	Travel	travel	0.086	1
		hotel	0.082	1
		tourism	0.009	1
		need	0.002	
		case	0.002	
	Other	contain	0.072	
		goldman sachs	0.012	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		cartoon	0.009	1
		need	0.002	
		case	0.002	
State and Local Government Administration	Local Government	city	0.267	1
		county	0.092	1
		town	0.076	1
		mayor	0.070	1
		city council	0.010	1
	State Government	state	0.396	1
		governor	0.067	1
		gov	0.035	1
		legislature	0.015	1
		state government	0.008	1
	Region: West	california	0.132	1
		montana	0.013	1
		colorado	0.011	1
		washington state	0.011	1
		carson	0.009	1
	Region: Midwest	chicago	0.068	1
		ohio	0.052	1
		iowa	0.044	1
		michigan	0.037	1
		detroit	0.027	1
	Region: South	north carolina	0.036	1
		miami	0.032	1
		south carolina	0.031	1
		alabama	0.031	1
		louisiana	0.025	1
	Region: Northeast	york city	0.064	1
		manhattan	0.055	1
		brooklyn	0.049	1
		massachusetts	0.032	1
		ny	0.024	1
	Region: Other	alaska	0.031	1
		hawaii	0.017	1
		ak	0.003	1
		anchorage	0.003	1
		honolulu	0.003	1
Weather and Natural Disasters	Weather and Natural Disasters	hurricane	0.043	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		weather	0.041	
		forecast	0.024	
		cloud	0.023	
		drought	0.019	
Fires	Fires	burn	0.073	
		smoke	0.048	
		flame	0.014	
		blaze	0.010	
		fire	0.009	
Sports and Recreation	General	run	0.191	1
		team	0.128	1
		game	0.098	1
		player	0.065	1
		sport	0.053	1
	Baseball	ball	0.052	1
		baseball	0.035	1
		yankee	0.025	1
		inning	0.010	1
		pitcher	0.009	1
	Basketball	basketball	0.039	1
		nba	0.035	1
		knicks	0.008	1
		need	0.002	
		case	0.002	
	Boxing	box	0.069	1
		method	0.037	
		rain	0.036	
		transcript	0.014	
		repay	0.008	
	Football	football	0.043	1
		super bowl	0.020	1
		nfl	0.010	1
		touchdown	0.007	1
		dictatorship	0.005	
	Golf	golf	0.025	1
		golf course	0.007	1
		really hard	0.006	
		golfer	0.005	1
		pga	0.003	1
	Hockey	majority	0.108	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		hockey	0.010	1
		nhl	0.004	1
		avatar	0.003	
		need	0.002	
	Olympics	olympic	0.055	1
		gold medal	0.008	1
		athlete	0.005	
		gymnast	0.004	1
		olympic game	0.004	1
	Motor Sports	underestimate	0.010	
		norwegian	0.008	
		european central bank	0.005	
		nascar	0.003	1
		case	0.003	
	Soccer	soccer	0.033	1
		flynn	0.013	
		official president	0.005	
		soccer team	0.004	1
		townhouse	0.003	
	Tennis	different	0.210	
		tennis	0.018	1
		poke	0.005	
		university student	0.004	
		mistrust	0.003	
	Bicycling	bike	0.024	1
		bicycle	0.010	1
		feast	0.006	
		cyclist	0.005	1
		need	0.002	
Death Notices	Death Notices	death	0.185	1
		dead	0.069	1
		survive	0.046	1
		funeral	0.020	1
		bury	0.018	1
Churches and Religion	General	god	0.105	1
		religious	0.047	1
		religion	0.021	1
		worship	0.018	1
		congregation	0.015	1
	Catholicism	pope	0.032	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		cardinal	0.020	1
		vatican	0.016	1
		catholic church	0.010	1
		archbishop	0.009	1
	Cults	cult	0.009	1
		replicate	0.008	
		jordanian	0.005	
		vox	0.003	
		need	0.002	
	Eastern Religions	hindu	0.010	1
		buddhist	0.008	1
		monk	0.008	1
		adult child	0.004	
		case	0.002	
	Islam	muslim	0.097	1
		islamic	0.038	1
		islam	0.025	1
		mosque	0.022	1
		quran	0.003	1
	Judaism	jewish	0.047	1
		jew	0.031	1
		orthodox	0.015	1
		rabbi	0.010	1
		synagogue	0.008	1
	Protestantism	church	0.155	1
		christian	0.112	1
		pastor	0.041	1
		jesus	0.032	1
		bible	0.028	1
	Megachurches	former federal	0.005	
		billy graham	0.004	1
		need	0.003	
		megachurch	0.003	1
		back	0.003	
	Atheism and Agnosticism	secular	0.017	1
		atheist	0.005	1
		compress	0.003	
		need	0.003	
		case	0.002	
Other	People	man	0.208	1

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		woman	0.190	1
		person	0.101	1
		girl	0.050	1
		boy	0.044	1
	Rules	rule	0.245	1
		regulation	0.065	1
		case	0.002	
		need	0.002	
		back	0.002	
	News	story	0.166	1
		editor	0.052	1
		newspaper	0.035	1
		article	0.015	1
		paper	0.004	1
	Change	percent	0.213	1
		add	0.130	1
		record	0.103	1
		decline	0.063	1
		drop	0.048	1
	Unseeded 1	commit	0.061	
		simple	0.052	
		equal	0.031	
		discovery	0.021	
		equity	0.017	
	Unseeded 2	base	0.200	
		probably	0.064	
		emerge	0.053	
		due	0.041	
		permit	0.029	
	Unseeded 3	engage	0.067	
		hide	0.040	
		hack	0.027	
		dance	0.018	
		sea level	0.015	
	Unseeded 4	clear	0.152	
		free	0.116	
		property	0.052	
		critical	0.050	
		healthy	0.027	
	Unseeded 5	less	0.186	

Table B.1: Modified Comparative Agenda Project Topic Structure with Lexical Priors *(continued)*

Major Topic	Minor Topic	Term	Beta	Prior
		unite	0.065	
		select	0.024	
		cohen	0.023	
		adjust	0.018	
	Unseeded 6	clintons	0.044	
		mask	0.042	
		via	0.042	
		possibly	0.032	
		gross	0.014	
	Unseeded 7	crisis	0.093	
		representative	0.081	
		presidential	0.039	
		pace	0.029	
		extremist	0.020	
	Unseeded 8	congress	0.143	
		declare	0.055	
		flag	0.036	
		atmosphere	0.021	
		sean	0.013	
	Unseeded 9	arrive	0.094	
		troop	0.051	
		easy	0.033	
		regular	0.030	
		staff member	0.022	
	Unseeded 10	official	0.304	
		screen	0.048	
		pride	0.017	
		slate	0.009	
		practical	0.008	
	Unseeded 11	interest	0.104	
		mark	0.103	
		equipment	0.037	
		tip	0.029	
		ap	0.017	
	Unseeded 12	inside	0.094	
		size	0.056	
		pray	0.028	
		afford	0.028	
		privilege	0.021	

APPENDIX C

SAMPLE OF CLIMATE CHANGE CO-OCCURRENCE STORIES

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes)

Major Topic	Tweet ID	Story headline and mentions
Macroeconomics	1552634329913753600	U.S. G.D.P. Report: G.D.P. Report Shows U.S. Economy Shrank Again
		"As a profession, we’ve been really focused on future economic impacts from climate change, because we’ve been focused on how you should be taxing carbon emissions,” said Derek Lemoine, an associate professor of economics at the University of Arizona.
		Other researchers are working on developing measures of economic growth that integrate not just production of goods and services — which themselves can accelerate climate change — but environmental and social elements as well.
Civil rights	1543273314524844033	Spurred by the Supreme Court, a Nation Divides Along a Red-Blue Axis
		On abortion, climate change, guns and much more, two Americas — one liberal, one conservative — are moving in opposite directions.
		The most immediate breaking point is on abortion, as about half the country will soon limit or ban the procedure while the other half expands or reinforces access to reproductive rights. But the ideological fault lines extend far beyond that one topic, to climate change, gun control and L.G.B.T.Q. and voting rights.
Health	1483864521823379458	New Research Shows How Health Risks to Children Mount as Temperatures Rise
		With climate change, heat waves and rising temperatures are becoming more frequent. And that has repercussions for human health.
		"We know that, due to climate change, days with extreme heat are going to be more frequent and more intense,” said Francesca Dominici, a biostatistician at Harvard’s T.H. Chan School of Public Health who has studied the effects of extreme heat on human health and was not involved in the new research.
Agriculture	1560397402368081920	Scientists Boost Crop Performance by Engineering a Better Leaf

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes) *(continued)*

Major Topic	Tweet ID	Story headline and mentions
		Now, researchers say that by using genetic modifications to increase the efficiency of photosynthesis, they significantly increased yields in a food crop, soybeans, providing a glimmer of potential that such methods could someday put more food on tables as climate change and other threats make it harder for vulnerable populations across the globe to feed their families. Human-caused climate change is threatening to exacerbate the problem, with increased droughts and storms causing more disruptions to food supplies.
Education	1513164920652771345	A College Fights 'Leftist Academics' by Expanding Into Charter Schools "The phrase 'climate change' doesn't appear at all, and 'global warming' occurs only once, at the sixth-grade level, as 'global warming theory,'" Glenn Branch, the organization's deputy director, wrote in an email. A spokeswoman for Hillsdale said the current science curriculum included texts that discuss climate change.
Environment	1535487020839190531	Some Monarch Butterfly Populations Are Rising. Is It Enough to Save Them? These declines have been attributed to a variety of factors, including climate change and logging near the overwintering sites. They found that in some regions, especially in parts of the Midwest, glyphosate use was associated with declines in abundance.;;But they also documented a countervailing force: climate change. In the northern part of the United States, increasing temperatures were correlated with increases in monarch abundance.
Energy	1565882296997224450	Hawaii Closes Its Last Coal-Fired Power Plant "It really is about reducing greenhouse gases," said Gov. David Ige, a Democrat, in an interview with The Associated Press. "And this coal facility is one of the largest emitters. Taking it offline means that we'll stop the 1.5 million metric tons of greenhouse gases that were emitted annually."
Transportation	1503079726235242496	California's Ambitious High-Speed Rail at a Crossroads The passage of Mr. Biden's $1 trillion infrastructure package, astronomical gas prices and California's insistence that the state lead the nation in addressing climate change make the moment seem perfect for providing oxygen to the plan.

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes) *(continued)*

Major Topic	Tweet ID	Story headline and mentions
		If completed, they say, the system would be an economic supercharger connecting two of the nation's biggest population centers and a desperately needed alternative to choked freeways and jammed airports as climate change becomes an ever urgent challenge.
Law, Crime, and Family Issues	1530056227417464834	The Rise and Fall of America's Environmentalist Underground
		While acknowledging that such attacks might fail, Malm nevertheless argues that the urgency of global warming — in the 16 years since Dibee's indictment, the world has collectively pumped about 500 billion more tons of carbon into the atmosphere — demands new tactics.
		As climate change, no longer an abstraction, has begun to transform American life in the form of heat, fire, floods and smoke, it is a story that may sound different to some listeners now than when prosecutors first told it.
Community Development and Housing Issues	1487830478405222402	The Ticking Clock for Miami's Condo Empire
		"... Because we're learning so much about sea-level rise and climate change and we're realizing that a lot of our old measures are outdated."
		He told me that he accepted the reality of climate change — he'd seen with his own eyes that the sea levels around his private dock were climbing. And he was as wary as anyone about the pace of development in Miami Beach, where, he stressed, the towers rise so high that some residents rarely catch a glimpse of the sun.
Banking, Finance, and Domestic Commerce	1522327833095331842	Stocks slide, erasing Wednesday's big gains, as volatility continues to reign.
		It also would let banks get credit for loans and investments that helped poor communities deal with climate change, like helping improve protections from floods and other natural disasters that are becoming more frequent as a result of global warming.
		In a statement, Shell's chief executive, Ben van Beurden, appeared to suggest that the disruption from the war in Ukraine had demonstrated that there was still a need for strong oil and gas businesses despite pressures to tackle climate change.
Defense	1565011541900120070	Many Developed Countries View Online Misinformation as 'Major Threat'
		Researchers asked 24,525 people from 19 countries with advanced economies to rate the severity of threats from climate change, infectious diseases, online misinformation, cyberattacks from other countries and the condition of the global economy.

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes) *(continued)*

Major Topic	Tweet ID	Story headline and mentions
		Climate change was the highest-rated concern for most countries, with a median of 75 percent of respondents saying it is a major threat.
Space, Science, Technology	1604257683795841024	For Planet Earth, This Might Be the Start of a New Age
		The official timeline of Earth's history — from the oldest rocks to the dinosaurs to the rise of primates, from the Paleozoic to the Jurassic and all points before and since — could soon include the age of nuclear weapons, human-caused climate change and the proliferation of plastics, garbage and concrete across the planet.
		Radionuclides are a convenient global marker, but they say nothing about climate change or other human effects, said Erle C. Ellis, an ecologist at the University of Maryland, Baltimore County.
Foreign trade	1602673451021619203	Europe Reaches Deal for Carbon Tax Law on Imports
		The European Union has taken a step closer to adopting a groundbreaking carbon tax law that would impose a tariff on imports from countries that fail to take strict steps to curb their greenhouse gas emissions.
		The United States is looking at similar legislation, and last week the Biden administration sent a proposal to the European Union to impose tariffs on steel and aluminum produced in ways that harm the environment as part of efforts to tackle climate change.
International Affairs and Foreign Aid	1539969029456007170	Extreme Weather Hits China With Massive Floods and Scorching Heat
		The two-pronged weather emergency that China is experiencing reflects a global trend of increasingly frequent and lengthy episodes of extreme weather driven by climate change.
		But in its pursuit of economic development, it has also become the world's largest polluter, with greenhouse gas emissions exceeding those of all developed nations combined.
Government Operations	1586382380524347392	Biden's Agenda Hangs in the Balance if Republicans Take Congress
		For their part, Republicans aim to roll back Mr. Biden's corporate tax increases, climate change spending, student loan forgiveness and I.R.S. expansion targeting wealthy tax cheats.
		They include a \$1.9 trillion pandemic stimulus package, a \$1 trillion plan to upgrade the nation's roads, bridges and other infrastructure, a \$739 billion package to fight climate change and curb prescription drug prices and a \$250 billion program to boost the semiconductor industry.

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes) *(continued)*

Major Topic	Tweet ID	Story headline and mentions
Public Lands and Water Management	1588100957933543425	Here's Where the U.S. Is Testing a New Response to Rising Seas Native American tribes are competing for the first federal grants designed to help move communities away from high water and other dangers posed by climate change. As climate change gets worse, tribes like Shoalwater Bay are being squeezed between existential threats and brutal financial arithmetic.
Arts and Entertainment	1485492593354022913	Can Works Like 'Don't Look Up' Get Us Out of Our Heads? In the doomsday smash and Bo Burnham's pandemic musical "Inside," themes of climate change, digital distraction and inequality merge and hit home. Though heavy with metaphors — most important, the comet signifying climate change — its message is clear and not open to interpretation: Wake up!
Weather and Natural Disasters	1597863043479879682	Hurricane Season Ends, Marked by Quiet August and Deadly September Dry spells. Flash droughts, the kind that arrive quickly and can lay waste to crops in a matter of weeks, are becoming more common and faster to develop around the world, and human-caused climate change is a major reason, a scientific study has found. The reason: An alarming decline of fish stocks linked to heavily engineered waterways and the supercharged heat and drought that come with climate change.
Fires	1530623714852802561	U.S. Forest Service Planned Burn Caused Largest New Mexico Wildfire In response to the fire investigators' findings, New Mexico's governor, Michelle Lujan Grisham, said the federal government must examine its fire management practices and how they account for climate change. She said climate change has made it more difficult to use prescribed fires because fire seasons have increased to seven to eight months from around three months.
Sports and Recreation	1548709616057552902	Golf's Birthplace Faces a Risky Future on a Warming Planet But golf has had little choice but to start weighing its own role in climate change — most notably through the vast, lush and thirsty courses that sometimes take the place of trees and then require fertilizer and mowing — while puzzling over how to preserve fairways and greens around the world. Citing climate change, the International Olympic Committee's president has said that Games organizers "may have to have a look into the overall calendar and whether there must be a shift."

Table C.1: Sample of Climate Change Associations in Non-Climate Change Coverage (@nytimes) *(continued)*

Major Topic	Tweet ID	Story headline and mentions
Churches and Religion	1478497751574974467	An Evangelical Climate Scientist Wonders What Went Wrong "As Christians, we’re naturally suspicious of people who believe differently from us” — doesn’t that seem like a recipe for a climate-change skeptic even before you add in political ideology?" "Where might cross-ideological conversations, particularly about climate change, happen for people who aren’t in a similar situation?"