2016

Mathematical Modeling of Public Opinion using Traditional and Social Media

Emily Cody
University of Vermont

Mathematical Modeling of Public Opinion using Traditional and Social Media

A Dissertation Presented

by

Emily Cody

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

October, 2016

Defense Date: June 2, 2016
Dissertation Examination Committee:

Chris Danforth, Ph.D., Advisor
Peter Dodds, Ph.D.
Josh Bongard, Ph.D.
Jennie Stephens, Ph.D., Chairperson
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

With the growth of the internet, data from text sources has become increasingly available to researchers in the form of online newspapers, journals, and blogs. This data presents a unique opportunity to analyze human opinions and behaviors without soliciting the public explicitly. In this research, I utilize newspaper articles and the social media service Twitter to infer self-reported public opinions and awareness of climate change. Climate change is one of the most important and heavily debated issues of our time, and analyzing large-scale text surrounding this issue reveals insights surrounding self-reported public opinion. First, I inquire about public discourse on both climate change and energy system vulnerability following two large hurricanes. I apply topic modeling techniques to a corpus of articles about each hurricane in order to determine how these topics were reported on in the post event news media. Next, I perform sentiment analysis on a large collection of data from Twitter using a previously developed tool called the "hedonometer". I use this sentiment scoring technique to investigate how the Twitter community reports feeling about climate change. Finally, I generalize the sentiment analysis technique to many other topics of global importance, and compare to more traditional public opinion polling methods. I determine that since traditional public opinion polls have limited reach and high associated costs, text data from Twitter may be the future of public opinion polling.

Citations

Material from this dissertation has been published in the following form:

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M.. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. PloS one, 10(8), e0136092.

AND

Cody, E. M., Stephens, J. C., Bagrow, J. P., Dodds, P. S., & Danforth, C. M.. (2016). Transitions in climate and energy discourse between Hurricanes Katrina and Sandy. Journal of Environmental Studies and Sciences, 10.1007/s13412-016-0391-8.

AND

Cody, E.M., Reagan, A. J., Dodds, P. S., & Danforth, C. M.. (2016). Public Opinion Polling with Twitter. *In Preparation*.

Dedication

To my friends, my family, and my fiancé

Acknowledgements

I would like to take this opportunity to thank those who supported me throughout the past four years emotionally, physically, and financially. I could not have accomplished what I have without my friends and colleagues at my side the entire way. Thank you to my officemates for always keeping the place social and friendly. Thank you to the IGERT administrator, Curtis Saunders, for ensuring our printers had ink and our refunds were processed quickly, and that the conference room was always reserved from 12-1 for group lunch. Thank you to Tom McAndrew for assistance with using the VACC and insightful conversations about research directions. Thank you to Mark Wagy for answering silly programming questions and for surviving four years at the desk next to me. Thank you to Andy Reagan, the data guru, for assisting me with any and all data collection questions. Thank you to Nick Allgaier and Cathy Bliss, who showed me that UVM was the place for me when I visited four years ago and continued to serve as mentors throughout my time here. I would also like to acknowledge the rest of the IGERT students and the Computational Story Lab crew, who I will always consider close friends.

A massive thank you goes out to my advisors, Chris Danforth and Peter Dodds, for all of their guidance, advice, and life lessons. You've both always believed in me more than I believed in myself. Thank you to my co-authors, Jim Bagrow who introduced me to data science and topic modeling, and Jennie Stephens who introduced me to the energy transition. Thank you to my committee, Chris Danforth, Peter Dodds, Jennie Stephens, and Josh Bongard for all of your guidance. And thank you to Jeff Marshall, the IGERT PI, for organizing the entire IGERT program.

I would also like to thank my family. Thanks to my parents, Lisa and Paul, for

supporting me in every life decision I have ever made. An extremely special thank you goes out to my fiancé, Matt, who moved to Vermont while I pursued my education, and puts up with more than any man should. And thank you to my cats, Yoda and Luke, who can make me smile on even the worst of days.

Finally, I would like to acknowledge my sources of funding. Thank you to the NSF for both the Integrated Graduate Education and Research Traineeship (IGERT) and Mathematics and Climate Research Network (MCRN) grants that supported my work for the past four years.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The adoption of mobile technology and the resulting emergence of computational social science has begun to revolutionize our understanding of human behavior. Analysis of text-based data can provide researchers with answers to real world problems involving self-reported human behaviors, opinions, and sentiments using automatic text processing techniques that require little to no knowledge of the content of the selected text. There is a massive amount of text data available to researchers ranging from social media posts, newspaper articles, and scientific papers. This research is focused on developing and applying mathematical techniques to large amounts of text data in order to solve real world, interdisciplinary problems.

This thesis uses sentiment analysis and machine learning techniques to infer characteristics about human behaviors and opinions surrounding issues of national and global interest. In previous works, topic models have been utilized to extract hidden concepts from large groups of documents without the use of manual coders.

Specifically, in [11] Deerwester *et al.* introduce Latent Semantic Analysis (LSA), a technique that organizes words and documents according to their hidden meanings

within a corpus. LSA uses singular value decomposition (SVD) to reduce a term-document matrix to a latent semantic space for term and document comparisons. LSA is often used in computational linguistics and information retrieval as a search engine technique [23], and to identify and separate different types of text [19, 5, 25].

In [6] Blei introduces Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique that is based on Bayesian statistics. LDA has since been used to study topics and trends within historical newspaper corpora over time [29, 38], find scientific topics in PNAS abstracts [16], and analyze historical trends in Computational Linguistics topics [17]. It has been proven to work as well or better than several comparable algorithms [9].

A third technique used to analyze human behaviors through text data is sentiment analysis. Specifically, I utilize the "hedonometer", an instrument designed to calculate a "happiness score" for a large collection of text [13]. In [20], Kloumann *et al.* collected happiness scores for 10,222 of the most frequently used English words in four disparate corpora using Amazon's Mechanical Turk. Each word was given a score from 1 (least happy) to 9 (most happy) depending on how users reported that that word made them feel. Since its development by Dodds *et al.* in [13], the hedonometer has been used to analyze the happiness of cities and states [27], the happiness of the English language as a whole [20], and the relationship between individuals' happiness and that of those they connect with [7].

Chapters 2 and 3 of this research involve implementing the machine learning and sentiment analysis techniques above to analyze text data related to climate change and energy consumption. Climate change is one of the largest global issues of our time. According to the IPCC Fifth Assessment report, humans are "very likely" to

be responsible for the increased warming of our planet [15]. While there is a scientific consensus on this issue [14, 2], the existence and cause of climate change continue to be heavily debated among politicians and the general public. This ongoing debate presents an opportunity to examine how various groups discuss this challenging issue. Since the general public learns most of what it knows about science from the mass-media [36], this research investigates climate change conversation within newspaper articles (Chapter 2) and the social media site Twitter (Chapter 3).

Research has shown a probable link between the increasing ocean temperature and the severity and frequency of hurricanes and tropical storms [15, 18, 24]. Damage and deaths caused by extreme weather events can serve as a tangible reminder of the consequences of climate change. Therefore, hurricanes have the potential to raise awareness and increase public concern about this global issue. Disruptions to our energy infrastructure also highlight the ramifications of severe natural disasters. These extreme events can serve as a teachable experience for those not previously engaged in climate change issues [28].

Extensive news coverage of extreme weather events has been found to increase public awareness of climate change by highlighting the risks it presents to our lives [4, 37]. Therefore, in Chapter 2 of this work, I analyze newspaper coverage of Hurricane Katrina (2005) and Hurricane Sandy (2012). In this analysis, I investigate the major topics in post-event reporting using topic modeling techniques LSA and LDA, to determine if climate change, energy consumption, and energy system vulnerability are among them. Furthermore, I seek to determine if either hurricane highlighted a link between these three topics. Climate change and energy consumption are obviously and intricately linked, however it is not always obvious that climate change and

increasing natural disasters will affect our energy infrastructure. Links between climate change and energy are typically focused on climate mitigation (reducing energy emissions), however climate change and energy are also linked in terms of increased energy system vulnerability in a changing climate [33]. Despite these links, climate change and energy are still often discussed in the media separately [32, 35]. Here, I aim to determine if and how this has changed since Hurricanes Katrina and Sandy.

Climate change opinions, however, do not arise solely from the hands of those who can publish articles on the subject. In the last decade, there has been a shift from the consumption of traditional mass media to the consumption of social media, and it has been shown that topics involving global warming are even more prominent on social media [8]. The social media site Twitter provides its users 140 characters to display their thoughts and opinions on any matter they choose to discuss. The majority of topics trending on Twitter are headline or persistent news [22], making Twitter a valuable source for analyzing climate change discussion. Previous works on the subject include an analysis of geo-tagged tweets before, during, and after Hurricane Sandy [21], a sentiment analysis exploring subjective, objective, positive, and negative tweets mentioning climate change [1], and an analysis of climate change hashtags used to locate pro/denialist communities on Twitter [34].

In Chapter 3, I collect 1.5 million tweets containing the word "climate" from September 2008 through July 2014 and utilize the hedonometer to determine how public opinion of climate change on Twitter varies in response to climate change news and events. In this chapter, I determine that Twitter is a valuable resource for analyzing public opinion on climate change by analyzing happiness time series and utilizing word shift graphs. In Chapter 4, similar methods are extended to a dataset of

10,000 terms to determine self-reported public opinions surrounding political topics, ideas, feelings, and commercial businesses.

Available public opinion data is very valuable to those in the computational social science field, however it is difficult to obtain for research purposes. In addition, traditional public opinion surveys can only research a limited number of people, and participants' opinions can vary in response to social influence [10, 31]. It is also difficult to obtain high resolution public opinion surveys, due to the nature of the data collection process. Twitter, however, has 320 million active monthly users displaying their opinions on both important and trivial subjects, and therefore has massive potential for the spread of awareness on major global issues [26]. We can collect data from Twitter in real time and can thus produce results with a very high temporal resolution. However, it is not required that Twitter users share their demographic information and thus Twitter does not represent and unbiased sample population. In traditional public opinion surveys, participants' demographic information is known and thus population biases can be avoided. In a traditional survey, there are typically a set of questions relating to a specific topic of interest, and thus researchers obtain the answers to exactly the questions they were looking for. Public opinions on social media are self-reported and thus may not answer one specific question. On Twitter, users can discuss any topic of interest and thus there is quit a bit of noise that should be taken into consideration. Previous works have created opinion polling resources by developing a sentiment estimation tool using wikipedia [12], a tool that correlates input data with time series on Twitter [3], and comparing Twitter data to traditional public opinion surveys [30].

In Chapter 4 of this work, I calculate *ambient happiness*, i.e. the happiness sur-

rounding a given term, on Twitter for each of the 10,000 words given happiness scores through Amazon Mechanical Turk (hereafter referred to as the labMT dataset). I compare ambient happiness time series to traditional public opinion polls and perform further analysis on the causes for shifts in happiness between two time periods. I compare the two methods of public opinion polling to determine if one may supplant or complement the other.

The purpose of this research is to explore the value of large scale text-based data using machine learning and text analysis techniques. I aim to determine what these datasets can tell us about human opinions and behaviors. Mining text data for natural language processing has many benefits to researchers, policymakers, marketers, and many others. In the future, public opinion polling may be used to complement and compare responses to traditional public opinion surveys. With both resources at our disposal, researches will have more evidence to draw specific conclusions about public opinions.

Bibliography

[1] Xiaoran An, Auroop R Ganguly, Yi Fang, Steven B Scyphers, Ann M Hunter, and Jennifer G Dy. Tracking climate change opinions from Twitter data. *Workshop on Data Science for Social Good*, 2014.

[2] William RL Anderegg, James W Prall, Jacob Harold, and Stephen H Schneider. Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27):12107–12109, 2010.

[3] Dolan Antenucci, Michael R. Andwerson, Penghua Zhao, and Michael Cafaerlla. A query system for social media signals. 2015.

[4] Allan Bell. Media (mis) communication on the science of climate change. *Public understanding of science*, 3(3):259–275, 1994.

[5] Yves Bestgen. Improving text segmentation using latent semantic analysis: A reanalysis of choi, wiemer-hastings, and moore (2001). *Computational Linguistics*, 32(1):5–12, 2006.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[7] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.

[8] Maxwell T Boykoff. *Who speaks for the climate?: Making sense of media reporting on climate change*. Cambridge University Press, 2011.

[9] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[10] Robert B Cialdini and Nathalie Garde. *Influence*, volume 3. A. Michel, 1987.

[11] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[12] Peter U Diehl, Bruno U Pedroni, Andrew Cassidy, Paul Merolla, Emre Neftci, and Guido Zarrella. TrueHappiness: Neuromorphic emotion recognition on TrueNorth. *arXiv preprint arXiv:1601.04183*, 2016.

[13] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752, 2011.

[14] Peter T Doran and Maggie Kendall Zimmerman. Examining the scientific consensus on climate change. *Eos, Transactions American Geophysical Union*, 90(3):22–23, 2009.

[15] Christopher B Field. *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[16] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[17] David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics, 2008.

[18] Daniel G. Huber and Jay Gulledge. *Extreme weather and climate change: Understanding the link, managing the risk*. Pew Center on Global Climate Change Arlington, 2011.

[19] Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics, 2006.

[20] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PLoS ONE*, 7(1):e29484, 2012.

[21] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Nowcasting disaster damage. *arXiv preprint arXiv:1504.06827*, 2015.

[22] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[23] Thomas K Landauer and Michael L Littman. Computerized cross-language document retrieval using latent semantic indexing, April 5 1994. US Patent 5,301,109.

[24] Michael E. Mann and Kerry A. Emanuel. Atlantic hurricane trends linked to climate change. *Eos, Transactions American Geophysical Union*, 87(24):233–241, 2006.

[25] Philip M McCarthy, Stephen W Briner, Vasile Rus, and Danielle S McNamara. Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In *Natural language processing and text mining*, pages 107–122. Springer, 2007.

[26] Yelena Mejova, Ingmar Weber, and Michael W Macy. *Twitter: A digital socioscope*. Cambridge University Press, 2015.

[27] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[28] Teresa A. Myers, Edward W. Maibach, Connie Roser-Renouf, Karen Akerlof, and Anthony A. Leiserowitz. The relationship between personal experience and belief in the reality of global warming. *Nature Climate Change*, 3(4):343–347, 2013.

[29] Robert K Nelson. Mining the dispatch. *Mining the dispatch*, 2010.

[30] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[31] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[32] Jennie C Stephens, Gabriel M Rand, and Leah L Melnick. Wind energy in us media: A comparative state-level analysis of a critical climate change mitigation technology. *Environmental Communication*, 3(2):168–190, 2009.

[33] Jennie C Stephens, Elizabeth J Wilson, Tarla R Peterson, and James Meadowcroft. Getting smart? Climate change and the electric grid. *Challenges*, 4(2):201–216, 2013.

[34] Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[35] Elizabeth J Wilson, Jennie C Stephens, Tarla Rai Peterson, and Miriam Fischlein. Carbon capture and storage in context: The importance of state policy and discourse in deploying emerging energy technologies. *Energy Procedia*, 1(1):4519–4526, 2009.

[36] Kris M Wilson. Mass media as sources of global warming knowledge. *Mass Comm Review*, 22:75–89, 1995.

[37] Kris M Wilson. Drought, debate, and uncertainty: Measuring reporters' knowledge and ignorance about climate change. *Public Understanding of Science*, 9(1):1–13, 2000.

[38] Tze-I Yang, Andrew J Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics, 2011.

Chapter 2

Transitions in climate and energy discourse between Hurricanes Katrina and Sandy

2.1 Abstract

Although climate change and energy are intricately linked, their explicit connection is not always prominent in public discourse and the media. Disruptive extreme weather events, including hurricanes, focus public attention in new and different ways, offering a unique window of opportunity to analyze how a focusing event influences public discourse. Media coverage of extreme weather events simultaneously shapes and reflects public discourse on climate issues. Here we analyze climate and energy newspaper coverage of Hurricanes Katrina (2005) and Sandy (2012) using topic models, mathematical techniques used to discover abstract topics within a set of doc-

uments. Our results demonstrate that post-Katrina media coverage does not contain a climate change topic, and the energy topic is limited to discussion of energy prices, markets, and the economy with almost no explicit linkages made between energy and climate change. In contrast, post-Sandy media coverage does contain a prominent climate change topic, a distinct energy topic, as well as integrated representation of climate change and energy, indicating a shift in climate and energy reporting between Hurricane Katrina and Hurricane Sandy.

2.2 Introduction

Climate change is one of the most challenging issues of our time. Anticipated climate disruptions, including a 4°C increase in the Earth's average temperature by the end of the 21st century [25] and more frequent and intense extreme weather events, result from increased atmospheric concentrations of greenhouse gases attributed primarily to fossil fuel burning for energy.

Given probable links between the increasing ocean temperature and the severity and frequency of hurricanes and tropical storms [33, 16, 24], extreme weather events have potential to raise awareness and increase public concern about climate change. The disruptions caused by hurricanes and other storms can also raise awareness and focus attention on energy system vulnerability. These extreme events can serve as a teachable experience for those not previously engaged with these issues [38]. Indeed, previous research has shown that after experiencing a large hurricane, citizens are more likely to adopt a pro-environmental belief system and support politicians who are climate change activists [46]. Populations living as far as 800 km from the path of

a hurricane report having experienced it in some way [23]. Extensive news coverage of extreme weather events has also been found to increase public awareness of climate change by highlighting tangible and specific risks [6, 50]. It has also been shown that individuals affected by a natural disaster are more likely to strengthen interactions on social media [42]. As climate change news is prominent on social media [13], these interactions provide another mechanism for raising climate change awareness following a natural disaster.

This research recognizes the complex relationship between the news media and public discourse on science and policy. The news media both shapes public perceptions and public discourse and reflects and represents public perceptions and public discourse [18, 17]. The media shapes public opinion of science by avoiding complex scientific language and displaying information for the layperson [37, 41, 45]. People are more likely to learn about environmental and other science related risks through the media than through any other source [14, 41]. Research indicates that news media establish the context within which future information will be interpreted [41]. In this research we analyze media coverage to characterize differences in the public discourse about climate change and energy after Hurricane Katrina and Hurricane Sandy.

Links between climate change and energy are often focused on climate mitigation, e.g., reducing greenhouse gas emissions from energy systems by shifting low-carbon energy systems. However, climate change and energy are also linked in terms of increased energy system vulnerability in a changing climate [48]. Hurricanes and other extreme weather events often cause disruptions to energy systems including infrastructure damage, fuel supply shortages, and increases in energy prices. Flooding and high wind speeds reveal multiple energy system vulnerabilities including evacuations

of oil rigs and power outages at refineries, which can contribute to energy supply shortages and price increases.

Despite the multiple linkages between climate change and energy systems, the issues of climate and energy are still often discussed in the media separately [47, 49]. Greater integration of the public discourse on climate change and energy could facilitate more sophisticated consideration of the opportunities for changing energy systems to prepare for climate change [25, 35].

A 2005 study on climate change in the media revealed that articles often frame climate change as a debate, controversy, or uncertainty, which is inconsistent with how the phenomenon is framed within the scientific community [2]. A recent 2015 linguistic study determined that the IPCC summaries, intended for non-scientific audiences, are becoming increasingly more complex and more difficult for people to understand [5], which highlights the critical interpretive role of the media in public discourse.

Here, we quantitatively compare media coverage of climate change, energy, and the links between climate and energy after Hurricanes Katrina and Sandy, two of the most disruptive and costly hurricanes to ever hit the United States [27, 9]. Since energy system disruption represents a tangible consequence of climate change, the linking of these two topics in post-hurricane newspaper coverage provides readers with a portal for climate change education and awareness. Newspaper media was selected for analysis rather than social media because in the rapidly changing media landscape the circulation patterns of these well-established newspapers have been relatively stable during the study period. Also, a 2014 study by the American Press Institute determined that 61% of Americans follow the news through print newspapers and

magazines alone. 69% of Americans use laptops and computers which includes online newspapers. 88% of Americans find their news directly from a news organization, as opposed to roughly 45% from social media and 30% from electronic news ads [34]. With this high percentage of Americans getting news from the media, analysis of climate change reporting provides insights on shifts in public discourse and awareness.

We apply two topic modeling techniques stemming from different areas of mathematics to a corpus (collection of text) of newspaper articles about each hurricane. A topic model uses word frequencies within a corpus to assign one or more topics to each text. For our present analysis, we employ Latent Semantic Analysis (LSA), which uses singular value decomposition to reduce a term-document matrix to latent semantic space, and Latent Dirichlet Allocation (LDA), a probabilistic bayesian modeling technique, which defines each hidden topic as a probability distribution over all of the words in the corpus (we provide more details in the Methods section, Sec. 2.3).

We apply a topic modeling approach as a way to assess the integration of climate change, energy and the links between climate and energy within post-hurricane media coverage. Topic modeling is a valuable tool for the kind of research we perform as it does not require manual coders to read thousands of articles. Instead, a specified number of topics are determined through analysis of the frequency of each word in each article in the corpus. The resulting model explains the corpus in detail by categorizing the articles and terms into topics.

We focus on the two most disruptive and costly hurricanes in U.S. history. In August 2005, Hurricane Katrina struck Louisiana as a Category 3 storm, affecting the Gulf Coast from central Florida to Texas, causing over 100 billion dollars in damage and roughly 1,800 deaths. Katrina destroyed or severely damaged much of New

Orleans and other heavily populated areas of the northern Gulf Coast, resulting in catastrophic infrastructure damage and thousands of job losses [27]. Hurricane Sandy hit the northeastern United States in October 2012. It was the largest hurricane of the 2012 Atlantic hurricane season, caused 233 reported deaths, and over 68 billion dollars in damage to residential and commercial facilities as well as transportation and other infrastructure [9]. Many businesses faced short term economic losses, while the travel and tourism industry experienced longer term economic difficulties. In the time shortly after Sandy hit, repairs and reconstructions were estimated to take four years [21].

We use this quantitative approach to assess the degree to which climate change or energy related topics are included in newspaper coverage following Hurricanes Sandy and Katrina. The individual words that define each topic reveal how climate change and energy were represented in post-event reporting, which in turn shapes public discourse.

We first describe the dataset and methods of analysis in Sec. 2.3. We then describe the results of each topic modeling technique for each hurricane and make comparisons between the two corpora in Sec. 2.4. We explore the significance of these results in Sec. 2.5&2.6.

2.3 Methods

2.3.1 Data Collection

We collected newspaper articles published in major U.S. newspapers in the year following each of the hurricanes. We chose the timespan of one year to capture the duration of media coverage following each hurricane and also to ensure we had enough articles from each hurricane to conduct a proper mathematical analysis. We identified newspaper articles through a search that included the name of the hurricane and either the word "hurricane" or "storm" in either the title or leading paragraphs of the article. To account for regional variation in post-hurricane reporting, we chose four newspapers spanning major regions of the United States: Northeast, New England, Midwest, and West. We chose the following four newspapers due to their high Sunday circulation, and because they are high-profile, established newspapers with high readership: The New York Times, The Boston Globe, The Los Angeles Times, and The Chicago Tribune are influential and well-respected nationally as well as locally. These four newspapers are consistently in the top 25 U.S. Sunday newspapers and were available for article collection through online databases. We collected articles appearing onwards from the first of the month the hurricane occurred in throughout the subsequent year using the ProQuest, LexisNexis, and Westlaw Campus Research online databases. The total number of articles collected and included in the corpora for analysis are 3,100 for Hurricane Katrina and 1,039 for Hurricane Sandy. We transform each corpus into a term-document matrix for the analysis.

2.3.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method of uncovering hidden relationships in document data [15]. LSA uses the matrix factorization technique Singular Value Decomposition (SVD) to reduce the rank of the term-document matrix, and merge the dimensions that share similar meanings. SVD creates the following matrices:

$$M = USV^T,$$

where the matrix M is the original $t \times d$ matrix (number of terms by number of documents), the columns of the matrix U are the eigenvectors of MM^T, the entries in the diagonal of the matrix S are the square roots of the eigenvalues of MM^T, and the rows of the matrix V^T are the eigenvectors of M^TM. Retaining the k largest singular values and setting all others to 0 gives the best rank k approximation of M. This rank reduction creates a $t \times k$ term matrix, U_kS_k, consisting of term vectors in latent semantic space as its columns, and a $k \times d$ document matrix, $S_kV_k^T$, consisting of document vectors as its rows. The documents and terms are then compared in latent semantic space using cosine similarity as the distance metric [7]. If two term vectors have cosine distances close to 1, then these terms are interpreted to be related to each other in meaning. We explain this process further in Fig. 2.1.

We load the documents into a term-document matrix and remove common and irrelevant terms. The terms we removed included terms common to the articles like "hurricane", "storm", "sandy", and "katrina", along with names of authors and editors of the articles. We then convert each frequency in the matrix to term frequency-

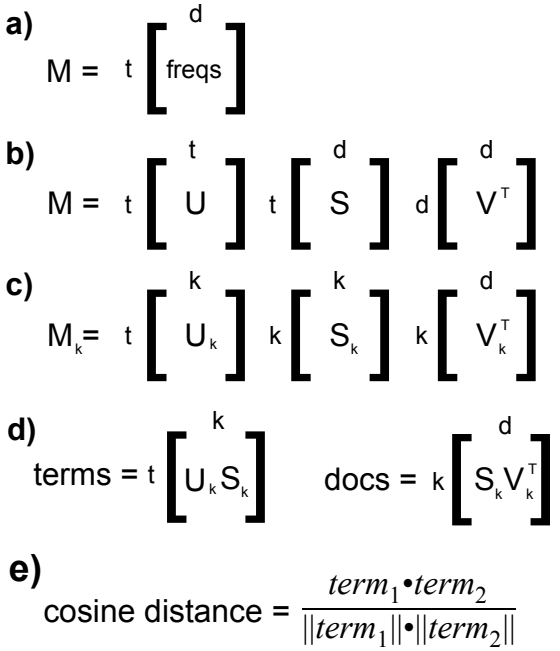

Figure 2.1: a) M is a $t \times d$ matrix where t and d are the number of terms and documents in the corpus. An entry in this matrix represents the number of times a specific term appears in a specific document. b) Singular Value Decomposition factors the matrix M into three matrices. The matrix S has singular values on its diagonal and zeros everywhere else. c) The best rank k approximation of M is calculated by retaining the k highest singular values. k represents the number of topics in the corpus. d) Each term and each document is represented as a vector in latent semantic space. These vectors make up the rows of the term matrix and the columns of the document matrix. e) Terms and documents are compared to each other using cosine similarity, which is determined by calculating the cosine of the angle between two vectors.

inverse document frequency (tf-idf) via the following transformation [4]:

$$w_{i,j} = \begin{cases} (1 + \log_2 f_{i,j}) \times \log_2 \frac{N}{n_i} & f_{i,j} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where the variable $w_{i,j}$ is the new weight in the matrix at location (i, j), $f_{i,j}$ is the current frequency in position (i, j), N is the number of documents in the corpus, and n_i is the number of documents containing word i. This weighting scheme places higher

weights on rarer terms because they are more selective and provide more information about the corpus, while placing lower weights on common words such as “the” and “and”.

We run LSA on the tf-idf term-document matrix for each hurricane. We then compare the documents and terms in the corpus to a given query of terms in latent semantic space. We transform the words that the query is composed of into term vectors, and calculate their centroid to give the vector representation of the query. If the query is only one word in length, then the vector representation of the query equals the vector representation of the word. We analyze three queries using LSA: “climate”, “energy”, and “climate, energy”. LSA gives the terms most related to this query vector, which we then use to determine how climate change and energy are discussed both separately and together in the media after Hurricanes Katrina and Sandy.

2.3.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA), a probabilistic topic model [11, 10], defines each hidden topic as a probability distribution over all of the words in the corpus, and each document’s content is then represented as a probability distribution over all of the topics. Fig. 2.2 gives illustrations of distributions for a potential LDA model.

LDA assumes that the documents were created via the following generative process. For each document:

1. Randomly choose a distribution of topics from a dirichlet distribution. This distribution of topics contains a nonzero probability of selecting each word in the corpus.

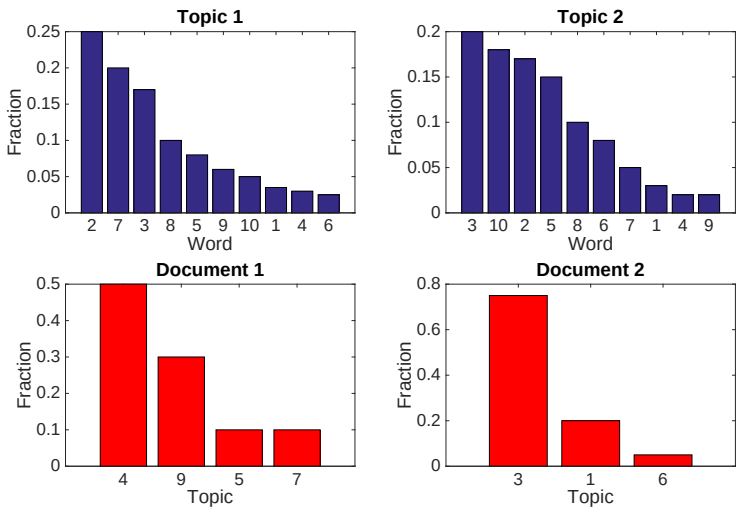

Figure 2.2: a) Examples of two topic distributions that may arise from an LDA model. In this example, each topic is made up of 10 words and each word contributes to the meaning of the topic in a different proportion. b) Examples of two document distributions that may arise from an LDA model. Document 1 is made up of four major topics, while document 2 is made up of 3 major topics.

2. For each word in the current document:

   a) Randomly select a topic from the topic distribution in part 1.

   b) Randomly choose a word from the topic just selected and insert it into the document.

3. Repeat until document is complete.

The distinguishing characteristic of LDA is that all of the documents in the corpus share the same set of k topics, however each document contains each topic in a different proportion. The goal of the model is to learn the topic distributions. The

generative process for LDA corresponds to the following joint distribution:

$$P(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$
$$\prod_{i=1}^{K} P(\beta_i) \prod_{d=1}^{D} P(\theta_d) \left(\prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\beta_{1:K}, z_{d,n}) \right),$$

where β_k is the distribution over the words, $\theta_{d,k}$ is the topic proportion for topic k in document d, $z_{d,n}$ is the topic assignment for the nth word in document d, and $w_{d,n}$ is the nth word in document d. This joint distribution defines certain dependences. The topic selection, $z_{d,n}$ is dependent on the topic proportions each the article, θ_d. The current word $w_{d,n}$ is dependent on both the topic selection, $z_{d,n}$ and topic distribution $\beta_{1:k}$. The main computational problem is computing the posterior. The posterior is the conditional distribution of the topic structure given the observed documents

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}.$$

The denominator of the posterior represents the probability of seeing the observed corpus under any topic model. It is computed using the sampling based algorithm, Gibbs Sampling.

We generate topic models for the Hurricane Sandy and Katrina articles using LDA-C, developed by Blei in [11]. We remove a list of common stop words from the corpus, along with common words specific to this corpus such as “Sandy”, “Katrina”, “hurricane”, and “storm”. After filtering through the words, we use a Porter word stemmer to stem the remaining words, so each word is represented in one form, while it may appear in the articles in many different tenses [44].

2.3.4 Determining the Number of Topics

The number of topics within a particular corpus depends on the size and scope of the corpus. In our corpora, the scope is already quite narrow as we only focus on newspaper articles about a particular hurricane. Thus, we do not expect the number of topics to be large, and to choose the number of topics for the analysis, we implement several techniques.

First, to determine k, the rank of the approximated term-document matrix used in LSA, we look at the singular values determined via SVD. The 100 largest singular values are plotted in Fig. 2.3 for Hurricanes Sandy and Katrina. The singular value decay rate slows considerably between singular values 20 and 30 for both matrices. We find that topics become repetitive above $k = 20$, and thus we choose $k = 20$ as the rank of the approximated term-document matrix in LSA.

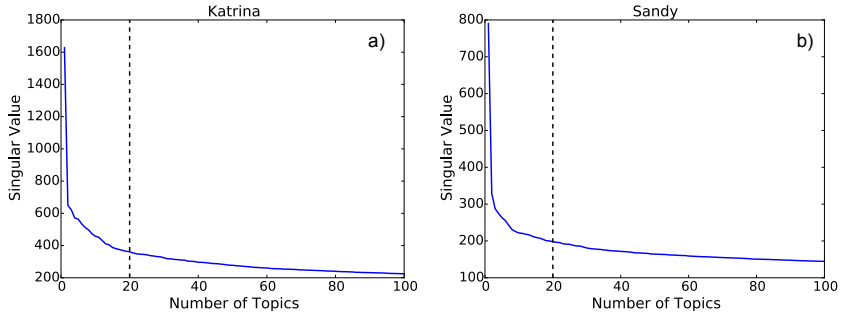

Figure 2.3: The 100 largest singular values in the (a) Hurricane Sandy and (b) Hurricane Katrina tf-idf matrices. The elbow around 20 topics (see dashed line) determines the value of k for SVD in LSA.

To determine the number of topics for LDA to learn we use the perplexity, a measure employed in [11] to determine how accurately the topic model predicts a sample of unseen documents. We compute the perplexity of a held out test set of

documents for each hurricane, and vary the number of learned topics on the training data. Perplexity will decrease with the number of topics and should eventually level out when increasing the number of topics no longer increases the accuracy of the model. The perplexity may begin to increase when adding topics causes the model to overfit the data. Perplexity is defined in [11] as

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\},$$

where the numerator represents the log-likelihood of unseen documents $\mathbf{w}_d$, and the denominator represents the total number of words in the testing set. We separate the data into 10 equal testing and training sets for 10 fold cross validation on each hurricane. We run LDA on each of the 10 different training sets consisting of 90% of the articles in each hurricane corpus. We then calculate the perplexity for a range of topic numbers on the testing sets, each consisting of 10% of the articles. We average the perplexity at each topic number over the testing sets, and plot the result in Fig. 2.4(a) & (b).

Figure 2.4 indicates that the optimal number of topics in the Hurricane Sandy corpus is roughly 20 distinct topics, while the optimal number in the Hurricane Katrina corpus is between 280 and 300 distinct topics. Compared to the Sandy corpus, the Hurricane Katrina corpus contains three times as many articles and about double the number of unique words (17,898 vs 9,521). On average, an article in the Hurricane Sandy corpus contains 270 words, while an article in the Hurricane Katrina corpus contains 376 words. The difference in these statistics may account for the difference in optimal topic numbers in Fig. 2.4. To test this hypothesis, we take 100 random samples of size 1039 (the size of the Sandy corpus) from the Katrina corpus

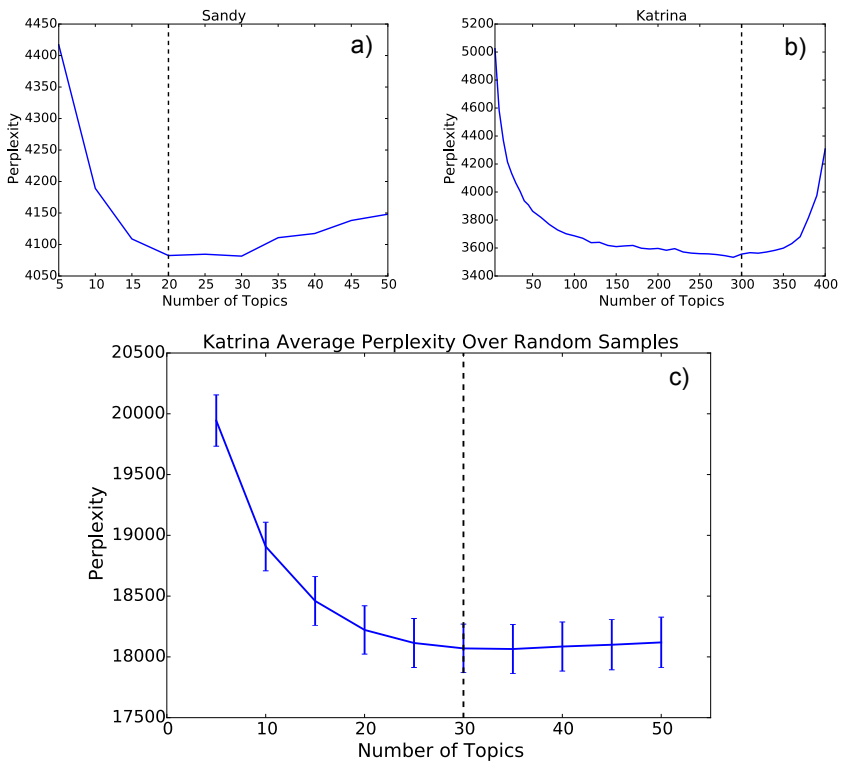

Figure 2.4: Average perplexity (over 10 testing sets) vs number of topics for the full (a) Sandy and (b) Katrina corpora. Perplexity measures how well the model can predict a sample of unseen documents. A lower perplexity indicates a better model. Dashed lines show the optimal number of topics. (c) The average perplexity over 100 random samples of 1039 (the size of the Sandy corpus) documents from the Katrina corpus. Each topic number is averaged first over 10 testing sets and then over 100 random samples from the full Katrina corpus. Topic numbers increase by 2. Error bars indicate the 95% confidence intervals.

and calculate the average perplexity over these samples. For each of the 100 random samples, we use 10 testing and training sets for 10 fold cross validation, as was done in the previous calculations of perplexity. We calculate the average perplexity over the 10 testing sets for each topic number, and then average over the 100 samples for each topic number, showing the result in Fig. 2.4(c). We find that on average, the optimal number of topics for a smaller Katrina corpus is around 30.

Based on the above analysis, we opt to use a 20-topic model for Hurricane Sandy and a 30-topic model for Hurricane Katrina in our LDA analysis of the post-event media coverage.

2.4 Results

2.4.1 Latent Semantic Analysis

We compute a topic model for each corpus using LSA as described in the preceding methods section. We provide 40 words most related to the three queries of interest in Tables 2.1 & 2.2. We list the 100 most related words to each query in the Supplementary Materials (see Tables A.1 & A.2). While it is not possible to objectively explain why each word ranks where it does in the following lists, we search for a common theme within the words to determine how climate and energy were discussed in the media following these hurricanes.

Hurricane Katrina

Within the Hurricane Katrina news media coverage, explicit reference to climate change was infrequent. The set of words most related to "climate" includes words such as "theory", "unlikely", "belief", and "possibility", indicating that linkages with climate change after Hurricane Katrina were tentative. The uncertain link between hurricanes and climate change is often present in political discussions, thus the appearance of the word "politician" in the "climate" list is not surprising. A direct quote from the article most related to the "climate" query reads:

Hurricane Katrina					
"climate"	Similarity	"energy"	Similarity	"climate,energy"	Similarity
climate	1.000	energy	1.000	energy	0.979
larger	0.866	prices	0.986	prices	0.952
destroy	0.861	exchange	0.968	deutsche	0.945
formally	0.848	consumers	0.966	price	0.943
theory	0.844	weinberg	0.966	underinvestment	0.943
sound	0.837	argus	0.964	signaling	0.941
gale	0.826	reidy	0.962	discounting	0.940
reinforced	0.817	splurge	0.960	java	0.940
journal	0.815	hummer	0.960	argus	0.939
sensitive	0.814	markets	0.959	hummer	0.938
unlikely	0.812	downers	0.958	oil	0.937
belief	0.809	highs	0.958	consumers	0.937
phenomenon	0.809	underinvestment	0.957	shocks	0.934
rail	0.800	exporting	0.954	weinberg	0.934
studying	0.796	price	0.954	markets	0.934
wealthy	0.795	reserves	0.954	profits	0.931
brings	0.792	signaling	0.953	reserves	0.931
barge	0.792	dampening	0.950	exchange	0.931
ancient	0.791	oil	0.950	peaks	0.931
masters	0.786	java	0.949	highs	0.929
politicians	0.785	cents	0.948	splurge	0.927
professor	0.783	deutsche	0.948	exporting	0.927
recommendations	0.782	gasoline	0.947	gasoline	0.923
thick	0.782	traders	0.946	dampening	0.923
marked	0.780	nariman	0.946	pinch	0.922
alter	0.779	discounting	0.945	oils	0.922
sounds	0.776	behravesh	0.944	soaring	0.922
hole	0.776	retailers	0.943	exported	0.920
peril	0.775	barrel	0.942	reidy	0.919
extremely	0.771	heating	0.942	output	0.919
avoided	0.770	oils	0.942	exporter	0.917
loose	0.770	shocks	0.941	easing	0.917
multi	0.769	idled	0.941	putins	0.917
appear	0.767	jolted	0.941	record	0.916
devastating	0.766	output	0.940	tumbling	0.916
draft	0.764	peaks	0.937	demand	0.915
possibility	0.764	profits	0.936	downers	0.915
roiled	0.759	soared	0.936	automaker	0.913
retracted	0.758	exported	0.936	heating	0.913
mismanagement	0.758	premcor	0.935	disruptions	0.913

Table 2.1: Results of LSA for Hurricane Katrina for 3 different queries. Words are ordered based on their cosine similarity with the query vector.

> "When two hurricanes as powerful as Katrina and Rita pummel the Gulf Coast so close together, many Americans are understandably wondering if something in the air has changed. Scientists are wondering the same thing. The field's leading researchers say it is too early to reach unequivocal conclusions. But some of them see evidence that global warming may be increasing the share of hurricanes that reach the monster magnitude of Katrina, and Rita" [12].

Words such as "studying", "professor", and "masters" also indicate that reporting on climate change focused on research and academics. The "climate" list does not contain words relating to energy or energy systems and does not focus on the science or consequences of climate change.

Within the 40 words most related to the "energy" query, the majority pertain to energy prices and the stock market. Within the "climate" and "energy" lists there is no overlap in the 40 most related words to these queries.

The "climate" and "energy" vectors are averaged to create the "climate, energy" query vector. The list of words most similar to this query is far more comparable to the "energy" list than the "climate" list. Of the 100 most related words to each query, there are 84 shared words between the "energy" and "climate, energy" lists. This list again focuses on energy prices and not at all on climate change or infrastructure vulnerability, indicating that discussions about climate change, energy, and power outages were independent of one another within media reporting following Hurricane Katrina.

Hurricane Sandy					
"climate"	Similarity	"energy"	Similarity	"climate,energy"	Similarity
climate	1.000	energy	1.000	climate	0.979
change	0.963	technologies	0.949	warmer	0.961
reduce	0.957	fuels	0.946	georgetown	0.956
warming	0.957	fossil	0.943	warming	0.955
reducing	0.956	hydroelectric	0.936	reduce	0.955
pressures	0.952	renewable	0.932	energy	0.952
georgetown	0.947	rogue	0.932	reducing	0.951
lowering	0.943	employing	0.921	pressures	0.948
talks	0.942	warmer	0.920	fossil	0.947
devise	0.938	supplying	0.918	fuels	0.946
expands	0.938	firing	0.913	change	0.946
outweigh	0.937	efficiency	0.911	technologies	0.945
warmer	0.937	streamlined	0.911	coal	0.943
plants	0.934	generating	0.908	global	0.942
drought	0.933	altering	0.906	hydroelectric	0.941
manipulation	0.929	coal	0.906	emissions	0.940
emissions	0.929	consumption	0.900	firing	0.937
global	0.929	adapt	0.898	outweigh	0.936
imperative	0.927	sparked	0.895	generating	0.933
arizona	0.924	dimming	0.894	carbon	0.930
attribute	0.923	georgetown	0.892	arizona	0.930
scientists	0.923	carbon	0.889	editorials	0.929
planet	0.920	masonry	0.888	plants	0.927
pollution	0.919	global	0.886	humanitys	0.926
curbing	0.918	erratic	0.885	altering	0.926
coal	0.917	searchable	0.884	manipulation	0.924
editorials	0.915	faster	0.882	pollution	0.923
targets	0.914	emissions	0.881	employing	0.923
oceans	0.912	skeptics	0.880	drought	0.922
vigil	0.912	proportion	0.877	extracted	0.921
scenarios	0.911	trillions	0.876	foretaste	0.920
extracted	0.911	foretaste	0.876	skeptics	0.919
humanitys	0.911	warming	0.875	lowering	0.919
distraction	0.910	reduce	0.875	dioxide	0.918
pentagon	0.910	editorials	0.875	efficiency	0.918
contiguous	0.909	humanitys	0.875	planet	0.917
controlling	0.908	eco	0.875	curbing	0.917
carbon	0.907	ton	0.874	consumption	0.915
dioxide	0.906	efficient	0.872	expands	0.914
extremes	0.905	cities	0.872	subtler	0.913

Table 2.2: Results of LSA for Hurricane Sandy for 3 different queries. Words are ordered based on their cosine similarity with the query vector.

Hurricane Sandy

In the Hurricane Sandy corpus, we find the word "climate" is most related to words describing climate change and global warming. We also see words related to energy such as "emissions", "coal", "carbon", and "dioxide". Including the top 100 words most related to "climate" we see more energy related words including "fossil", "hydroelectric", "technologies", and "energy" itself. This list differs substantially from that of the Hurricane Katrina analysis.

The word "energy" in the Hurricane Sandy corpus is most related to words describing climate change, such as the contributions of fossil fuels and the potential of renewable ("hydroelectric", "renewable") energy resources. This list of words focuses largely on how energy consumption is contributing to climate change, and, unlike the Katrina corpus, considerably overlaps with the list of "climate" words.

Of the 100 words most related to "energy", 58 of them are also listed in the 100 words most related to "climate". Of the 20 documents most related to the word "energy", 15 of them are also listed in the 20 documents most related to "climate". Many of these articles discuss harmful emissions, renewable energy, and fossil fuels.

In the Hurricane Sandy corpus, the "climate, energy" query is again most related to the climate change and global warming related terms. There are 87 shared terms in the "climate" and "climate, energy" lists and 66 shared terms in the "energy" and "climate, energy" related lists. This result illustrates that when climate change was discussed in the media following Hurricane Sandy, energy related themes were often present.

2.4.2 Latent Dirichlet Allocation

We generate LDA models for both the Sandy and Katrina corpora using 20 topics and 30 topics for Sandy and Katrina respectively (see Methods). The 20 most probable words in 10 selected topic distributions are given in Tables 2.3 & 2.4. The full models are given in the Supplementary Materials (see Tables A.3 & A.4). In addition to creating a distribution of topics over words, LDA also creates a distribution of documents over topics. Each topic is present in each document with some nonzero probability. We counted the number of times each topic appeared as one of the top two ranked topics in an article and divided this number by the number of articles in the corpus. Fig. 2.5 summarizes the overall results of LDA for Katrina (a) and Sandy (b) by giving the proportion of articles that each topic appears in with high probability. We determined the topic names by manually analyzing the probability distribution of words in each topic. We go into more detail on the topics of importance in the following sections.

Hurricane Katrina

In Table 2.3 we give 10 of the 30 topics in the LDA model for Hurricane Katrina. In the Hurricane Katrina model, we see topics relating to deaths, relief, insurance, flooding, and energy. We also see location specific topics such as sporting events, Mardi Gras, and music. A major topic that is absent from this model is climate change. Similar to the results we saw for the Katrina LSA model, the energy topic (Topic 8) in the Katrina LDA model contains words relating to energy prices, the market, and the economy. In addition to a missing climate change topic, there is

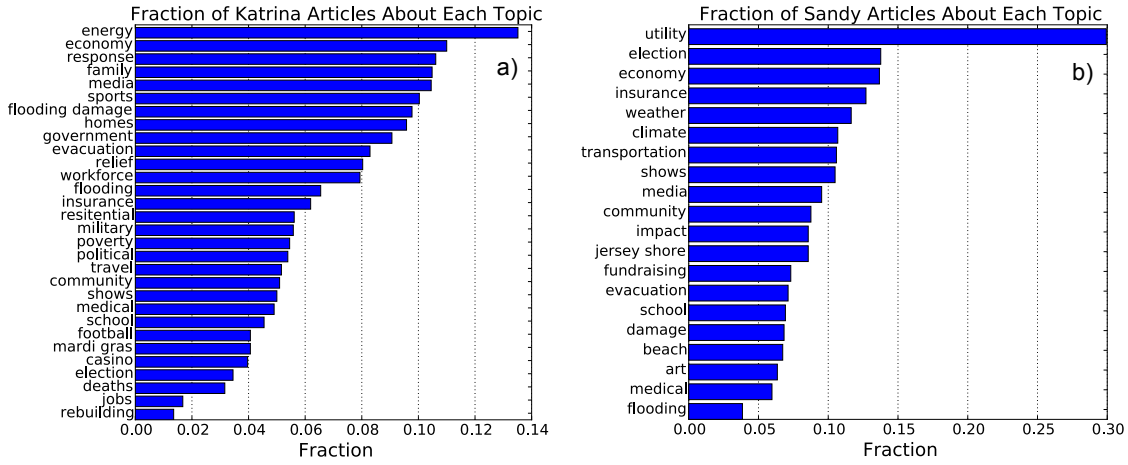

Figure 2.5: The proportion of articles ranking each topic as the first or second most probable topic, i.e., the proportion of articles that each topic appears in with high probability in the (a) Hurricane Katrina and (b) Hurricane Sandy corpora. The topics order is by decreasing proportions.

no mention of the climate within Topic 8 either, indicating that Hurricane Katrina did not only lack in climate change reporting but it also did not highlight the link between climate change and energy.

Hurricane Sandy

In the Hurricane Sandy LDA model, we see topics related to medics, insurance, fundraisers, government, damage, power outages, and climate change. Unlike the Katrina model, we find that Topic 2 clearly represents climate change. Words such as "flood", "weather", and "natural" indicate that the reporting on climate change within articles about Hurricane Sandy discussed how climate change is contributing to weather extremes and natural disasters. There was also considerable reporting on the rising sea levels, which are expected to contribute to the intensity of hurricanes and tropical storms [36].

Hurricane Katrina				
7: deaths	8: energy	12: relief	13: family	14: mardi gras
bodi	price	red	famili	gras
death	oil	cross	home	mardi
offici	percent	donat	children	french
state	energi	relief	day	restaur
home	gas	organ	live	parad
die	gasolin	volunt	back	street
victim	rate	victim	school	back
peopl	market	fund	mother	peopl
famili	week	peopl	friend	quarter
parish	product	million	peopl	time
st	month	chariti	im	home
louisiana	consum	disast	call	day
identifi	report	american	hous	citi
morgu	economi	money	stay	make
relat	compani	group	time	club
coron	increas	rais	dont	louisiana
dr	gulf	effort	work	cook
dead	fuel	food	life	krew
found	expect	org	son	hotel
remain	gallon	shelter	left	celebr
16: shows	19: travel	21: insurance	28: evacuation	29: response
music	ship	insur	hous	fema
jazz	airlin	flood	evacue	respons
band	show	damag	fema	feder
musician	news	billion	peopl	agenc
art	time	state	offici	brown
cultur	northrop	compani	home	disast
museum	network	loss	houston	govern
perform	travel	mississippi	feder	emerg
play	air	home	agenc	secur
festiv	nbc	homeown	hotel	offici
artist	million	pay	trailer	homeland
song	broadcast	claim	famili	hous
work	report	cost	state	depart
show	abc	allstat	shelter	report
time	cruis	area	emerg	manag
concert	program	properti	live	chertoff
includ	film	louisiana	month	white
orchestra	channel	industri	apart	bush
event	televis	feder	govern	plan
record	navi	polici	assist	investig

Table 2.3: The 20 most probable words within 10 of the 30 topic distributions given by LDA for Hurricane Katrina. The words are stemmed according to a Porter stemmer, where for example "flooded", "flooding", and "floods" all become "flood".

Dispersed throughout the weather related words in Topic 2, we see the words "energy", "power", and "develop", indicating that power outages and energy system development were often discussed within articles that mentioned climate change, highlighting a link between climate change and the energy disruption caused by Hurricane Sandy. Extending the number of words in Topic 2 we find more energy related words including "infrastructure" (23), "carbon" (28), "resilience" (35), and "emissions" (37). A list of the 100 most probable words in Topic 2 is given in the Supplementary Information. While "carbon" and "emissions" are clearly linked to climate change, words like "infrastructure" and "resilience" indicate a link between climate change discussion and energy system vulnerability.

Topic 0 also contains words pertaining to energy systems. This topic, however, does not contain any words pertaining to climate change. Topic 0 is about electricity ("company", "electricity", "system"), power outages ("power","utility", "service"), and communication ("verizon", "phone", "network"). One benefit of LDA is that the model not only creates distributions of words over topics, but also distributions of topics over documents. Of the 162 articles that are made up of more than 1% Topic 2, 24 of them also contain Topic 0, demonstrating that these two topics were sporadically reported on in the same article. For example, an article in *The New York Times* entitled "Experts Advise Cuomo on Disaster Measures" discusses how New York City can better prepare for drastic outages caused by extreme weather and directly quotes Governor Cuomo's concerns about climate change:

> " 'Climate change is dramatically increasing the frequency and the severity of these situations,' Mr. Cuomo said. 'And as time goes on, we're more and more realizing that these crises are more frequent and worse than

Hurricane Sandy				
0: utility	1: election	2: climate	3: community	7: transportation
power	obama	climat	hous	train
util	romney	flood	home	author
servic	presid	chang	water	station
compani	campaign	protect	beach	line
author	elect	build	car	servic
electr	state	rise	live	tunnel
island	republican	sea	flood	jersey
custom	vote	water	peopl	gas
state	polit	risk	point	transport
system	governor	level	fire	power
grid	voter	energi	street	damag
long	day	natur	rockaway	subway
verizon	poll	power	back	street
nation	democrat	weather	day	manhattan
work	peopl	develop	insur	offici
phone	debat	make	damag	transit
commiss	candid	cost	resid	long
network	presidenti	state	work	system
con	time	plan	famili	day
edison	nation	surg	neighborhood	island
8: medical	9: insurance	12: impact	13: media	15: fundraising
hospit	insur	wind	show	concert
home	compani	power	time	perform
patient	percent	day	stewart	ticket
health	sale	close	peopl	music
medic	month	weather	make	show
nurs	market	coast	photo	million
evacu	busi	expect	live	money
emerg	increas	servic	twitter	benefit
center	million	travel	call	hall
dr	loss	area	work	rais
peopl	industri	offici	news	song
citi	home	peopl	stori	peopl
offici	report	state	includ	night
resid	expect	damag	inform	work
island	billion	flood	magazin	relief
day	rate	nation	photograph	refund
care	week	massachusett	design	springsteen
bird	retail	center	post	jersey
mayor	consum	report	print	sale
mold	claim	hour	page	band

Table 2.4: The 20 most probable words within 10 of the 20 topic distributions given by LDA for Hurricane Sandy. The words are stemmed according to a Porter stemmer.

anyone had predicted.' " [26]

Although the models for each hurricane generate some similar topics, there are some topics in one model that do not appear in the other. Both models give topics on politics, community, government aid, fundraisers, insurance, family, travel, medics, flooding, damage, evacuations, and energy. The Hurricane Katrina model also gives topics relating to sporting events, Mardi Gras, music, military, and the death toll, while the Sandy model gives topics relating to museums, beaches, weather, Broadway, and climate change. Many of the topics only appearing in one of the models appear there due to the hurricane's location. The climate change topic, however, appears only in the Hurricane Sandy corpus and its absence in the Hurricane Katrina corpus cannot be simply be a consequence of the different locations of the hurricanes.

2.5 Discussion

Through this analysis using topic models, we discover that climate change and energy were often discussed together within coverage of Hurricane Sandy, whereas the climate change topic is largely absent in post Hurricane Katrina reporting. This difference can be attributed in part to changing public perceptions about climate change over time. As early as 2001, the scientific consensus that climate change is occurring and resulting from human activity was legitimized by the IPCC assessment reports [19]. A 2003 national study on climate change risk perceptions, however, revealed that while most Americans demonstrate awareness of climate change, 68% considered it only a moderate risk issue more likely to impact areas far from the United States [31]. In Fall 2008 (years after Hurricane Katrina), 51% of Americans were either alarmed or

concerned about global warming [32], and in March 2012 (months before Hurricane Sandy), this number decreased to 39% [30]. In April 2013, 38% of Americans believed that people around the world are currently affected or harmed by the consequences of climate change [29]. Those in the "alarmed" and "concerned" categories are also far more likely to report that they experienced a natural disaster within the last year [30], implying a potential relationship between personal experience of consequences and the perception of climate change risks [38]. Participants in the Yale School of Forestry & Environmental Studies "Americans and Climate Change" conference in 2005 determined that since science is the main source of climate change information, there is room for misinterpretation and disconnects in society's understanding of the issue [1].

The 2004 and 2005 Atlantic hurricane seasons were among the costliest in United States history [8]. In 2004, scientists began to propose that the intensity of the latest hurricane season may be linked to global warming. However, the state of climate science at the time could not support such a hypothesis, and linkages between global warming and the impacts of hurricanes were deemed premature [43]. Media coverage of climate change often presents the scientific consensus and has influenced public opinion and risk perceptions on climate change [3]. Complexity and uncertainty within the scientific community regarding the link between climate change and hurricanes may be why climate change does not appear as a prominent topic in the 2005 news media analysis of Hurricane Katrina.

Conversely, media reporting following Hurricane Sandy did connect explicitly with climate change. By the time Hurricane Sandy occurred in 2012, climate science research had progressed and begun exploring the link between hurricanes and global

warming [33, 16, 24]. The Yale Project on Climate Change and Communications poll in March 2012 showed that a large majority of Americans believed at that time that certain weather extremes and natural disasters are caused by global warming [28]. This evolution of climate change research and public awareness is reflected in the different coverage of climate change after Hurricane Sandy.

Also unique to Hurricane Sandy coverage was the presence of climate and energy topics together. While Hurricane Katrina reporting focused on the increase in energy prices following the storm, this increase in price was not explicitly linked to the consequences of climate change within media reporting. Hurricane Katrina caused massive disruptions in oil and gas production in the Gulf of Mexico, which caused large spikes in the cost of oil and natural gas. During Katrina, 2.6 million customers lost power in Louisiana, Mississippi, Alabama, Florida, and Georgia [39]. The destruction caused by Katrina (followed shortly after by Hurricane Rita) encouraged drilling companies to upgrade their infrastructure to better withstand the forceful waves and wind from a large hurricane [20]. During Hurricane Sandy, 8.66 million customers lost power from North Carolina to Maine, and it took 10 days for the utilities to restore power to 95% of these affected customers. Reporting on these outages is reflected in the LDA climate change topic. Flooding and power outages at refineries, pipelines, and petroleum terminals in the New York Harbor area lead to gasoline shortages and prices increases [40]. These impacts illustrated some of the consequences of climate change and an increase in severity of natural disasters. Hurricane Sandy news reporting not only highlighted the consequences of climate change but also the relationship between climate change, energy, and energy system vulnerability.

2.6 Conclusion

Given that the media both shapes and reflects public discourse, this analysis characterizing stark differences in media coverage between Hurricane Katrina and Hurricane Sandy demonstrates a shift in public discourse on climate change and energy systems. Although energy systems were disrupted in both storms, the connections between energy and climate change were made much more explicitly in the post-Hurricane Sandy news coverage as compared to the post-Hurricane Katrina coverage. This shift is likely to represent multiple changes including: (1) increased public awareness and concern about climate change, (2) improved scientific understanding of the link between hurricane intensity and climate change, and (3) greater understanding of the energy system risks associated with climate change. The ways that climate and energy are connected in the media coverage also reflects a larger shift toward increasing attention towards climate change adaptation in addition to climate mitigation [22].

Our investigation presents a mathematical approach to assessing public discourse of climate and energy, one that could be applied to assessing news media of other key areas in environmental studies. This analysis focuses on Hurricanes Katrina and Sandy due to their disruption and societal impact as focusing events. Future research could expand to investigate how energy and climate are presented in other climate and energy related media coverage over time.

Bibliography

[1] Daniel R. Abbasi. *Americans and Climate Change: Closing the Gap between Science and Action*. Yale school of forestry & environmental studies publication series, 2006.

[2] Liisa Antilla. Climate of scepticism: US newspaper coverage of the science of climate change. *Global environmental change*, 15(4):338–352, 2005.

[3] Liisa Antilla. Self-censorship and science: A geographical review of media coverage of climate tipping points. *Public Understanding of Science*, 2008.

[4] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[5] Ralf Barkemeyer, Suraje Dessai, Beatriz Monge-Sanz, Barbara Gabriella Renzi, and Giulio Napolitano. Linguistic analysis of IPCC summaries for policymakers and associated coverage. *Nature Climate Change*, page 10.1038/nclimate2824, 2015.

[6] Allan Bell. Media (mis) communication on the science of climate change. *Public understanding of science*, 3(3):259–275, 1994.

[7] Michael W Berry and Murray Browne. *Understanding search engines: Mathematical modeling and text retrieval*, volume 17. Siam, 2005.

[8] John L Beven, Lixion A Avila, Eric S Blake, Daniel P Brown, James L Franklin, Richard D Knabb, Richard J Pasch, Jamie R Rhome, and Stacy R Stewart. Atlantic hurricane season of 2005. *Monthly Weather Review*, 136(3):1109–1173, 2008.

[9] Eirc S. Blake, Tom B. Kimberlian, Robert J. Berg, John P. Cangialosi, and John L. Beven. Tropical cyclone report, Hurricane Sandy. *National Hurricane Center*, 2013.

[10] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[12] Ronald Brownstein. Hard choices blow in the winds of Katrina, and now Rita. *The Los Angeles Times*, Sep 26 2005.

[13] Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLoS ONE*, 10(8), 2015.

[14] Julia B Corbett and Jessica L Durfee. Testing public (un) certainty of science media representations of global warming. *Science Communication*, 26(2):129–151, 2004.

[15] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[16] Christopher B Field. *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[17] William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, pages 1–37, 1989.

[18] Doris A Graber. *Mass media and American politics*. SAGE, 2009.

[19] David J Griggs and Maria Noguer. Climate change 2001: The scientific basis. contribution of working group i to the third assessment report of the intergovernmental panel on climate change. *Weather*, 57(8):267–269, 2002.

[20] Rob Heidrick. Hurricane season could bring higher energy prices. *Texas Enterprise*, 2013.

[21] David K. Henry, Sandra Cooke-Hull, Jacqueline Savukinas, Fenwick Yu, Nicholas Elo, and Bradford Vac Arnum. Economic impact of Hurriane Sandy: Potential economic activity lost and gained in New Jersey and New York. Technical report, U.S. Department of Commerce, 09 2013.

[22] David J Hess. Transitions in energy systems: The mitigation–adaptation relationship. *Science as Culture*, 22(2):197–203, 2013.

[23] Peter D Howe, Hilary Boudet, Anthony Leiserowitz, and Edward W Maibach. Mapping the shadow of experience of extreme weather events. *Climatic Change*, 127(2):381–389, 2014.

[24] Daniel G. Huber and Jay Gulledge. *Extreme weather and climate change: Understanding the link, managing the risk*. Pew Center on Global Climate Change Arlington, 2011.

[25] IPCC. Climate change 2014 mitigation of climate change, intergovernmental panel on climate change. 2014.

[26] Thomas Kaplan. Experts advise Cuomo on disaster measures. *The New York Times*, January 4 2013.

[27] RD Knabb, JR Rhome, and DP Brown. Tropical cyclone report. Hurricane Katrina. National Hurricane Center. *Miami, FL*, 2006.

[28] A Leiserowitz, E Maibach, C Roser-Renouf, and JD Hmielowski. Extreme weather, climate & preparedness in the American mind. *Yale University and George Mason University. New Haven, CT.)(Report)*, 2012.

[29] Anthony Leiserowitz, Edward Maibach, Connie Roser-Renouf, Geoff Feinberg, and Peter Howe. Climate change in the American mind: Americans' global warming beliefs and attitudes in April, 2013. *Yale University and George Mason University. New Haven, CT: Yale Project on Climate Change Communication*, 2013.

[30] Anthony Leiserowitz, Edward Maibach, Connie Roser-Renouf, and Nicholas Smith. Global warming's six americas, March 2012 and November 2011. *Yale University and George Mason University*, 2012.

[31] Anthony A Leiserowitz. American risk perceptions: Is climate change dangerous? *Risk analysis*, 25(6):1433–1442, 2005.

[32] Edward W Maibach, Anthony Leiserowitz, Connie Roser-Renouf, and CK Mertz. Identifying like-minded audiences for global warming public engagement campaigns: An audience segmentation analysis and tool development. *PLoS ONE*, 6(3):e17571, 2011.

[33] Michael E. Mann and Kerry A. Emanuel. Atlantic hurricane trends linked to climate change. *Eos, Transactions American Geophysical Union*, 87(24):233–241, 2006.

[34] Media Insight Project. How Americans get their news. *The Personal News Cycle*, 2014.

[35] Bert Metz. *Controlling climate change*. Cambridge University Press, 2009.

[36] William K Michener, Elizabeth R Blood, Keith L Bildstein, Mark M Brinson, and Leonard R Gardner. Climate change, hurricanes and tropical storms, and rising sea level in coastal wetlands. *Ecological Applications*, 7(3):770–801, 1997.

[37] David Murray, Joel B Schwartz, and S Robert Lichter. *It ain't necessarily so: How media make and unmake the scientific picture of reality*. Rowman & Littlefield, 2001.

[38] Teresa A. Myers, Edward W. Maibach, Connie Roser-Renouf, Karen Akerlof, and Anthony A. Leiserowitz. The relationship between personal experience and belief in the reality of global warming. *Nature Climate Change*, 3(4):343–347, 2013.

[39] U.S. Department of Energy. Hurricane Katrina situation report #11. *Office of Electricity Delivery and Energy Reliability (OE)*, 2005.

[40] U.S. Department of Energy. Comparing the impacts of northeast hurricanes on energy infrastructure. *Office of Electricity Delivery and Energy Reliability (OE)*, 2013.

[41] TR Peterson and JL Thompson. Environmental risk communication: Responding to challenges of complexity and uncertainty. *Handbook of risk and crisis communication (pp. 591Á606). New York: Routledge*, 2009.

[42] Tuan Q Phan and Edoardo M Airoldi. A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences*, 112(21):6595–6600, 2015.

[43] Roger A Pielke Jr, Chris Landsea, Max Mayfield, Jim Laver, and Richard Pasch. Hurricanes and global warming. *Bulletin of the American Meteorological Society*, 86(11):1571–1575, 2005.

[44] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[45] Susanna Hornig Priest. *Doing media research: An introduction*. Sage, 2009.

[46] Laurie A. Rudman, Meghan C. McLean, and Martin Bunzl. When truth is personally inconvenient, attitudes change the impact of extreme weather on implicit support for green politicians and explicit climate-change beliefs. *Psychological science*, 2013.

[47] Jennie C Stephens, Gabriel M Rand, and Leah L Melnick. Wind energy in us media: A comparative state-level analysis of a critical climate change mitigation technology. *Environmental Communication*, 3(2):168–190, 2009.

[48] Jennie C Stephens, Elizabeth J Wilson, Tarla R Peterson, and James Meadowcroft. Getting smart? Climate change and the electric grid. *Challenges*, 4(2):201–216, 2013.

[49] Elizabeth J Wilson, Jennie C Stephens, Tarla Rai Peterson, and Miriam Fischlein. Carbon capture and storage in context: The importance of state policy and discourse in deploying emerging energy technologies. *Energy Procedia*, 1(1):4519–4526, 2009.

[50] Kris M Wilson. Drought, debate, and uncertainty: Measuring reporters' knowledge and ignorance about climate change. *Public Understanding of Science*, 9(1):1–13, 2000.

Chapter 3

Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll

3.1 Abstract

The consequences of anthropogenic climate change are extensively debated through scientific papers, newspaper articles, and blogs. Newspaper articles may lack accuracy, while the severity of findings in scientific papers may be too opaque for the public to understand. Social media, however, is a forum where individuals of diverse backgrounds can share their thoughts and opinions. As consumption shifts from old media to new, Twitter has become a valuable resource for analyzing current events and headline news. In this research, we analyze tweets containing the word "climate" collected between September 2008 and July 2014. Through use of a previously

developed sentiment measurement tool called the Hedonometer, we determine how collective sentiment varies in response to climate change news, events, and natural disasters. We find that natural disasters, climate bills, and oil-drilling can contribute to a decrease in happiness while climate rallies, a book release, and a green ideas contest can contribute to an increase in happiness. Words uncovered by our analysis suggest that responses to climate change news are predominately from climate change activists rather than climate change deniers, indicating that Twitter is a valuable resource for the spread of climate change awareness.

3.2 Introduction

After decades receiving little attention from non-scientists, the impacts of climate change are now widely discussed through a variety of mediums. Originating from scientific papers, newspaper articles, and blog posts, a broad spectrum of climate change opinions, subjects, and sentiments exist. Newspaper articles often dismiss or sensationalize the effects of climate change due to journalistic biases including personalization, dramatization and a need for novelty [6]. Scientific papers portray a much more realistic and consensus view of climate change. These views, however, do not receive widespread media attention due to several factors including journal paywalls, formal scientific language, and technical results that are not easy for the general public to understand [6].

According to the IPCC Fifth Assessment report, humans are "very likely" (90-100% probability) to be responsible for the increased warming of our planet [13], and this anthropogenic global warming is responsible for certain weather extremes

[15]. In April 2013, 63% of Americans reported that they believe climate change is happening. This number, however, drops to 49% when asked if climate change is being caused by humans. The percentage drops again to 38% when asked if people around the world are currently being harmed by the consequences of climate change [24]. These beliefs and risk perceptions can vary by state or by county [18]. By contrast, 97% of active, publishing, climate change scientists agree that "human activity is a significant contributing factor in changing mean global temperatures" [12, 2]. The general public learns most of what it knows about science from the mass-media [44]. Coordination among journalists, policy actors, and scientists will help to improve reporting on climate change, by engaging the general public and creating a more informed decision-making process [5].

One popular source of climate information that has not been heavily analyzed is social media. The Pew Research Center's Project for Excellence in Journalism in January of 2009 determined that topics involving global warming are much more prominent in the new, social media [5]. In the last decade, there has been a shift from the consumption of traditional mass media (newspapers and broadcast television) to the consumption of social media (blog posts, Twitter, etc.). This shift represents a switch in communications from "one-to-many" to "many-to-many" [5]. Rather than a single journalist or scientist telling the public exactly what to think, social media offers a mechanism for many people of diverse backgrounds to communicate and form their own opinions. Exposure is a key aspect in transforming a social problem into a public issue [9], and social media is a potential avenue where climate change issues can be initially exposed.

Here we study the social media site Twitter, which allows its users 140 characters

to communicate whatever they like within a "tweet". Such expressions may include what individuals are thinking, doing, feeling, etc. Twitter has been used to explore a variety of social and linguistic phenomena [7, 27, 26], and used as a data source to create an earthquake reporting system in Japan [39], detect influenza outbreaks [3], and analyze overall public health [36]. An analysis of geo-tagged Twitter activity (tweets including a latitude and longitude) before, during, and after Hurricane Sandy using keywords related to the storm is given in [22]. They discover that Twitter activity positively correlates with proximity to the storm and physical damage. It has also been shown that individuals affected by a natural disaster are more likely to strengthen interactions and form close-knit groups on Twitter immediately following the event [37]. Twitter has also been used to examine human sentiment through analysis of variations in the specific words used by individuals. In [11], Dodds et al. develop the "hedonometer", a tool for measuring expressed happiness – positive and negative sentiment – in large-scale text corpora. Since its development, the hedonometer has been implemented in studies involving the happiness of cities and states [30], the happiness of the English language as a whole [21], and the relationship between individuals' happiness and that of those they connect with [4].

The majority of the topics trending on Twitter are headlines or persistent news [23], making Twitter a valuable source for studying climate change opinions. For example, in [1], subjective vs objective and positive vs negative tweets mentioning climate change are coded manually and analyzed over a one year time period. In [43], various climate hashtags are utilized to locate pro/denialist communities on Twitter. In the present study, we apply the hedonometer to a collection of tweets containing the word "climate". We collected roughly 1.5 million such tweets from Twitter's

gardenhose API (a random 10% of all messages) during the roughly 6 year period spanning September 14, 2008 through July 14, 2014. This time period represents the extent of our database at the time of writing. Each collected tweet contains the word "climate" at least once. We include retweets in the collection to ensure an appropriately higher weighting of messages authored by popular accounts (e.g. media, government). We apply the hedonometer to the climate tweets during different time periods and compare them to a reference set of roughly 100 billion tweets from which the climate-related tweets were filtered. We analyze highest and lowest happiness time periods using word shift graphs developed in [11], and we discuss specific words contributing to each happiness score.

3.3 Methods

The hedonometer is designed to calculate a happiness score for a large collection of text, based on the happiness of the individual words used in the text. The instrument uses sentiment scores collected by Kloumann et al. and Dodds et al. [21, 11], where 10,222 of the most frequenly used English words in four disparate corpora were given happiness ratings using Amazon's Mechanical Turk online marketplace. Fifty participants rated each word, and the average rating becomes the word's score. Each word was rated on a scale from 1 (least happy) to 9 (most happy) based on how the word made the participant feel. We omit clearly neutral or ambiguous words (scores between 4 and 6) from the analysis. In the present study, we use the instrument to measure the average happiness of all tweets containing the word "climate" from September 14, 2008 to July 14, 2014 on the timescales of day, week, and month. The

word “climate” has a score of 5.8 and was thus not included when calculating average happiness. For comparison, we also calculate the average happiness score surrounding 5 climate related keywords.

We recognize that not every tweet containing the word “climate” is about climate change. Some of these tweets are about the economic, political, or social climate and some are ads for climate controlled cars. Through manual coding of a random sample of 1,500 climate tweets, we determined that 93.5% of tweets containing the word “climate” are about the earth’s climate or climate change. We calculated the happiness score for both the entire sample and the sample with the non-earth related climate tweets removed. The scores were 5.905 and 5.899 respectively, a difference of 0.1%. This difference is small enough to conclude that the non-earth related climate change tweets do not substantially alter the overall happiness score.

Based on the happiness patterns given by the hedonometer analysis, we select specific days for analysis using word shift graphs. We use word shift graphs to compare the average happiness of two pieces of text, by rank ordering the words that contribute the most to the increase or decrease in happiness. In this research, the comparison text is all tweets containing the word “climate”, and the reference text is a random 10% of all tweets. Hereafter, we refer to the full reference collection as the “unfiltered tweets”.

Finally, we analyze four events including three natural disasters and one climate rally using happiness time series and word shift graphs. These events include Hurricane Irene (August 2011), Hurricane Sandy (October 2012), a midwest tornado outbreak (May 2013), and the Forward on Climate Rally (February 2013).

3.4 Results

Fig. 3.1 gives the raw and relative frequencies of the word "climate" over the study period. We calculate the relative frequencies by dividing the daily count of "climate" by the daily sum of the 50,000 most frequently used words in the gardenhose sample. From this figure, we can see that while the raw count increases over time, the relative frequency decreases over time. This decrease can either be attributed to reduced engagement on the issue since the maximum relative frequency in December 2009, during Copenhagen Climate Change Conference, or an increase in overall topic diversity of tweets as Twitter grows in popularity. The observed increase in raw count can largely be attributed to the growth of Twitter during the study period from approximately 1 million tweets per day in 2008 to approximately 500 million in 2014. In addition, demographic changes in the user population clearly led to a decrease in the relative usage of the word "climate".

Fig. 3.2 shows the average happiness of the climate tweets by day, by week, and by month during the 6 year time span. The average happiness of unfiltered tweets is shown by a dotted red line. Several high and low dates are indicated in the figure. The average happiness of tweets containing the word "climate" is consistently lower than the happiness of the entire set of tweets.

Several outlier days, indicated on the figure, do have an average happiness higher than the unfiltered tweets. Upon recovering the actual tweets, we discover that on March 16, 2009, for example, the word "progress" was used 408 times in 479 overall climate tweets. "Progress" has a happiness score of 7.26, which increases the average happiness for that particular day. Increasing the time period for which the average

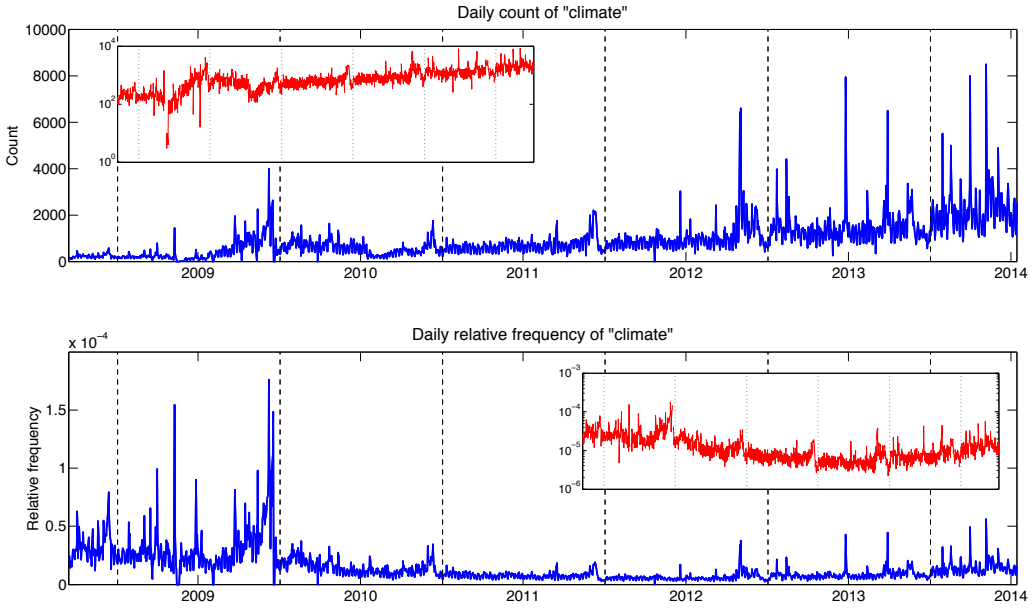

Figure 3.1: The daily raw frequencies (top) and relative frequencies (bottom) of the word "climate" on Twitter from September 14, 2008 to July 14, 2014. The insets (in red) show the same quantity with a logarithmically spaced y-axis.

happiness is measured (moving down the panels in Fig. 3.2), the outlier days become less significant, and there are fewer time periods when the climate tweets are happier than the reference tweets. After averaging weekly and monthly happiness scores, we see other significant dates appearing as peaks or troughs in Fig. 3.2. For example, the week of October 28, 2012 appears as one of the saddest weeks for climate discussion on Twitter. This is the week when Hurricane Sandy made landfall on the east coast of the U.S. For the same reason, October 2012 also appears as one of the saddest months for climate discussion.

The word shift graph in Fig. 3.3 shows which words contributed most to the shift in happiness between climate tweets and unfiltered tweets. The total average happiness of the reference text (unfiltered tweets) is 5.99 while the total average happiness of the comparison text (climate tweets) is 5.84. This change in happiness is due to the

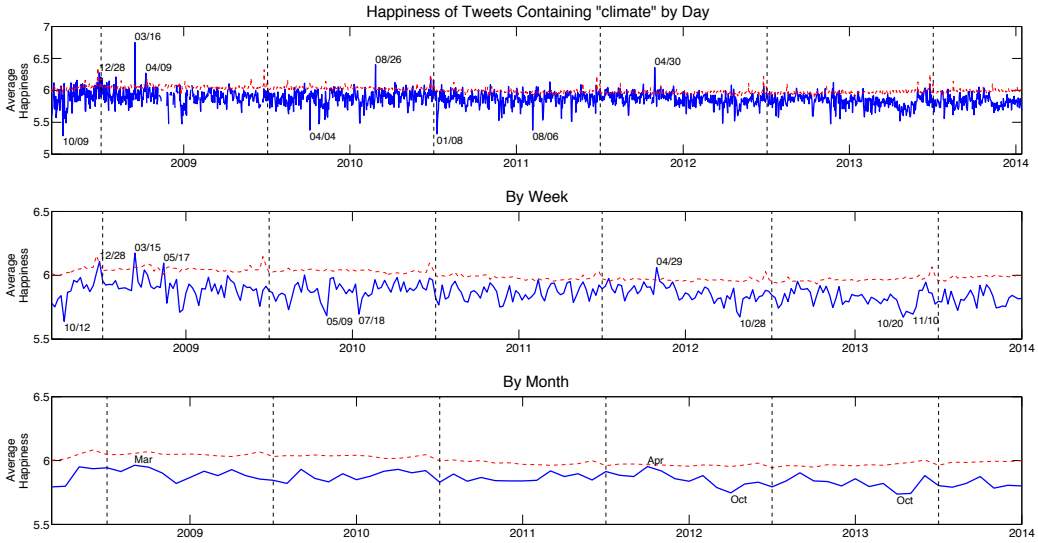

Figure 3.2: Average happiness of tweets containing the word "climate" from September 2008 to July 2014 by day (top), by week (middle), and by month (bottom). The average happiness of all tweets during the same time period is shown with a dotted red line. Several of the happiest and saddest dates are indicated on each plot, and are explored in subsequent figures.

fact that many positively rated words are used less and many negatively rated words are used more when discussing the climate.

The word "love" contributes most to the change in happiness. Climate change is not typically a positive subject of discussion, and tweets do not typically profess love for it. Rather, people discuss how climate change is a "fight", "crisis", or a "threat". All of these words contribute to the drop in happiness. Words such as "pollution", "denial", "tax", and "war" are all negative, and are used relatively more frequently in climate tweets, contributing to the drop in happiness. The words "disaster" and "hurricane" are used more frequently in climate tweets, suggesting that the subject of climate change co-occurs with mention of natural disasters, and strong evidence exists proving Twitter is a valid indicator of real time attention to natural disasters [38].

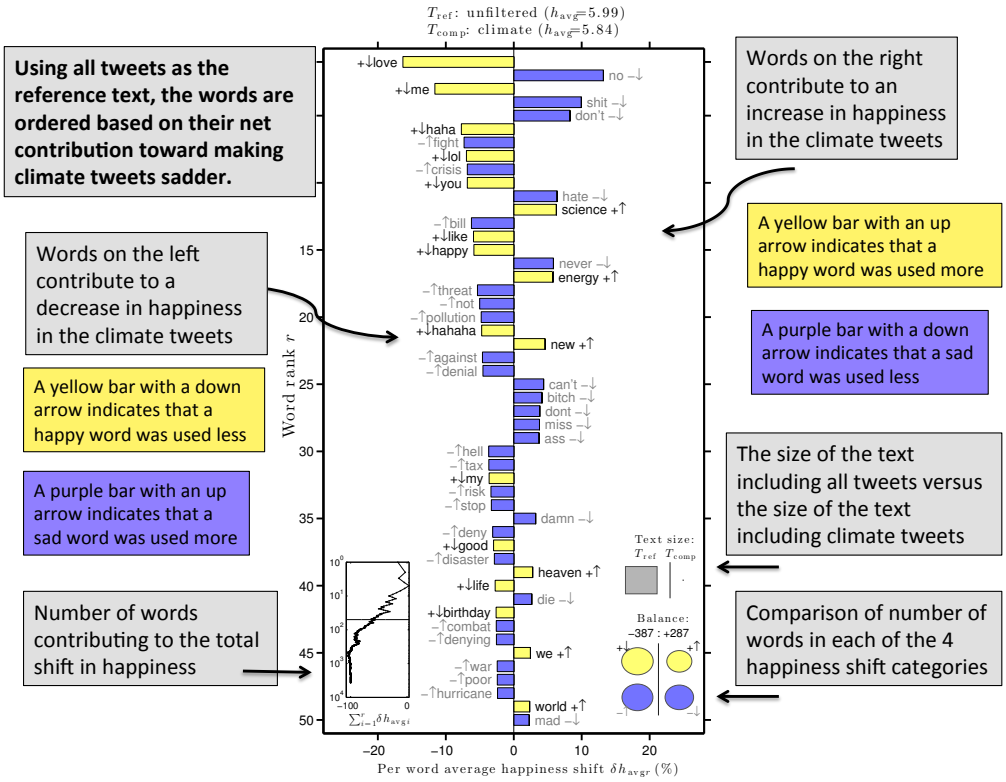

Figure 3.3: A word shift graph comparing the happiness of tweets containing the word "climate" to all unfiltered tweets. The reference text is roughly 100 billion tweets from September 2008 to July 2014. The comparison text is tweets containing the word "climate" from September 2008 to July 2014. A yellow bar indicates a word with an above average happiness score. A purple bar indicates a word with below average happiness score. A down arrow indicates that this word is used less within tweets containing the word "climate". An up arrow indicates that this word is used more within tweets containing the word "climate". Words on the left side of the graph are contributing to making the comparison text (climate tweets) less happy. Words on the right side of the graph are contributing to making the comparison text more happy. The small plot in the lower left corner shows how the individual words contribute to the total shift in happiness. The gray squares in the lower right corner compare the sizes of the two texts, roughly 10^7 vs 10^{12} words. The circles in the lower right corner indicate how many happy words were used more or less and how many sad words were used more or less in the comparison text.

On the positive side, we see that relatively less profanity is used when discussing the climate, with the exception of the word "hell". We also see that "heaven" is used more often. From our inspection of the tweets, it is likely that these two words appear

because of a famous quote by Mark Twain: “Go to heaven for the climate and hell for the company” [42]. Of the 97 non-earth related climate tweets from our 1,500 tweet sample, 8 of them referenced this quote. The word ”energy” is also used more during climate discussions. This indicates that there may be a connection between energy related topics and climate related topics. As energy consumption and types of energy sources can contribute to climate change, it is not surprising to see the two topics discussed together.

Using the first half of our dataset, Dodds et. al. [11] calculated the average happiness of tweets containing several individual keywords including “climate”. They found that tweets containing the word “climate” were, on average, similar in ambient happiness to those containing the words “no”, “rain”, “oil”, and “cold” (see Table 2 [11]). In the following section, we compare the happiness score of tweets containing the word “climate” to that of 5 other climate-related keywords.

3.4.1 Climate Related Keywords

The diction used to describe climate change attitudes on Twitter may vary by user. For example, some users may consistently use “climate change” and others may use “global warming”. There are also cohorts of users that utilize various hashtags to express their climate change opinions. In order to address this, we collected tweets containing 5 other climate related keywords to explore the variation in sentiment surrounding different types of climate related conversation. As in [43], we choose to analyze the keywords “global warming” (5.72), “globalwarming” (5.81), “climaterealists” (5.79), “climatechange” (5.86), and “agw” (5.73, standing for “anthropogenic global warming”). Search terms lack spaces in the cases where they are climate related

hashtags.

Tweets including the "global warming" keyword contain more negatively rated words than tweets including "climate". There is more profanity within these tweets and there are also more words suggesting that climate change deniers use the term "global warming" more often than "climate change". For example, there is more usage of the words "stop", "blame", "freezing", "fraud", and "politicians" in tweets containing "global warming". These tweets also show less frequent usage of positive words "science" and "energy", indicating that climate change science is discussed more within tweets containing "climate". We also see a decrease in words such as "crisis", "bill", "risk", "denial", "denying", "disaster", and "threat". The positively rated words "real" and "believe" appear more in "global warming" tweets, however so does the word "don't", again indicating that in general, the Twitter users who who don't acknowledge climate change use the term "global warming" more frequently than "climate change". A study in 2011 determined that public belief in climate change can depend on whether the question uses "climate change" or "global warming" [40].

Tweets containing the hashtag "globalwarming" also contain words indicating that this is often a hashtag used by deniers. The word contributing most to the decrease in happiness between "climate" and "globalwarming" is "fail", possibly referencing an inaccurate interpretation of the timescale of global warming consequences during cold weather. We see an increase in negative words "fraud", "die", "lie", "blame", "lies", and again a decrease in positive, scientific words. There is also an increase in several cold weather words including "snow", "freezing", "christmas", "december", indicating that the "globalwarming" hashtag may often be used sarcastically. Similarly, Tweets including the hashtag "climaterealists" use more words like "fraud", "lies", "wrong",

and “scandal” and less “fight”, “crisis”, “pollution”, “combat”, and “threat”.

The hashtag “agw” represents a group that is even more so against anthropogenic climate change. We see an increase in “fraud”, “lie”, “fail”, “wrong”, “scare”, “scandal”, “conspiracy”, “crime”, “false”, and “truth”. This particular hashtag gives an increase in positive words “green” and “science”, however based on the large increase in the aforementioned negative words, we can deduce that these terms are being discussed in a negative light. The “climatechange” hashtag represents users who are believers in climate change. There is an increase in positive words “green”, “energy”, “environment”, “sea”, “oceans”, “nature”, “earth”, and “future”, indicating a discussion about the environmental impacts of climate change. There is also an increase in “pollution”, “threat”, “risk”, “hunger”, “fight”, and “problem” indicating that the “climatechange” hashtag is often used when tweeting about the fight against climate change.

With the exception of the “globalwarming” hashtag, our analysis of these keywords largely agrees with what is found in [43]. Our analysis, however, compares word frequencies within tweets containing these hashtags with word frequencies within tweets containing the word “climate”. We find that more skeptics use “global warming” in their tweets than “climate”, while it may be the case that ”global warming” and “globalwarming” hashtag are also used by activists.

3.4.2 Analysis of Specific Dates

While Fig. 3.3 shows a shift in happiness for all climate tweets collected in the 6 year period, we now move to analyzing specific climate change-related time periods and events that correspond to spikes or dips in happiness. It is important to note

that tweets including the word "climate" represent a very small fraction of unfiltered tweets (see gray squares comparing text sizes in bottom right of Fig. 3.3). While our analysis may capture specific events pertaining to climate change, it may not capture everything, as Twitter may contain background noise that we can't easily analyze.

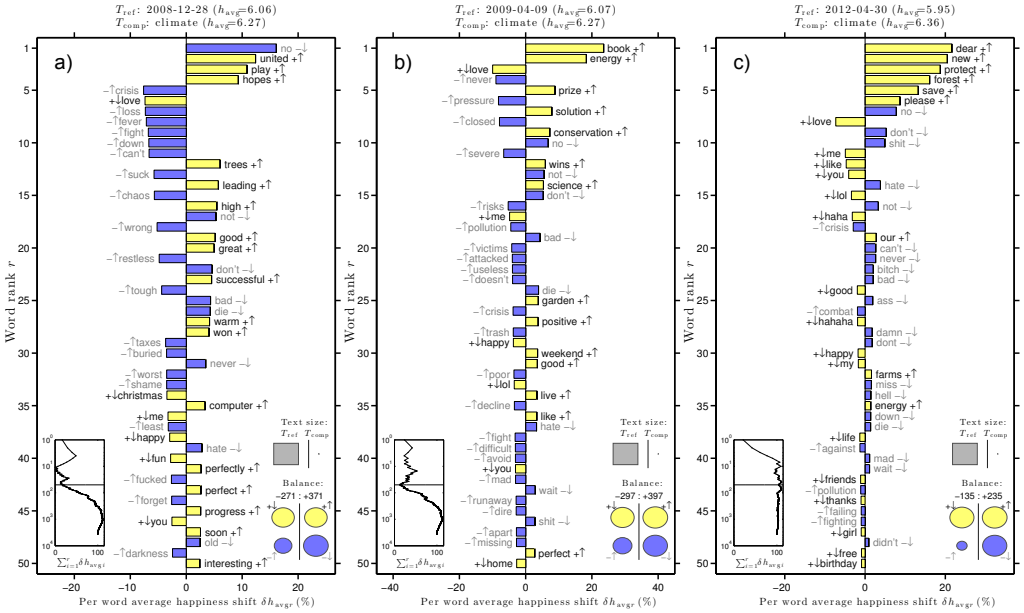

Figure 3.4: Word shift graphs for three of the happiest days in the climate tweet time series.

Fig. 3.4 gives word shift graphs for three of the happiest days according to the hedonometer analysis. These dates are indicated in the top plot in Fig. 3.2. The word shift graphs use unfiltered tweets as the reference text and climate tweets as the comparison text for the date given in each title. Fig. 3.4(a) shows that climate tweets were happier than unfiltered tweets on December 28, 2008. This is due in part to a decrease in the word "no", and an increase in the words "united", "play", and "hopes". On this day, there were "high hopes" for the U.S. response to climate change. An example tweet by OneWorld News is given in Fig. 3.5(a) [33].

Fig. 3.4(b) shows that climate tweets were happier than unfiltered tweets on April

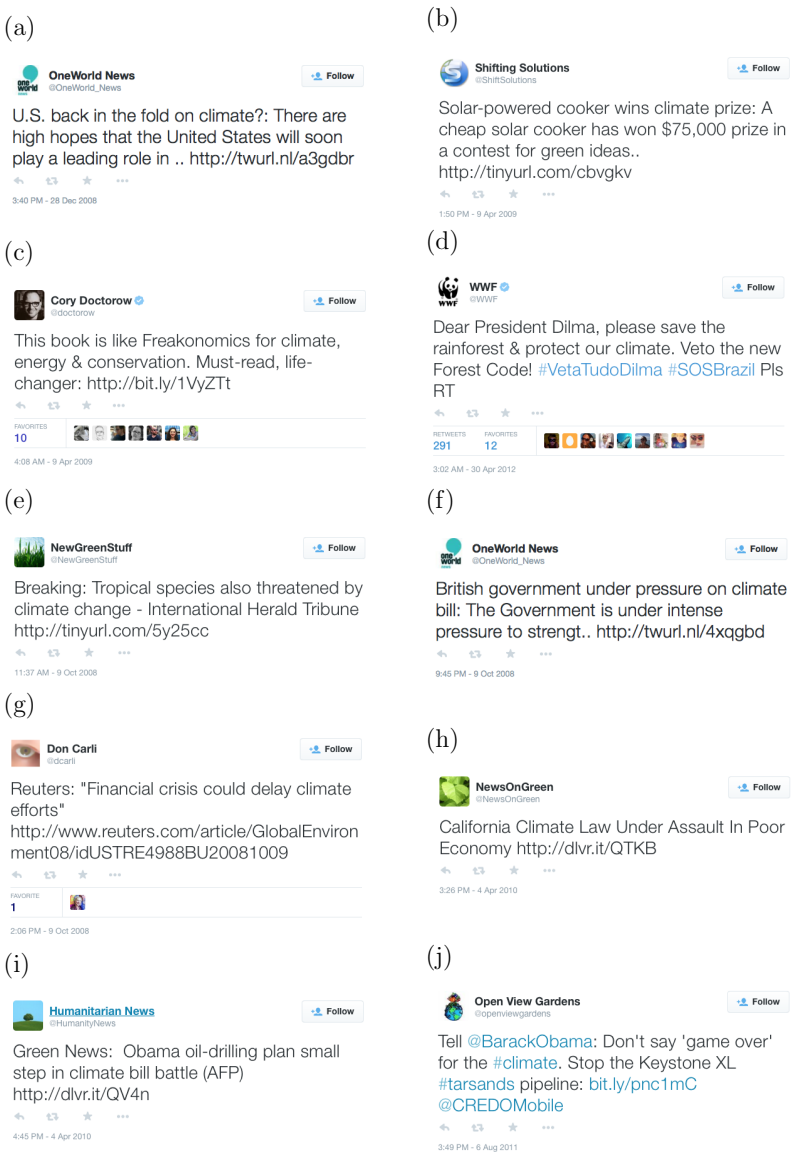

Figure 3.5: Example tweets on the happiest and saddest days for climate conversation on Twitter

9, 2009, largely due to the increase in positive words "book", "energy", and "prize". Twitter users were discussing the release of a new book called *Sustainable Energy Without the Hot Air* by David JC MacKay [28]. Also on this date, users were posting about a Climate Prize given to a solar-powered cooker in a contest for green ideas.

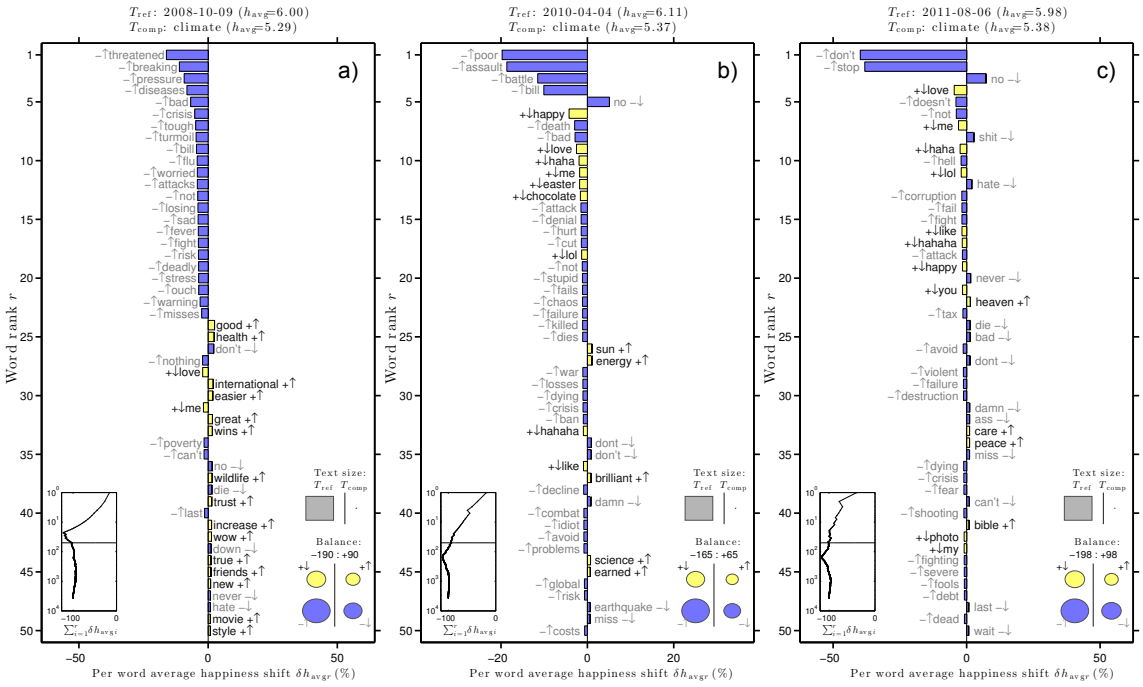

Figure 3.6: Word shift graphs for 3 of the saddest days in the climate tweet time series.

Example tweets include Fig. 3.5(b) and (c) [41, 10]. Finally, Fig. 3.4(c) shows that climate tweets were happier than unfiltered tweets on April 30, 2012. This is due to the increased usage of the words “dear”, “new”, “protect”, “forest”, “save”, and “please”. On this date, Twitter users were reaching out to Brazilian president Dilma to save the Amazon rainforest, e.g., Fig. 3.5(d) [45].

Similarly, Fig. 3.6 gives word shift graphs for three of the saddest days according to the hedonometer analysis. These dates are indicated in the top panel in Fig. 3.2. Fig. 3.6(a) shows an increase in many negative words on October 9, 2008. Topics of conversation in tweets containing “climate” include the threat posed by climate change to a tropical species, a British climate bill, and the U.S. economic crisis. Example tweets include Fig. 3.5(e-g) [31, 34, 8].

Fig. 3.6(b) shows an increase in negative words “poor”,“assault”, “battle”, and

"bill" on April 4, 2010. Popular topics of conversation on this date included a California climate law and President Obama's oil-drilling plan. Example tweets include Fig. 3.5(h) and (i) [35, 32]. Finally, Fig. 3.6(c) shows that the words "don't" and "stop" contributed most to the decrease in happiness on August 6, 2011. A topic of conversation on this date was the Keystone XL pipeline, a proposed extension to the current Keystone Pipeline. An example tweet is given in Fig. 3.5(j) [16].

This per day analysis of tweets containing "climate" shows that many of the important issues pertaining to climate change appear on Twitter, and demonstrate different levels of happiness based on the events that are unfolding. In the following section, we investigate specific climate change events that may exhibit a peak or a dip in happiness. First, we analyze the climate change discussion during several natural disasters that may have raised awareness of some of the consequences of climate change. Then, we analyze a non-weather related event pertaining to climate change.

3.4.3 Natural Disasters

Natural disasters such as hurricanes and tornados have the potential to focus society's collective attention and spark conversations about climate change. A person's belief in climate change is often correlated with the weather on the day the question is asked [46, 25, 17]. A study using "climate change" and "global warming" tweets showed that both weather and mass media coverage heavily influence belief in anthropogenic climate change [20]. In this section, we analyze tweets during three natural disasters: Hurricane Irene, Hurricane Sandy, and a midwest tornado outbreak that damaged many towns including Moore, Oklahoma and Rozel, Kansas. Fig. 3.7 gives the frequencies of the words "hurricane" and "tornado" within tweets that contain the word

"climate". Each plot labels several of the spikes with the names of the hurricanes (top) or the locations (state abbreviations) of the tornado outbreaks (bottom). This figure indicates that before Hurricane Irene in August 2011, hurricanes were not commonly referenced alongside climate, and before the April 2011 tornado outbreak in Alabama and Mississippi, tornados were not commonly referenced alongside climate.

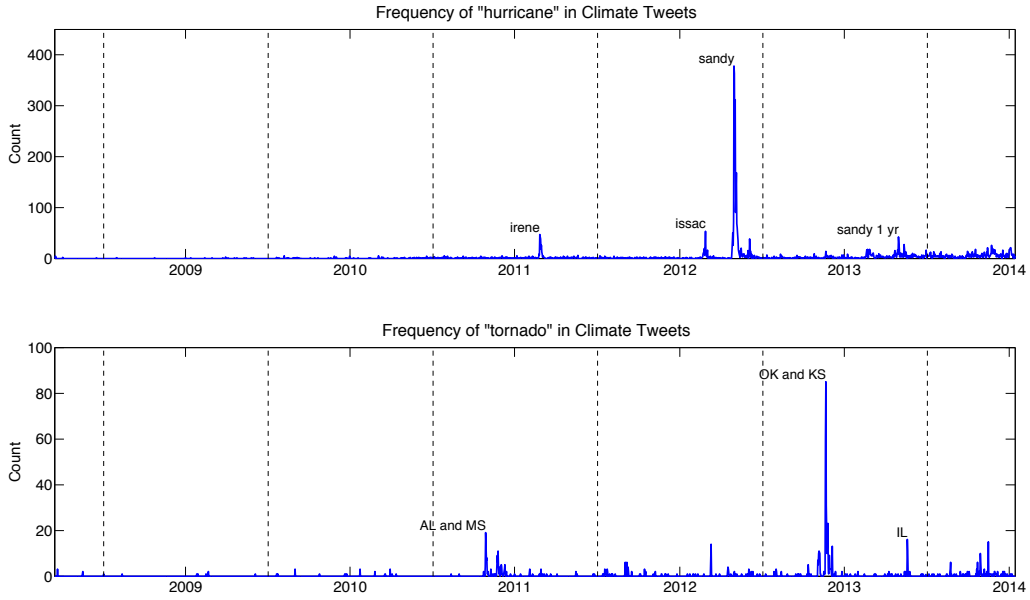

Figure 3.7: Frequency of the word "hurricane"' (top) and "tornado" (bottom) within tweets containing the word "climate". Several spikes have been identified with the hurricane or tornado that took place during that time period.

This analysis, however, will not capture every hurricane or tornado mentioned on Twitter, only those that were referenced alongside the word "climate". Hurricane Arthur, for example, occurred in early July, 2014 and does not appear as a spike in Fig. 3.7. This particular hurricane did not cause nearly as much damage or as many fatalities as the hurricanes that do appear in Fig. 3.7, and perhaps did not draw enough attention to highlight a link between hurricanes and climate change on Twitter. Additionally, a large tornado outbreak in Kentucky, Alabama, Indiana, and

Ohio occurred in early March 2012 and does not appear as a spike in our analysis.

Fig. 3.7 shows that the largest peak in the word "hurricane" occurred during Hurricane Sandy in October 2012. Fig. 3.8 provides a deeper analysis for the climate time series during hurricane Sandy. The time series of the words "hurricane" and "climate" as a fraction of all tweets before, during, and after Hurricane Sandy hit are given in Fig. 3.8(a) and (c). Spikes in the frequency of usage of these words is evident in these plots. The decay of each word is fitted with a power law in Fig. 3.8(b) and (d). A power law is a functional relationship of the following form:

$$f(t - t_{event}) = \alpha(t - t_{event})^{-\gamma} \tag{3.1}$$

Here, t is measured in days, and t_{event} is the day Hurricane Sandy made landfall. $f(t)$ represents the relative frequency of the word "hurricane" (top) or "climate" (bottom), and α and γ are constants.

Using the power law fit, we calculate the first three half lives of the decay. Letting M equal the maximum relative frequency, the time at which the first half life of the power law relationship occurs is calculated by equation 3.2:

$$t_{\frac{1}{2}} = \left(\frac{M}{2\alpha}\right)^{-\frac{1}{\gamma}} \tag{3.2}$$

The first three half lives of the decay in the frequency of the word "hurricane" during hurricane Sandy are 1.57, 0.96, and 1.56 additional days. Since the decay is not exponential, these half lives are not constant. The first half life indicates that after about a day and a half, "hurricane" was already tweeted only half as often. The second half life indicates that after one more day, "hurricane" was tweeted only one fourth

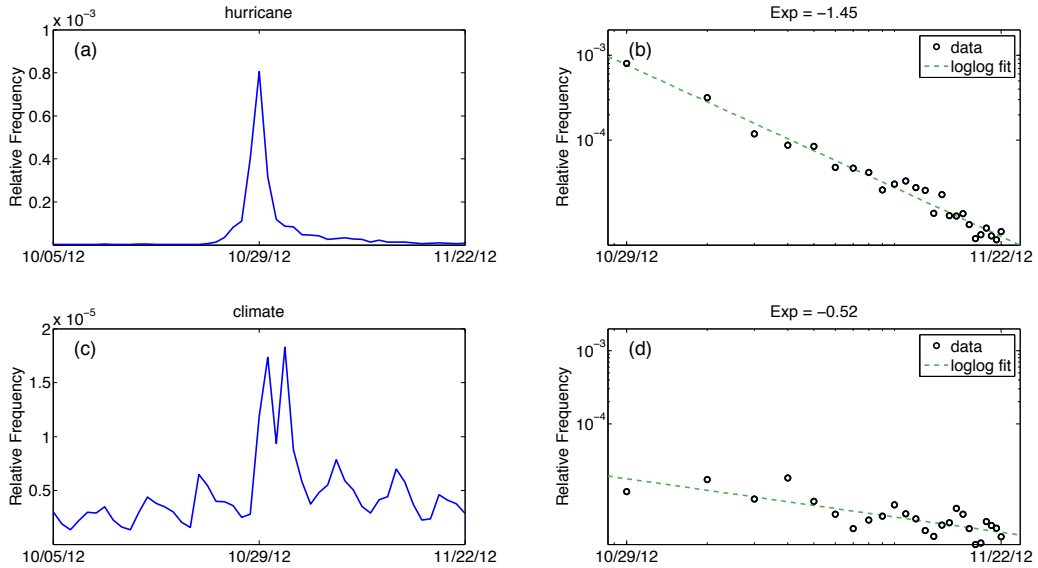

Figure 3.8: Decay rates of the words "hurricane" (top) and "climate" (bottom). The left plots gives the time series of each word during hurricane Sandy. The right plots gives the power law fit for the decay in relative frequency, x-axes are spaced logarithmically. The power law exponents are given in the titles of the figures.

as often, and so on. Thus, it did not take long for the discussion of the hurricane to decrease. The half lives, however, of the word "climate" are much larger at 8.19, 22.58, and 84.85 days.

Fig. 3.9 gives happiness time series plots for three natural disasters occurring in the United States. These plots show that there is a dip in happiness on the day that the disasters hit the affected areas, offering additional evidence that sentiment is depressed by natural disasters [1]. The word shift graphs indicate which words contributed to the dip in happiness. The circles on the bottom right of the word shift plots indicate that for all three disasters, the dip in happiness is due to an increase in negative words, more so than a decrease in positive words. During a natural disaster, tweets mentioning the word "climate" use more negative words than tweets not mentioning the word "climate".

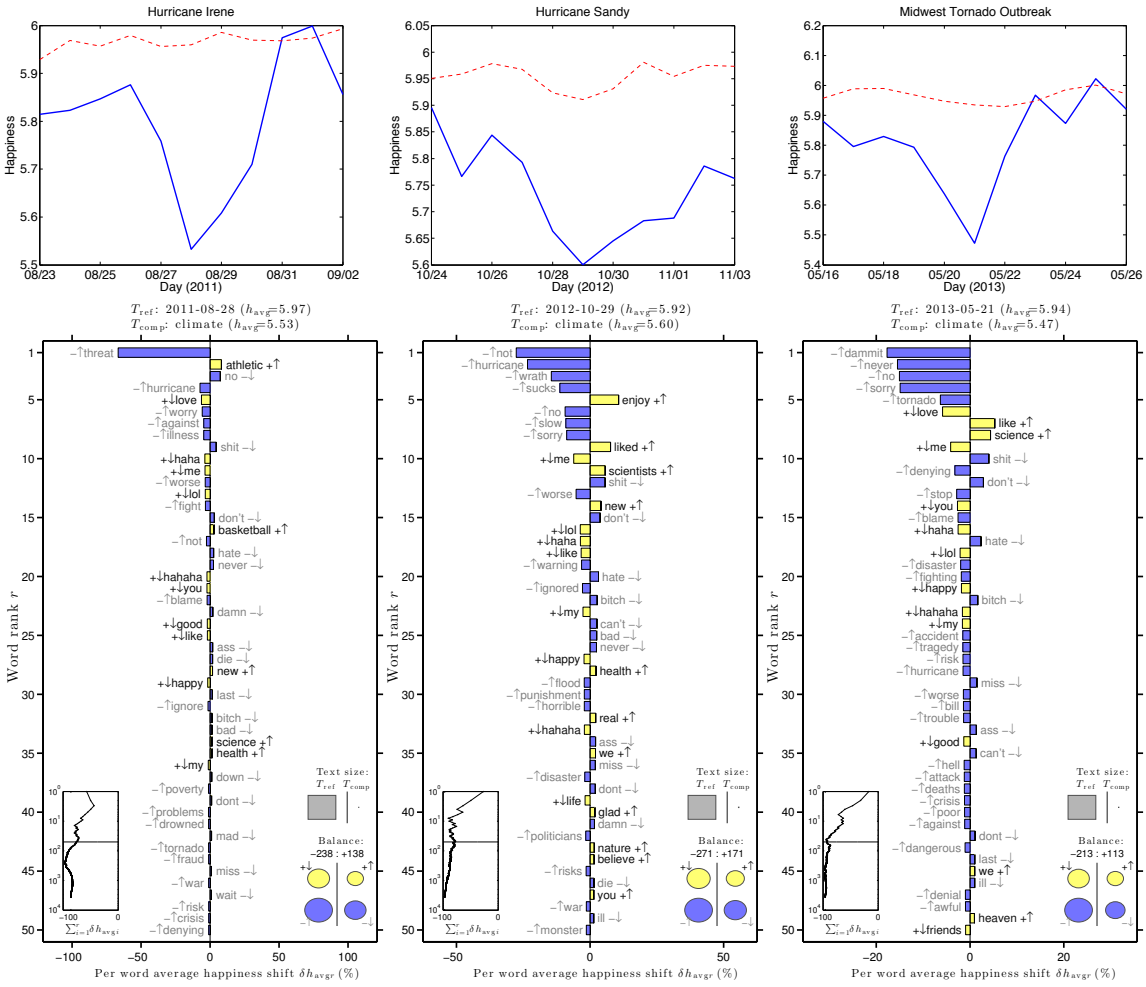

Figure 3.9: Happiness time series plots for tweets containing the word "climate" one week before and one week after three natural disasters in the United States (top) and word shift graphs indicating what words contributed most to the drop in happiness during the natural disasters (bottom). The word shift graphs compare the climate tweets to unfiltered tweets on the day of the natural disaster.

3.4.4 Forward on Climate Rally

In this section, we analyze tweets during the Forward on Climate Rally, which took place in Washington D.C. on February 17, 2013. The goal of the rally, one of the largest climate rallies ever in the United States, was to convince the government

to take action against climate change. The proposed Keystone pipeline bill was a particular focus. Fig. 3.10 shows that the happiness of climate tweets increased slightly above the unfiltered tweets during this event, which only occurs on 8% of days in Fig. 3.2.

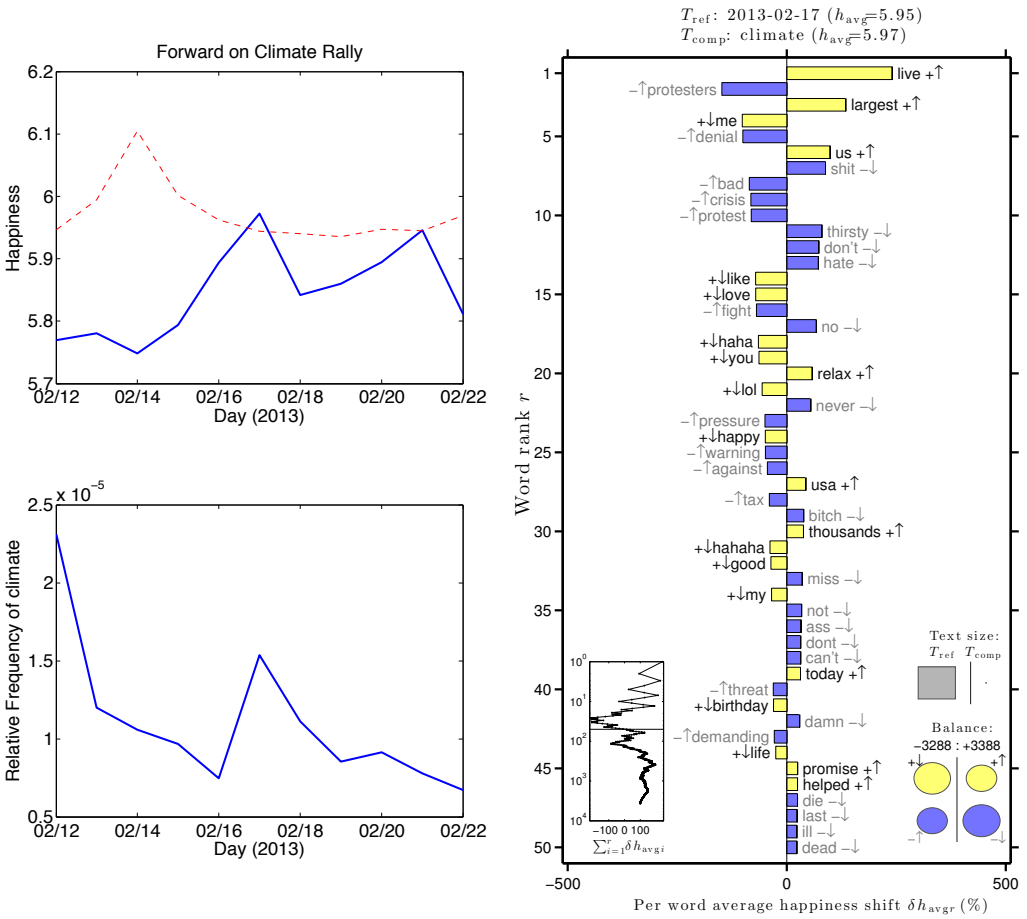

Figure 3.10: Left: Happiness time series plot for unfiltered tweets (red dashed) and tweets containing the word "climate" (blue solid) one week before and one week after the Forward on Climate Rally. Right: word shift plot for climate tweets versus unfiltered tweets on the day of the rally.

Despite the presence of negative words such as "protestors", "denial", and "crisis", the Forward on Climate Rally introduced positive words such as "live", "largest", and

"promise". The Keystone pipeline bill was eventually vetoed by President Obama.

3.5 Conclusion

We have provided a general exploration of the sentiment surrounding tweets containing the word "climate" in response to natural disasters and climate change news and events. The general public is becoming more likely to use social media as an information source, and discussion on Twitter is becoming more commonplace. We find that tweets containing the word "climate" are less happy than all tweets. In the United States, climate change is a topic that is heavily politicized; the words "deny", "denial", and "deniers" are used more often in tweets containing the word "climate". The words that appear in our climate-related tweets word shift suggest that the discussion surrounding climate change is dominated by climate change activists rather than climate change deniers, indicating that the twittersphere largely agrees with the scientific consensus on this issue. The presence of the words "science" and "scientists" in almost every word shift in this analysis also strengthens this finding (see also [1]). The decreased "denial" of climate change is evidence for how a democratization of knowledge transfer through mass media can circumvent the influence of large stakeholders on public opinion.

In examining tweets on specific dates, we have determined that climate change news is abundant on Twitter. Events such as the release of a book, the winner of a green ideas contest, or a plea to a political figure can produce an increase in sentiment for tweets discussing climate change. For example, the Forward on Climate Rally demonstrates a day when the happiness of climate conversation peaked above the

background conversation. On the other hand, consequences of climate change such as threats to certain species, extreme weather events, and climate related legislative bills can cause a decrease in overall happiness of the climate conversation on Twitter due to an increase in the use words such as “threat”, “crisis”, and “battle”.

Natural disasters are more commonly discussed within climate-related tweets than unfiltered tweets, implying that some Twitter users associate climate change with the increase in severity and frequency of certain natural disasters [29, 19, 14]. During Hurricane Irene, for example, the word “threat” was used much more often within climate tweets, suggesting that climate change may be perceived as a bigger threat than the hurricane itself. The analysis of Hurricane Sandy in Fig. 3.8 demonstrates that while climate conversation peaked during Hurricane Sandy, it persisted longer than the conversation about the hurricane itself.

While climate change news is prevalent in traditional media, our research provides an overall analysis of climate change discussion on the social media site, Twitter. Through social media, the general public can learn about current events and display their own opinions about global issues such as climate change. Twitter may be a useful asset in the ongoing battle against anthropogenic climate change, as well as a useful research source for social scientists, an unsolicited public opinion tool for policy makers, and public engagement channel for scientists.

Bibliography

[1] Xiaoran An, Auroop R Ganguly, Yi Fang, Steven B Scyphers, Ann M Hunter, and Jennifer G Dy. Tracking climate change opinions from Twitter data. *Workshop on Data Science for Social Good*, 2014.

[2] William RL Anderegg, James W Prall, Jacob Harold, and Stephen H Schneider. Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27):12107–12109, 2010.

[3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[4] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.

[5] Maxwell T Boykoff. *Who speaks for the climate?: Making sense of media reporting on climate change*. Cambridge University Press, 2011.

[6] Maxwell T Boykoff and Jules M Boykoff. Climate change and journalistic norms: A case-study of US mass-media coverage. *Geoforum*, 38(6):1190–1204, 2007.

[7] Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, 2012.

[8] Don Carli. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/dcarli/status/953288121.

[9] James W Dearing and Everett M Rogers. *Agenda-setting*, volume 6. Sage Publications, 1996.

[10] Cory Doctorow. *Twitter*, 2009 (accessed March 19, 2015). https://twitter.com/doctorow/status/1482803994.

[11] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752, 2011.

[12] Peter T Doran and Maggie Kendall Zimmerman. Examining the scientific consensus on climate change. *Eos, Transactions American Geophysical Union*, 90(3):22–23, 2009.

[13] Christopher Field and Maarten Van Aalst. *Climate change 2014: Impacts, adaptation, and vulnerability*, volume 1. IPCC, 2014.

[14] Christopher B Field. *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[15] E. M. Fischer and R. Knutti. Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nature Climate Change*, advance online publication, April 2015.

[16] Open View Gardens. *Twitter*, 2011 (accessed March 19, 2015). https://twitter.com/openviewgardens/status/99975488293978112.

[17] Lawrence C Hamilton and Mary D Stampone. Blowin' in the wind: Short-term weather and belief in anthropogenic climate change. *Weather, Climate, and Society*, 5(2):112–119, 2013.

[18] Peter D Howe, Matto Mildenberger, Jennifer R Marlon, and Anthony Leiserowitz. Geographic variation in opinions on climate change at state and local scales in the usa. *Nature Climate Change*, 2015.

[19] Daniel G. Huber and Jay Gulledge. *Extreme weather and climate change: Understanding the link, managing the risk*. Pew Center on Global Climate Change Arlington, 2011.

[20] Andrei P Kirilenko, Tatiana Molodtsova, and Svetlana O Stepchenkova. People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Global Environmental Change*, 30:92–100, 2015.

[21] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PLoS ONE*, 7(1):e29484, 2012.

[22] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Nowcasting disaster damage. *arXiv preprint arXiv:1504.06827*, 2015.

[23] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[24] Anthony Leiserowitz, Edward Maibach, Connie Roser-Renouf, Geoff Feinberg, and Peter Howe. Climate change in the American mind: Americans' global warming beliefs and attitudes in April, 2013. *Yale University and George Mason University. New Haven, CT: Yale Project on Climate Change Communication*, 2013.

[25] Ye Li, Eric J Johnson, and Lisa Zaval. Local warming daily temperature change influences belief in global warming. *Psychological Science*, 2011.

[26] Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PloS one*, 9(5):e94093, 2014.

[27] Yu-Ru Lin, Drew Margolin, Brian Keegan, and David Lazer. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*, pages 737–748. International World Wide Web Conferences Steering Committee, 2013.

[28] David MacKay. *Sustainable Energy-without the hot air*. UIT Cambridge, 2008.

[29] Michael E. Mann and Kerry A. Emanuel. Atlantic hurricane trends linked to climate change. *Eos, Transactions American Geophysical Union*, 87(24):233–241, 2006.

[30] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[31] NewGreenStuff. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/NewGreenStuff/status/953099924.

[32] Humanitarian News. *Twitter*, 2010 (accessed March 19, 2015). https://twitter.com/HumanityNews/status/11612292989.

[33] OneWorld News. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/OneWorld_News/status/1083004712.

[34] OneWorld News. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/OneWorld_News/status/953758970.

[35] NewsOnGreen. *Twitter*, 2010 (accessed March 19, 2015). https://twitter.com/NewsOnGreen/status/11608867076.

[36] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *ICWSM*, pages 265–272, 2011.

[37] Tuan Q. Phan and Edoardo M. Airoldi. A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences*, 2015.

[38] Joseph T Ripberger, Hank C Jenkins-Smith, Carol L Silva, Deven E Carlson, and Matthew Henderson. Social media and severe weather: Do tweets provide a valid indicator of public attention to severe weather risk communication? *Weather, Climate, and Society*, 6(4):520–530, 2014.

[39] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[40] Jonathon P Schuldt, Sara H Konrath, and Norbert Schwarz. "global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly*, page nfq073, 2011.

[41] Shifting Solutions. *Twitter*, 2009 (accessed March 19, 2015). https://twitter.com/ShiftSolutions/status/1485975759.

[42] Mark Twain, Michael Barry Frank, Robert Pack Browning, Lin Salamo, Frederick Anderson, and Mark Twain. *Mark Twain's Notebooks & Journals, Volume III:(1883-1891)*, volume 8. Univ of California Press, 1980.

[43] Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[44] Kris M Wilson. Mass media as sources of global warming knowledge. *Mass Comm Review*, 22:75–89, 1995.

[45] WWF. *Twitter*, 2012 (accessed March 19, 2015). https://twitter.com/WWF/status/196902312797671424.

[46] Lisa Zaval, Elizabeth A Keenan, Eric J Johnson, and Elke U Weber. How warm days increase belief in global warming. *Nature Climate Change*, 4(2):143–147, 2014.

Chapter 4

Public Opinion Polling with Twitter

4.1 Abstract

Compared to traditional opinion polling techniques, sentiment analysis of text-based data from social media has many advantages. Solicited public opinion surveys reach a limited subpopulation of willing participants and are expensive to conduct, leading to poor time resolution and a restricted pool of expert-chosen survey topics. In this study, we demonstrate that public opinion polling with Twitter correlates well with traditional measures, and has predictive power for issues of global importance. We also examine Twitter's potential to provide unsolicited public opinion polls for topics seldom surveyed, including ideas, personal feelings and perceptions of commercial enterprises. To make possible a full examination of our work and to enable others' research, we make public over 10,000 data sets, each a seven-year series of daily word counts for tweets containing a frequently used search term.

4.2 Introduction

Readily available public opinion data is valuable to researchers, policymakers, marketers, and many other groups, but is difficult to generate. Solicited public opinion polls can be expensive, prohibitively time consuming, and may only reach a limited number of people on a limited number of days. Polling accuracy evidently relies on accessing representative populations and high response rates. Poor temporal sampling will weaken any polls value as individual opinions vary in time and in response to social influence [39, 11]. Public opinion data can be used to determine public awareness, to predict outcomes of events, and to infer characteristics of human behaviors.

With the continued rise of social media as a communication platform, the ability to construct unsolicited public opinion polls has become a possibility for researchers though parsing of massive text-based datasets. Social media provides access to public opinions in real time, and has been proven to play a role in human behavior [21].

With its open platform, Twitter has proved to be a boon for many research enterprises [33], having been used to explore a variety of social and linguistic phenomena [9, 27, 26]; harnessed as a data source to create an earthquake reporting system in Japan [38]; made possible detection of influenza outbreaks [6]; and used to analyze overall public health [34]. Predictions made using Twitter have focused on elections [20, 40], the spread of disease [37], crime [42], and the stock market [31]. These studies demonstrate a proof-of-concept, avoiding the more difficult tasks of building operational systems for continued forecasting. Twitter unsuccessfully predicts the stock market in [8], where OpinionFinder is used to evaluate tweets containing direct expressions of emotion as in “I feel”, or “I am feeling”. Eliminating indirect expres-

sions of emotion is likely why the experiment did not result in a significant correlation between Twitter mood and Dow Jones Industrial Average.

Despite limitations, which we address later, Twitter data reveals unprecedented view of human behavior and opinion related to major issues of global importance[28]. In a previous study [13], we analyzed the sentiment surrounding climate change conversation on Twitter. We discovered that sentiment varies in response to climate change news and events, and that the conversation is dominated by activists. Another study by Helmuth et. al analyzed tweets by United States Senators to determine which research oriented science organizations and which senators are best at getting science-related findings into the hands of the general public [22]. Twitter is also often used to analyze public opinion of political issues [16, 7, 41], and in several previous works as an opinion polling resource. In an application using neural networks called TrueHappiness, users enter one of 300,000 words to obtain a sentiment estimation based on this word's usage in a massive Wikipedia data set, and on previously collected sentiment scores for 10,222 words on Amazon's Mechanical Turk [17, 15], hereafter referred to as the labMT word set. In another application called RACCOON, a user enters a query term and a rough sketch of what its time series may look like to obtain words or phrases on Twitter that correlate well with the inputs [5]. Google Correlate is a similar tool that discovers Google searches for terms or phrases that match well with real-world time series [2]. Financial term searches from Google Trends was shown by Preis et. al [35] to correlate with Dow Jones economic indices.

We argue that Twitter is a better source for opinion mining than Wikipedia, used in TrueHappiness, due to the personal nature of each post. RACCOON uses only user estimates for correlations and not actual survey data. Google Correlate uses

only frequencies of Google searches to compare time series and may only be useful in specific situations. Successful predictions using Twitter tend to use context or sentiment analysis of surrounding words to achieve high accuracy. Failed predictions make use of only frequencies, for example Google Flu Trends. This flu tracking algorithm often over estimated flu trends because it relied to heavily on simple Google searches.

In many studies that use text data from Twitter, the results are not compared to actual polling data, leaving the conclusions open to interpretation. One example work which does make a direct comparison is [32], where the authors use a Twitter data set from 2008 and 2009 to compare sentiments on Twitter, calculated with OpinionFinder, with daily and monthly public opinion polls. Our approach here is analogous to that of [32], but we use the sentiment analysis techniques developed in [13] to investigate public opinion regarding over 10,000 search terms.

Specifically, for each of 10,222 of the most frequently used English words, we calculate daily, weekly, monthly, and yearly sentiment time series of co-occurring words from September 2008 to November 2015. We compare many of these happiness time series to actual polling data, which is not typically available at such a high resolution in time. We investigate a wide range of topics, including politics, the economy, and several commercial organizations. Given the size of the dataset, we were unable to exhaustively compare all search terms to solicited opinion polls, and have released the data publicly along with this paper.

Here, we aim to determine if public opinion polling with Twitter can be used to complement and compare responses to traditional public opinion surveys. With both resources at our disposal, researches will have more evidence to draw specific

conclusions about public opinions.

4.3 Methods

We implement the "hedonometer", an instrument designed to calculate a happiness score for a large-scale text, based on the happiness of individual words in the text. The hedonometer uses previously assessed happiness scores for the labMT word set, which contains the most frequently used English words in four disparate corpora [24]. We choose the hedonometer to obtain lexical coverage of Twitter text and produce meaningful word shift graphs [36].

The words were scored in isolation by human subjects in previous work on a scale from 1 (least happy) to 9 (most happy). We remove neutral and ambiguous words (scores between 4 and 6) from the analysis. For details regarding stop words, see Dodds et. al, [17].

We use the hedonometer to calculate what we refer to as *ambient happiness* scores for each of the labMT words (first defined in [17]). We determine ambient happiness of a given word, w_j, by calculating the average happiness of the words that appear in tweets with that word, i.e.,

$$h_{\text{amb}}(w_j) = \frac{\sum\limits_{i=1, i\neq j}^{N} h_{\text{avg}}(w_i) f_i}{\sum\limits_{i=1, i\neq j}^{N} f_i}. \tag{4.1}$$

Here, w_i is a word that appears in a tweet with word w_j, $h_{\text{avg}}(w_i)$ is the surveyed happiness score of word i, f_i is the frequency of word i, and N is the number of words in labMT (with stop words removed) appearing in tweets containing word j. Note

that we do not include the frequencies or happiness scores of the given word (w_j) in the calculation of ambient happiness.

For example, $h_{\text{avg}} = 8.42$ for "love" and the ambient happiness for tweets containing "love" is 6.17. For "hate", $h_{\text{avg}} = 2.34$ and we find the ambient happiness of "hate" is 5.75. As seen in the Appendix in Fig. B.1, we find that due to averaging, ambient happiness covers a smaller range of scores than labMT happiness.

We use the ambient happiness scores to create time series for each of the words in the labMT word set, and we correlate the happiness time series with polling data at various temporal resolutions.

4.3.1 Data

We collected tweets from Twitter's gardenhose API from September 2008 to November 2015. During this time period, the volume of tweets grew by three orders of magnitude, but the random percentage of all public tweets fell from 50% to 10%. For each word in the labMT dictionary, e.g. "Obama", we subsample the gardenhose for all tweets matching the word. We then tabulate the daily frequencies of labMT words in this term-defined collection of tweets, resulting in temporal counts of the words co-occuring with "Obama". For example, the resulting collection of counts for "Obama" is a 2,628 (days) by 10,222 (words) matrix with entry (i, j) representing the frequency of labMT word j appearing in a tweet containing the term "Obama" on day i. This collection of counts is posted on the online Appendix for this paper.

Fig. 4.1 gives the average daily ambient happiness of "Obama", along with the average daily happiness of all tweets during the same time period. Along with a general slow decline, we see spikes in happiness each year on August 4th, the President's

birthday, with the largest spike occurring on October 9, 2009 when President Obama was awarded the Nobel Peace Prize. We see a strong dip shortly after on October 26, 2009 when President Obama declares a state of emergency for the H1N1 virus. We see spikes in relative frequency of "Obama" on both election days in 2008 and in 2012.

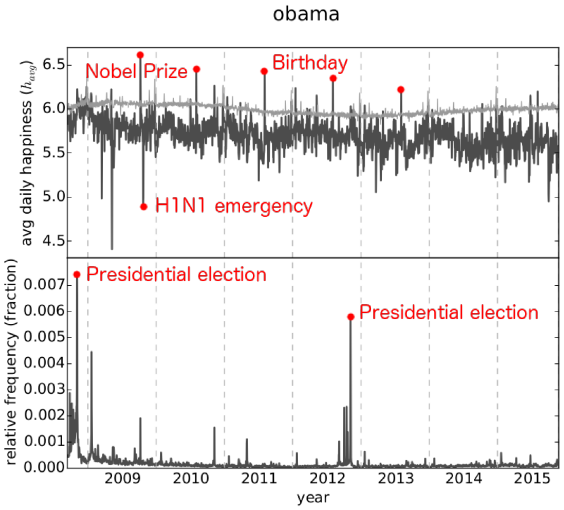

Figure 4.1: Average daily happiness of tweets containing "Obama" (top) with the relative frequency of "Obama" tweets (bottom). Spikes in happiness include President Obama's birthday (August 4th) and his winning of the Nobel Prize (10/09/2009). Dips include a state of emergency for the H1N1 virus. Spikes in relative frequency occur on election days in 2008 and 2012.

To compare our findings with solicited opinions, we collected yearly and quarterly polling data from Gallup [1]. We focus on quarterly data, as it is the highest resolution we were able to obtain. The yearly analysis provides us with only 7 data points, and results are in the Appendix. We compare President Obama's job approval rating on Gallup and on Pollster [3], which allows for daily data collection through their API. Finally, we use the University of Michigan's Index of Consumer Sentiment data, which is collected monthly [4].

4.4 Results

4.4.1 Unsolicited Public Opinions

Here we present happiness time series for several words for which we find interesting patterns. Figure 4.2 gives examples of ambient happiness and relative frequency time series for a few selected words. Happiness associated with certain religious words, e.g. "church" and "muslim" has decreased in recent years, with dips corresponding to several mass shootings including the Charleston shooting in June 2015 and the Chapel Hill shooting in February 2015. The relative frequency of "church" peaks each year on Easter Sunday. We see that ambient happiness of "snow" is seasonal, with the highest happiness during the northern hemisphere summer and lowest during the winter, while the relative frequency is highest during the winter and lowest during the summer. The saddest day for "snow" was June 14, 2015 when a popular Game of Thrones character is presumed dead. The ambient happiness scores of "democrat" and "republican" are on a slow decline, with relative frequencies peaking during presidential and midterm elections. President Obama's press conference after the Sandy Hook shooting is the saddest day for "democrat", while the saddest day for "republican" coincides with an incident involving the Egyptian Republican Guard. Ambient happiness of "love" peaks around the holidays each year, and the relative frequency was increasing until recently. While ambient happiness of "love" peaks on Christmas each year, the relative frequency of "love" peaks on Valentine's Day each year, which could be due to the difference in labMT scores for "christmas" and "valentines" (7.96 and 7.30 respectively).

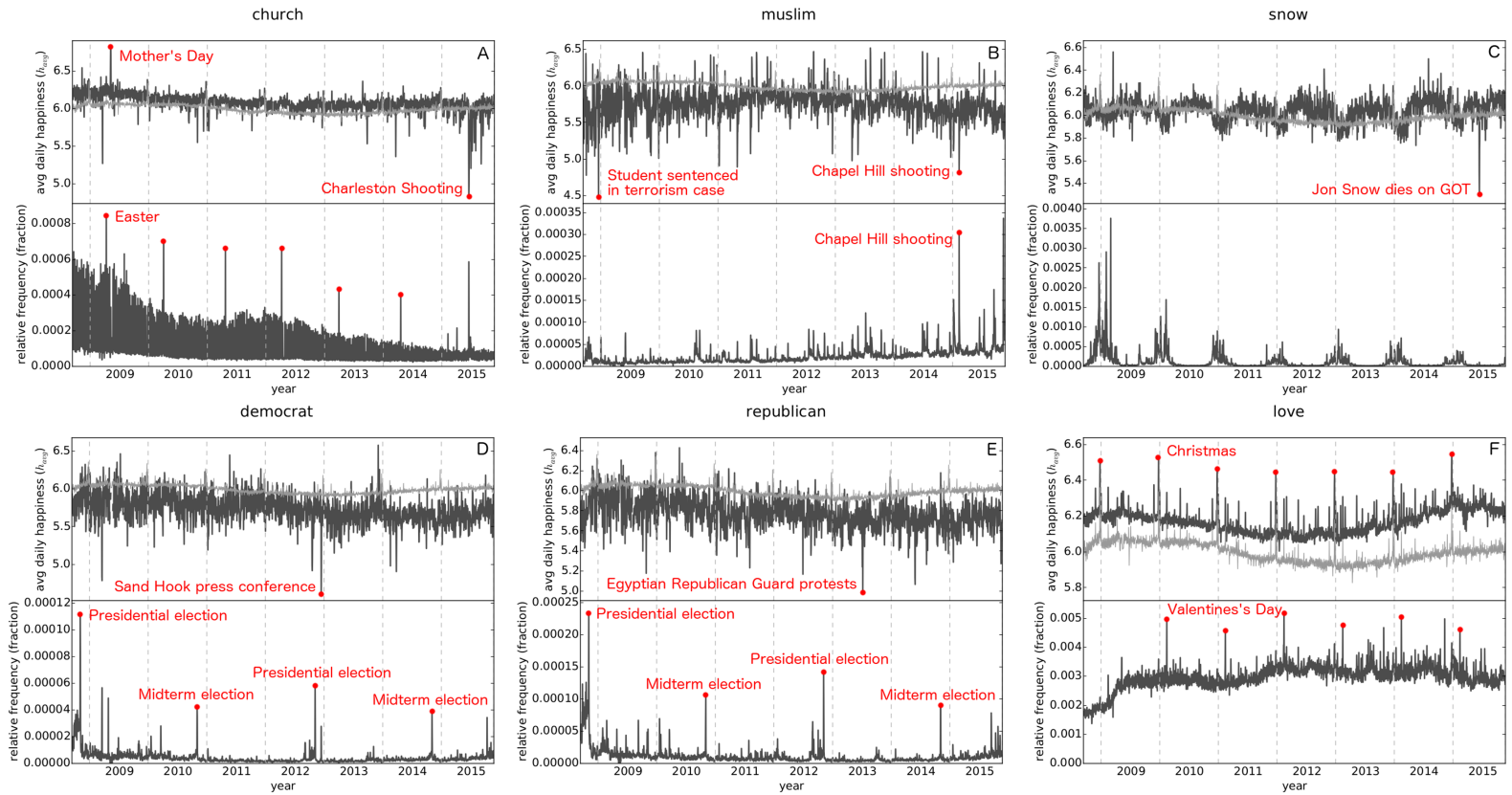

Figure 4.2: Six examples of ambient happiness time series (top, dark gray) along with relative frequency (bottom). Twitter's overall average happiness trend is in light gray for each plot. Relative frequency is approximated by dividing the total frequency of the word by the total frequency of all labMT words on a given day. (A) "church": There is a large spike in happiness on Mother's day and a large dip following the Charleston church shooting in June 2015. There are spikes in relative frequency each Sunday, and yearly on Easter Sunday. (B) "muslim": Two dips correspond to a sentencing in a terrorism case in late 2008, and the shooting at Chapel Hill in February 2015. (C) "snow": Sentiment and relative frequency are seasonal, with a large dip when a main character dies on the HBO show Game of Thrones. (D) "democrat": Overall sentiment gradually decreases with a large dip after president Obama's press conference following the Sandy Hook shooting. There are spikes in relative frequency on election days. (E) "republican": Overall sentiment gradually decreases with a large dip after protests of the Egyptian Republican Guard. (F) "love": Sentiment peaks each year on Christmas while relative frequency peaks each year on Valentine's Day. Weekly and monthly ambient happiness time series for each of these six terms are given in the Appendix (Figs. B.5 and B.6) and time series for nearly 10,000 terms can be found in the online Appendix for the paper.

In traditional polls, there may be large differences in public opinion from one time period to the next. In a yes/no or multiple choice survey question it is impossible to use that data to determine why differences occur. Here we use word shift graphs to determine the cause of a shift in ambient happiness.

A word shift graph ranks words by their contributing factor to the change in happiness between two pieces of text. For example, in Fig. 4.3 we investigate why the ambient happiness of "snow" is higher in the northern hemisphere summer (when its relative frequency is lowest) and lower in the winter (when its relative frequency is highest).

The word shift graph in Fig. 4.3 compares "snow" tweets during the winter months (December, January, February) to "snow" tweets in the summer months (June, July, August). English speaking countries like Australia and New Zealand will necessarily be included in the wrong season, however their contribution is small.

We find that Twitter users loathe the snow during the winter, and miss the snow during the summer, as indicated by the increase in the word "hate", negatives, and profanity during the winter months and the decrease in the word "love". The influence of the Disney classic "Snow White" is also visible, appearing to be referenced more often in summer months due to its motion picture release on June 1, 2012.

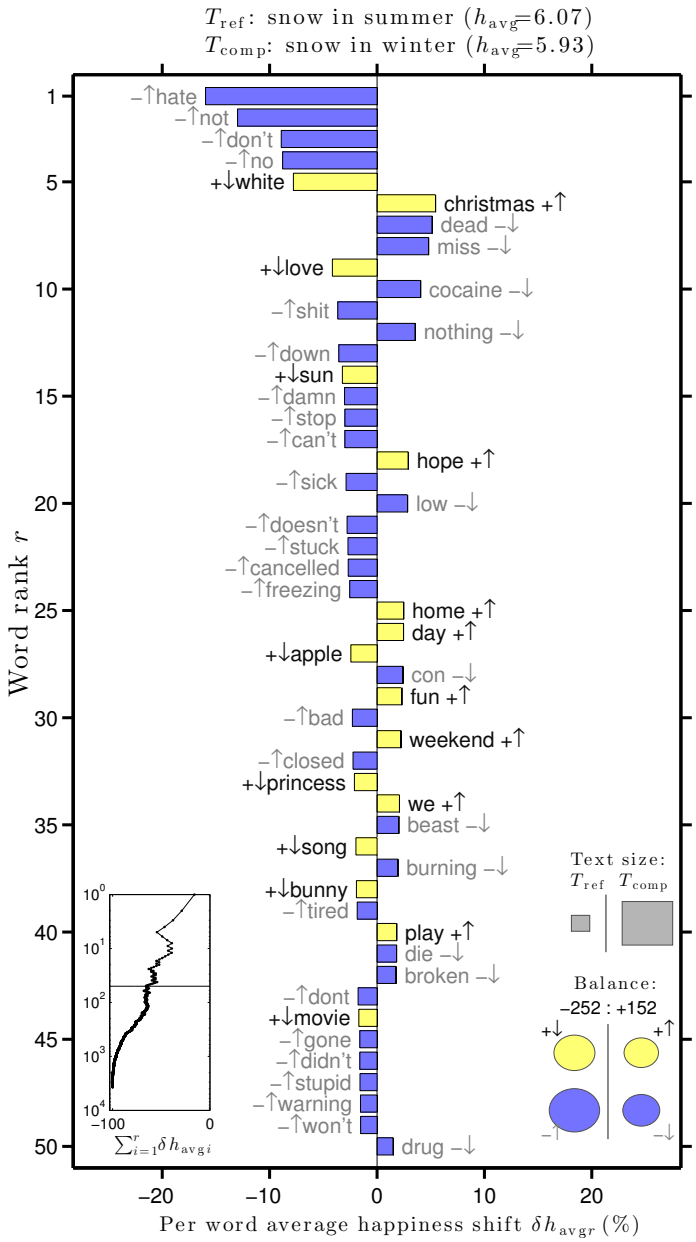

Figure 4.3: A word shift graph comparing tweets that contain the word "snow" during the summer months (reference text) and winter months (comparison text). A purple bar indicates a relatively negative word, a yellow bar indicates a relatively positive word, both with respect to the reference text's average happiness. An up arrow indicates that word was used more in the comparison text. A down arrow indicates that word was used less in the comparison text. Words on the left contribute to a decrease in happiness in the comparison text. Words on the right contribute to an increase in happiness in the comparison text. The circles in the lower right corner indicate how many happy words were used more or less and how many sad words were used more or less in the comparison text.

4.4.2 President Obama's Job Approval Rating

We next investigate the relationship between President Obama's Job Approval Rating from two public opinion polling resources and the ambient happiness of "Obama" tweets. President Obama's quarterly job approval rating is freely available on gallup.com [1], and President Obama's daily job approval rating is freely available on pollster.com [3].

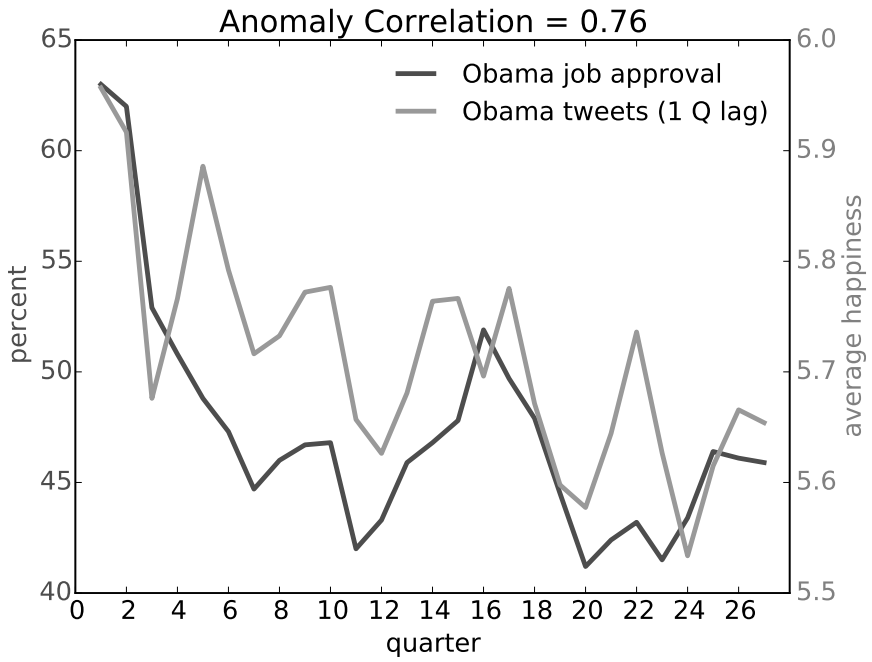

Figure 4.4: Average quarterly happiness of tweets containing "Obama" on a one quarter lag with Obama's quarterly job approval rating. The high positive correlation indicates opinions on Twitter precede timely solicited surveys.

We correlate the average quarterly happiness of tweets containing the word "Obama" with President Obama's quarterly job approval rating and find a strong positive correlation (see Appendix Fig. B.3). However, we find the correlation is much stronger in Fig. 4.4, which gives the happiness time series at a one quarter lag. Similarly, we

find a strong positive correlation between the daily approval rating available on Pollster and the daily ambient happiness of "Obama" (see Appendix Fig. B.4a) with an improvement in the correlation when the tweets are lagged by 30 days in Appendix Fig. B.4b. This indicates that real time Twitter data has the potential to predict solicited public opinion polls.

Figure 4.4 shows that President Obama's highest approval rating in all three sources was during his first quarter (January-March, 2009). His lowest approval rating was during his 23rd quarter (July-September, 2014). Fig. 4.5 shows which words contributed most to this shift in ambient happiness. Tweets containing the word "Obama" discuss war and terrorism more often during his 23rd quarter than his first quarter.

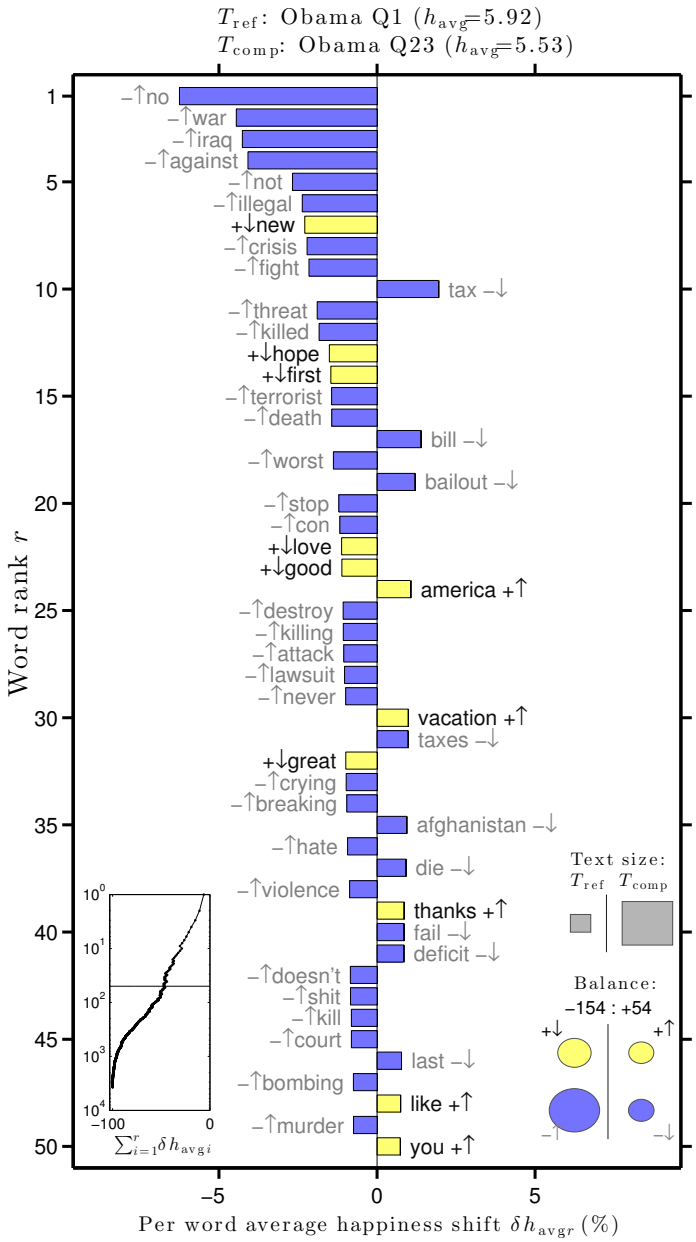

Figure 4.5: A word shift graph comparing tweets that contain the word "Obama" during the first quarter of his presidency, 2009/01–2009/03, (reference text) and 23rd quarter of his presidency, 2014/07–2015/09, (comparison text). Tweets referred to war and terrorism more often in quarter 1.

4.4.3 Index of Consumer Sentiment

Next, we investigate a monthly poll on consumer sentiment designed by the University of Michigan [4]. This poll asks participants five questions about their current and future financial well being and calculates an Index of Consumer Sentiment (ICS) based on responses. In Fig. 4.6 we correlate this monthly time series with the ambient happiness of the word "job". We find that the correlation is much stronger starting in 2011 (Fig. 4.6b), and even stronger still when the ambient happiness is lagged one month (Fig. 4.6c). In Fig. 4.6d we correlate the ICS with the relative frequency of the word "job" on Twitter. We find a strong negative correlation, indicating that it is more likely that a user will tweet the word "job" when they are searching for one.

4.4.4 Business Sentiment Shifts

In this section, we investigate the changes in Twitter sentiment surrounding two businesses, Walmart and McDonalds. We examine the ambient happiness time series to determine how sentiment changes in response to events that took place at specific stores. Fig. 4.7 gives the ambient happiness and relative frequency of the words "walmart" and "mcdonalds".

Many of the spikes in the "walmart" ambient happiness time series correspond to free giveaways to which Twitter users are responding. A dip in November 2008 corresponds to the trampling to death of a Walmart employee on Black Friday (the day after Thanksgiving, notorious in the U.S. for shopping). Shootings that took place in Walmart stores in 2014 are shown with orange dots in Fig. 4.7a. In June 2014 the Jerad and Amanda Miller Las Vegas shootings ended with 5 deaths (including

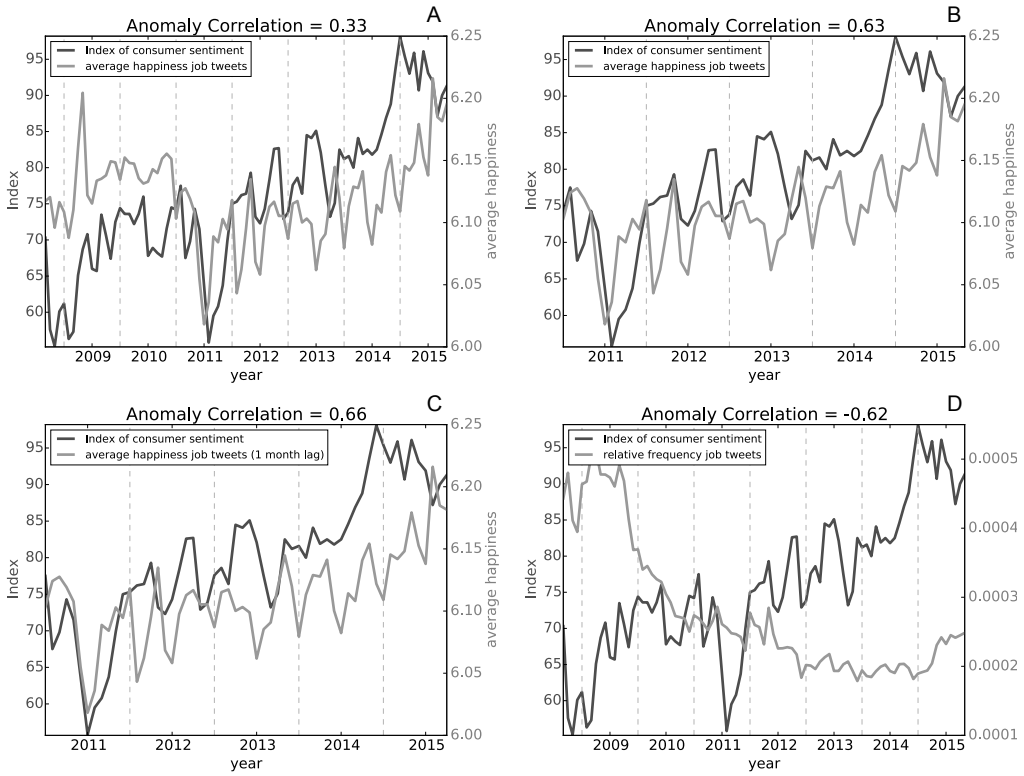

Figure 4.6: (A) Ambient happiness of "job" with the Index of Consumer Sentiment. We see a small positive correlation getting stronger after 2011. (B) Ambient happiness of "job" with ICS starting in 2011. (C) Ambient happiness of "job" is lagged by one month. (D) ICS with relative frequency of "job".

themselves) in a Nevada Walmart. In September 2014, the police officer who shot John Crawford in an Ohio Walmart was indicted. In December 2014, a 2 year old accidentally shot and killed his mother in an Idaho Walmart. We also see a dip in happiness on the day Tracy Morgan sues Walmart over a fatal crash with one of their tractor trailers in July 2014.

The happiest day in the "mcdonalds" ambient happiness time series is Valentine's Day in 2015. Upon reading some example tweets from this day, we find that McDonalds was a popular ironic destination for Valentine's Day dinner that year among Twitter users. A second spike corresponds to a prestigious award given to the Mc-

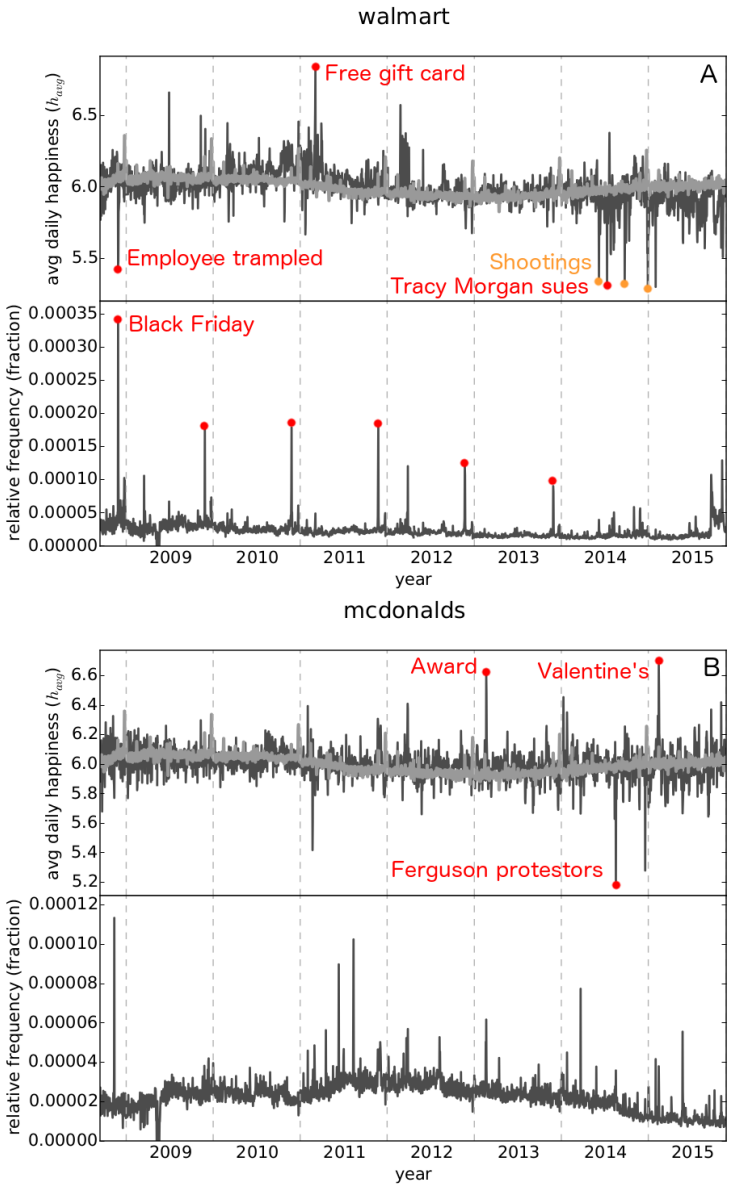

Figure 4.7: The ambient happiness and relative frequency time series for (A) "walmart" and (B) "mcdonalds". Dips in sentiment correspond to deaths, lawsuits, and protests, while spikes in happiness correspond to awards, giveaways, and holidays. Spikes in the relative frequency of "walmart" appear largely on Black Friday. Time series for nearly 10,000 other terms can be found on the online Appendix for the paper.

Donalds enterprise in February 2013. McDonalds was given the "Top Toilet Award" for the cleanliness of its restrooms. The saddest day for McDonalds on Twitter was August 18, 2014, the day that Ferguson protesters broke into a McDonalds to steal milk to relieve tear gas victims.

In Fig. 4.8 we explore the monthly ambient happiness of "walmart" and "mcdonalds". We find that the ambient happiness of "walmart" reaches its maximum in March 2011, and its minimum in October 2015, and the ambient happiness of "mcdonalds" reaches its maximum in February 2015 and its minimum shortly after in May 2015. To investigate the texture behind these observations, we use word shift graphs to compare the happiest and saddest months for each business in Fig. 4.9.

Figure 4.8: Monthly ambient happiness of (A) "walmart" and (B) "mcdonalds".

In November 2015 (comparison text Fig. 4.9a), there were Black Friday altercations at many Walmarts throughout the country, often caught on camera, leading to

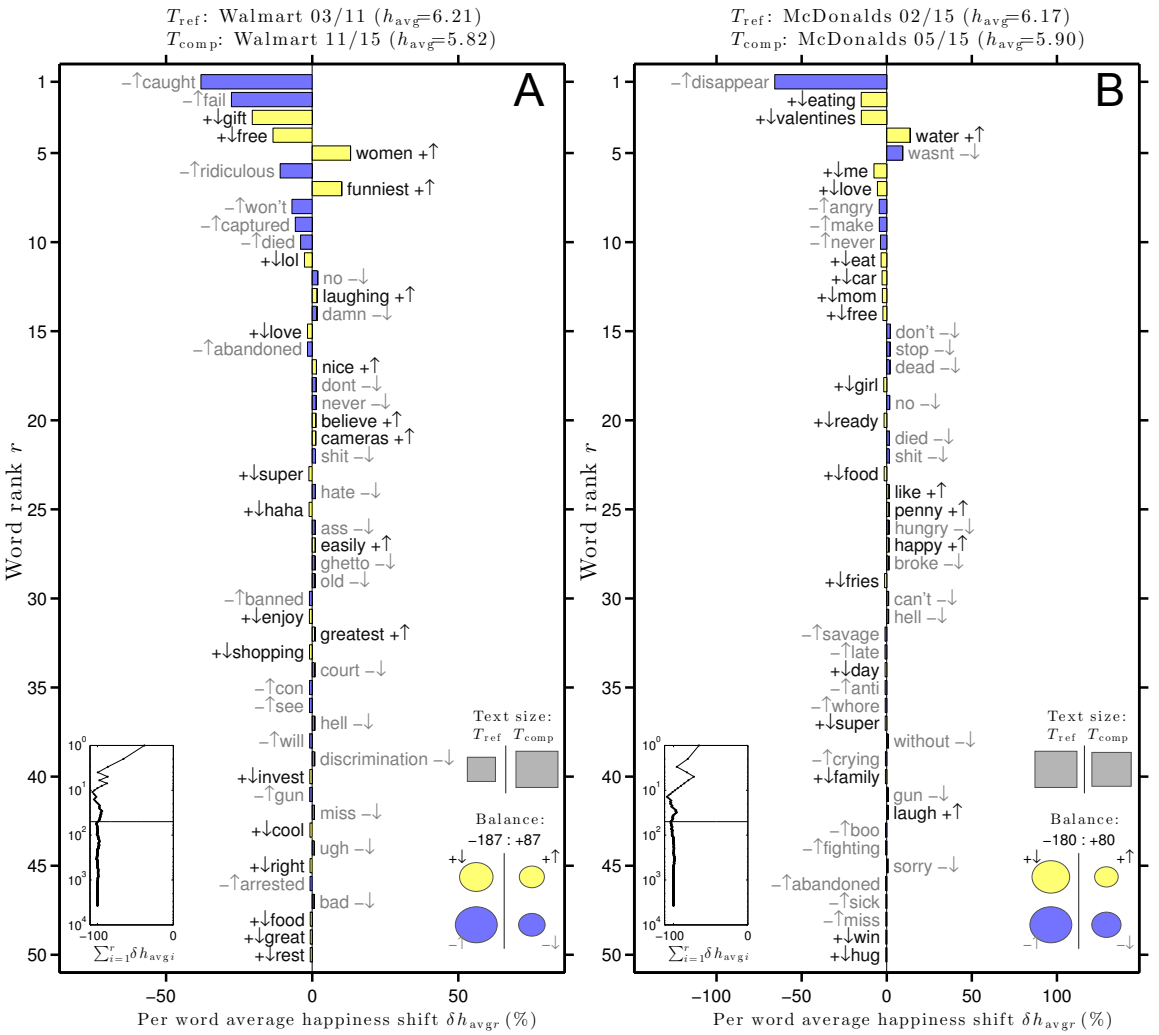

Figure 4.9: Word shift graphs comparing the happiest and saddest months for (A) "walmart" and (B) "mcdonalds". The happiest month represents the reference text and the saddest month represents the comparison text.

an increase in negative words such as "caught", "fail", "ridiculous", and "captured". Twitter users were happier about Walmart in March 2011 (reference text Fig. 4.9a) due in part to a free gift card giveaway. Happier tweets included the words "lol", "love", "haha", and "super". Surprisingly, we actually see more curse words during the happiest month than the saddest month.

The happiest month for McDonalds was February 2015 (reference text Fig. 4.9b) when a surprising number of Twitter users were spending Valentine's Day there, hence the decrease in the words "valentines" and "love". The decrease in happiness in May 2015 is in large part in an increase in the word "disappear". During this time, a video of a Michigan McDonalds employee performing a practical joke, in which he claims he's going to make a penny disappear in a bottle of water, went viral. Using word shifts, we are thus able to determine "disappear" was not a true indicator for negativity. The next step would be to add "disappear" to out stop word list and reevaluate the time series. In general, word shifts are very centavo diagnostics that allow us to make sense of apparent sentiment patterns, and to adjust the hedonometer as needed.

4.5 Limitations

Here we present positive correlations between ambient happiness on Twitter and solicited public opinion surveys. The methods used in this work present several limitations that we address here. First, we are unable to compare ambient happiness to public opinion surveys for the majority of subjects discussed on Twitter. Ambient happiness surrounding commercial businesses can be compared to stock prices, however this analysis provides very poor and meaningless correlations. Stock prices take into account much more than the success of one business, and thus would not be an accurate representation of public opinion.

We acknowledge that daily analysis can uncover very small events that are retweeted many times and cause a spike or a dip in ambient happiness. Twitter contains both

noise and insignificant conversation and reactions to major events. Depending on the needs of the researcher, one might not care about the video at McDonalds that went viral in May 2015. This noise also makes it difficult to compare high resolution ambient happiness to low resolution opinion surveys. Fig. B.7 in the Appendix shows our attempt to compare ambient happiness to six topics available on Gallup. Gallup surveys take place over several days, once a year, while ambient happiness is calculated by averaging all tweets for a full 365 days. We are therefore unable to report which method is a more accurate indicator of public opinion. Instead, we conclude that public opinion polling with Twitter has the potential to complement traditional public opinion surveys.

Finally, without knowing user information from each tweet, we cannot know if we are using an unbiased sample of the human population. In this work, however, we claim our conclusions pertain only to the Twitter universe, and not the human race as a whole. Twitter users are not even always human. Previous work shows that Twitter contains many bots, which send tweets automatically, often to advertise a product or a political campaign [23]. We do not eliminate these tweets in this work, however many methods for uncovering them have been suggested [18, 10, 14, 12].

4.6 Conclusion

The objective of this research was to determine the extent to which ambient happiness on Twitter can be used as a reliable public opinion polling device. A central motivation is that solicited public opinion polling data is difficult to obtain at a high temporal resolution, if at all [29].

With data from Twitter we can investigate topics other than political or global issues, which are the focus of a large majority of solicited surveys. We can use ambient happiness to determine how people feel about seemly neutral topics like "snow", or how they are using the words "love", and "feel". We also have shown that Twitter users respond to various kinds of events taking place at commercial businesses, and thus ambient happiness could be used in market analysis to predict or improve a company's sales.

Of the available polling data we were able to obtain, we find that ambient happiness of selected words correlates well with solicited public opinions. Often times, the correlation increases when the tweets are lagged, indicating that real time Twitter data has the potential to predict solicited public opinion polls.

Not only can tweets precede survey responses, but we can use individual words within tweets to determine why one time period is happier than another, something that is not possible in traditional polls due to the multiple choice aspect of most surveys. Several other advantages of using tweets for public opinion polling include the ability to track movement [19] and make maps [30, 25] using geolocated tweets. Data collection itself is largely algorithmic, and does not rely on the responses of participants.

We find that for many topics, Twitter is a valuable resource for mining public opinions without solicited surveys. We encourage readers to explore the data online here. Social media may be the future of public opinion polling, revealing important signals complementary to traditional surveys.

Bibliography

[1] Gallup Trends. http://www.gallup.com/poll/trends.aspx. Accessed: 2016-03-08.

[2] Google Correlate. https://www.google.com/trends/correlate,.

[3] Pollster API. http://elections.huffingtonpost.com/pollster/api. Accessed: 2016-03-08.

[4] University of Michigan index of consumer sentiment. http://www.sca.isr.umich.edu/tables.html. Accessed: 2016-03-08.

[5] Dolan Antenucci, Michael R. Andwerson, Penghua Zhao, and Michael Cafaerlla. A query system for social media signals. 2015.

[6] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[7] Pablo Barberá. Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data.

[8] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[9] Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, 2012.

[10] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9(6):811–824, 2012.

[11] Robert B Cialdini and Nathalie Garde. *Influence*, volume 3. A. Michel, 1987.

[12] Eric M Clark, Jake Ryland Williams, Chris A Jones, Richard A Galbraith, Christopher M Danforth, and Peter Sheridan Dodds. Sifting robotic from organic text: A natural language approach for detecting automation on Twitter. *Journal of Computational Science*, 2015.

[13] Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLoS ONE*, 10(8), 2015.

[14] John P Dickerson, Vadim Kagan, and VS Subrahmanian. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 620–627. IEEE, 2014.

[15] Peter U Diehl, Bruno U Pedroni, Andrew Cassidy, Paul Merolla, Emre Neftci, and Guido Zarrella. TrueHappiness: Neuromorphic emotion recognition on TrueNorth. *arXiv preprint arXiv:1601.04183*, 2016.

[16] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449, 2013.

[17] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752, 2011.

[18] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *arXiv preprint arXiv:1407.5225*, 2014.

[19] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, 3, 2013.

[20] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, page 0894439313493979, 2013.

[21] Maria Glenski and Tim Weninger. Rating effects on social news posts and comments. *arXiv preprint arXiv:1606.06140*, 2016.

[22] Brian Helmuth, Tarik C Gouhier, Steven Scyphers, and Jennifer Mocarski. Trust, tribalism and tweets: Has political polarization made science a "wedge issue"? *Climate Change Responses*, 3(1):1, 2016.

[23] Philip N Howard and Bence Kollanyi. Bots,# Strongerin, and# Brexit: Computational propaganda during the UK-EU referendum. *Available at SSRN 2798311*, 2016.

[24] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PLoS ONE*, 7(1):e29484, 2012.

[25] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3):e1500779, 2016.

[26] Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PloS one*, 9(5):e94093, 2014.

[27] Yu-Ru Lin, Drew Margolin, Brian Keegan, and David Lazer. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*, pages 737–748. International World Wide Web Conferences Steering Committee, 2013.

[28] Yelena Mejova, Ingmar Weber, and Michael W Macy. *Twitter: A digital socioscope*. Cambridge University Press, 2015.

[29] Greg Miller. Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815, 2011.

[30] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[31] Anshul Mittal and Arpit Goel. Stock prediction using Twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 2012.

[32] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[33] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[34] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *ICWSM*, pages 265–272, 2011.

[35] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, 3, 2013.

[36] Andrew Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs, 2015.

[37] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, pages 9–17. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), 2009.

[38] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[39] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[40] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[41] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T Jost, Jonathan Nagler, and Joshua Tucker. Social media and political communication: a survey of Twitter users during the 2013 Italian general election. *Rivista italiana di scienza politica*, 43(3):381–410, 2013.

[42] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

Chapter 5

Conclusion

In this work, I have explored the benefits of utilizing large scale text data from traditional and social media sources to draw conclusions about human opinions and behaviors. Compared to traditional multiple choice surveys, text-based data can be easier and cheaper to obtain, and can provide researchers with information for further analysis using techniques from machine learning, computational linguistics, and information retrieval. Data of this variety allows us greater spatial and temporal visibility into human behavior, and additional instrumentation on the dashboard of society.

The majority of this research focuses on public opinions on climate change. Using LSA and LDA on two corpora consisting of newspaper articles, I discover that news media reporting following Hurricane Sandy focused partly on the consequences of climate change. Not only was climate change awareness higher following this natural disaster, but post-event reporting also highlighted the link between climate change, hurricanes, and energy system vulnerability. This shift in climate change awareness may be due to an increase in climate change related research, but may also be partly

the effect of several confounding factors including the timing and the location of Hurricane Sandy. Hurricane Sandy hit New York City which is a more heavily populated and higher educated area than the location of Hurricane Katrina. More educated people are more likely to link severe hurricanes to climate change. Hurricanes are also far less common in the northern United States than in the south, and they are far less common in October than during the summer months. These two factors may have highlighted the risks of climate change more so than a southern, summer hurricane.

Using the hedonometer, I explore the value of climate change discussions on social media. Through happiness time series and word shift plots, I discover that climate change news is abundant on Twitter, and public opinion shifts in response to climate change news and events. Natural disasters cause the ambient happiness of "climate" to fall, while climate change rallies cause the ambient happiness of "climate" to rise. I also discover that hurricanes are discussed relatively more frequency within tweets also mentioning climate change, than those that do not.

More generally, I demonstrate that Twitter is a valuable data source for public opinion on many subjects, from significant global and national debates to simple feelings and ideas. Ambient happiness of specific words often correlates well with traditional public opinion surveys on political issues and general well being. I also find that ambient happiness on Twitter often precedes traditional survey data, indicating Twitter's potential predictive power.

The work presented in this dissertation contributes to several fields including machine learning, computational linguistics, environmental sciences, and environmental communication. We illustrate that machine learning techniques LSA and LDA are

valuable in analyzing public opinions and discourse within newspaper articles. These techniques have not yet been utilized to explore climate change awareness. I also demonstrate the power of the hedonometer to detect changes in public opinion of climate change, politics, commercial businesses, and human behaviors. I provide an online appendix to Chapter 4 which gives 30,000 happiness time series along with the data utilized in this chapter so this research can be expanded by other interested researchers. In the past, environmental communications studies have used manual coders to extract topics from collections of text. This work is one of the first to utilize mathematical modeling techniques to draw similar conclusions without having to read an entire corpus of articles.

There are a multitude of directions that this research could go in the future. There are several limitations to my current approaches which can be improved through future research. Determining the number of topics to use in a topic model is a difficult and heavily researched subject. It is often a subjective choice, and results can vary depending on the selection of this variable. The Hierarchical Dirichlet Process (HDP) is another approach to clustering words to create topics within a corpus, and has the unique benefit of learning the number of topics itself. HDP can be used to compare human selected topic numbers to computer selected topic numbers.

One approach I did not take within this work was to separate the articles by newspaper, and analyze any differences in prominent topics based on location of the newspaper. Perhaps the climate change topic present in the media following Hurricane Sandy is dominated by a single source. Any of the three topic modeling techniques could be used to accomplish this. Similarly, one could compare climate change opinions by location by using geo-tagged tweets.

Due to the long period of time between Hurricanes Katrina and Sandy, I am unable to conclude that Hurricane Sandy is the reason for the shift in public discourse. I do conclude that this shift happened sometime between the two disasters. To determine when and why this shift took place, a deeper analysis using more hurricanes and potentially other types of natural disasters over time may help to strengthen the argument of this chapter. With more data, we can pinpoint when the shift in public discourse took place.

While I do find some survey data with which to compare ambient happiness results, it would be very beneficial to obtain solicited survey data at the daily resolution so that many more topics on Twitter could be analyzed. This can be accomplished with a subscription to gallup.com or Amazon's Mechanical Turk. Future researchers could conduct their own surveys and compare the results to self-reported public opinions on Twitter.

Finally, this dissertation has kept analysis data sources from different mediums separate. In the future, I am eager to compare opinions within traditional media, social media, and scientific journals by implementing techniques similar to those used in this work.

The overall significance of my work is three fold. First, I determined that climate change opinions can be moulded by observable consequences, news, and events. Specifically, that Hurricane Sandy highlighted both climate change risks and energy system vulnerability. Second, I demonstrated that there is a plethora of valuable data available to researchers for analysis of public opinions. We can learn so much about current events and global debates by researching how the general public is discussing it within different portals. Finally, I illustrate that human computer interaction is a

very powerful and time saving tool when analyzing massive amounts of text. Rather than reading thousands of time consuming articles, we can use computer generated algorithms to draw similar conclusions. If we combine new tools from Computational Social Science with traditional methods of public opinion polling, we can now quantify human behaviors, opinions, and actions at an unprecedented scale..

Appendix A

Supplementary Materials for Chapter 2

Hurricane Katrina LSA												
	"climate"	Similarity	"energy"	Similarity	"climate, energy"	Similarity		"climate"	Similarity	"energy"	Similarity	"climate, energy"	Similarity
1	climate	1.000	energy	1.000	energy	0.979	51	supposedly	0.746	conocophillips	0.929	retailers	0.908
2	larger	0.866	prices	0.986	prices	0.952	52	boogie	0.746	jumped	0.928	citroen	0.907
3	destroy	0.861	exchange	0.968	deutsche	0.945	53	theories	0.746	citroen	0.927	behravesh	0.907
4	formally	0.848	consumers	0.966	price	0.943	54	nurtured	0.745	tumbling	0.926	traders	0.906
5	theory	0.844	weinberg	0.966	underinvestment	0.943	55	raw	0.745	mercantile	0.925	producers	0.905
6	sound	0.837	argus	0.964	signaling	0.941	56	topics	0.744	production	0.924	idled	0.905
7	gale	0.826	reidy	0.962	discounting	0.940	57	sounded	0.743	embargo	0.923	products	0.905
8	reinforced	0.817	splurge	0.960	java	0.940	58	cynthia	0.742	putins	0.922	tenth	0.904
9	journal	0.815	hummer	0.960	argus	0.939	59	deadly	0.742	shutdowns	0.920	export	0.904
10	sensitive	0.814	markets	0.959	hummer	0.938	60	sacrifice	0.741	reserve	0.920	commodity	0.904
11	unlikely	0.812	downers	0.958	oil	0.937	61	certain	0.740	crude	0.920	imports	0.903
12	belief	0.809	highs	0.958	consumers	0.937	62	cataclysmic	0.740	arabica	0.920	adjusting	0.903
13	phenomenon	0.809	underinvestment	0.957	shocks	0.934	63	nor	0.740	pretax	0.919	yergin	0.902
14	rail	0.800	exporting	0.954	weinberg	0.934	64	reconstructed	0.739	mobil	0.919	artificially	0.902
15	studying	0.796	price	0.954	markets	0.934	65	assessments	0.739	soaring	0.919	nariman	0.902
16	wealthy	0.795	reserves	0.954	profits	0.931	66	haunting	0.738	uncharted	0.919	cents	0.902
17	brings	0.792	signaling	0.953	reserves	0.931	67	continuing	0.737	imports	0.919	tightness	0.902
18	barge	0.792	dampening	0.950	exchange	0.931	68	transforming	0.737	chevrons	0.919	subjective	0.902
19	ancient	0.791	oil	0.950	peaks	0.931	69	william	0.737	exxon	0.917	doha	0.901
20	masters	0.786	java	0.949	highs	0.929	70	regard	0.736	manifold	0.917	spikes	0.900
21	politicians	0.785	cents	0.948	splurge	0.927	71	vicinity	0.736	trading	0.916	winter	0.898
22	professor	0.783	deutsche	0.948	exporting	0.927	72	booming	0.735	suisse	0.916	exxon	0.897
23	recommendations	0.782	gasoline	0.947	gasoline	0.923	73	audiences	0.735	automaker	0.916	uncharted	0.897
24	thick	0.782	traders	0.946	dampening	0.923	74	advocacy	0.734	tepid	0.915	chairmans	0.897
25	marked	0.780	nariman	0.946	pinch	0.922	75	mass	0.733	futures	0.915	soared	0.897
26	alter	0.779	discounting	0.945	oils	0.922	76	remarkable	0.733	geopolitical	0.915	conocophillips	0.896
27	sounds	0.776	behravesh	0.944	soaring	0.922	77	breaking	0.732	record	0.914	clamping	0.895
28	hole	0.776	retailers	0.943	exported	0.920	78	facts	0.732	yergin	0.914	exporters	0.895
29	peril	0.775	barrel	0.942	reidy	0.919	79	constituents	0.731	clamping	0.914	bps	0.895
30	extremely	0.771	heating	0.942	output	0.919	80	isolated	0.730	retail	0.914	crimp	0.895
31	avoided	0.770	oils	0.942	exporter	0.917	81	vibrant	0.703	hess	0.913	cutback	0.894
32	loose	0.770	shocks	0.941	easing	0.917	82	unequivocal	0.730	pinch	0.912	global	0.894
33	multi	0.769	idled	0.941	putins	0.917	83	recommended	0.729	chairmans	0.911	pretax	0.893
34	appear	0.767	jolted	0.941	record	0.916	84	unprotected	0.728	closings	0.911	disrupted	0.893
35	devastating	0.766	output	0.940	tumbling	0.916	85	inundated	0.727	depository	0.910	liquefied	0.892
36	draft	0.764	peaks	0.937	demand	0.915	86	ears	0.726	disrupted	0.909	premcor	0.892
37	possibility	0.764	profits	0.936	downers	0.915	87	exuberant	0.725	winter	0.909	jumped	0.891
38	roiled	0.759	soared	0.936	automaker	0.913	88	greenhouse	0.725	sunoco	0.909	mobil	0.891
39	retracted	0.758	exported	0.936	heating	0.913	89	powers	0.725	chevron	0.908	arabica	0.890
40	mismanagement	0.758	premcor	0.935	disruptions	0.913	90	alarms	0.724	doha	0.908	bros	0.890
41	plot	0.757	disruptions	0.934	atm	0.911	91	comment	0.723	bros	0.907	mercantile	0.890
42	produced	0.757	exporter	0.934	tepid	0.911	92	brokers	0.722	commodity	0.907	analyst	0.887
43	becomes	0.755	easing	0.933	chevrons	0.911	93	deny	0.722	commodities	0.906	gas	0.887
44	decades	0.753	crimp	0.932	jolted	0.911	94	pianos	0.722	wholesalers	0.905	geopolitical	0.886
45	consider	0.752	dent	0.932	embargo	0.909	95	baker	0.721	refiner	0.905	interruptions	0.886
46	wealthier	0.752	demand	0.932	pricing	0.909	96	ethnic	0.720	soar	0.903	squeeze	0.886
47	dismissed	0.751	roasters	0.930	roasters	0.908	97	cyclical	0.720	analyst	0.903	chevron	0.885
48	repeated	0.751	tightness	0.930	dent	0.908	98	relieve	0.720	products	0.903	crude	0.885
49	delays	0.750	atm	0.929	production	0.908	99	studies	0.720	bps	0.903	nations	0.884
50	unique	0.749	pricing	0.929	barrel	0.908	100	spread	0.720	thurtell	0.902	derivatives	0.884

Table A.1: Results of LSA for Hurricane Katrina for 3 different queries. Words are ordered based on their cosine distance from the query vector. Includes the 100 words most similar to the query.

Hurricane Sandy LSA													
	"climate"	Similarity	"energy"	Similarity	"climate, energy"	Similarity		"climate"	Similarity	"energy"	Similarity	"climate, energy"	Similarity
1	climate	1.000	energy	1.000	climate	0.979	51	fuels	0.895	extracted	0.862	deniers	0.906
2	change	0.963	technologies	0.949	warmer	0.961	52	kerry	0.894	abundance	0.860	vigil	0.904
3	reduce	0.957	fuels	0.946	georgetown	0.956	53	hydroelectric	0.893	tackle	0.860	proportion	0.904
4	warming	0.957	fossil	0.943	warming	0.955	54	pollute	0.893	regulating	0.858	targets	0.902
5	reducing	0.956	hydroelectric	0.936	reduce	0.955	55	technologies	0.891	outweigh	0.858	mover	0.901
6	pressures	0.952	renewable	0.932	energy	0.952	56	altering	0.890	envisioned	0.857	scientists	0.899
7	georgetown	0.947	rogue	0.932	reducing	0.951	57	regulating	0.890	miserably	0.856	automobiles	0.899
8	lowering	0.943	employing	0.921	pressures	0.948	58	mover	0.889	subtler	0.856	devise	0.898
9	talks	0.942	warmer	0.920	fossil	0.947	59	believing	0.886	upending	0.855	controlling	0.898
10	devise	0.938	supplying	0.918	fuels	0.946	60	enhancement	0.885	pollution	0.855	modification	0.896
11	expands	0.938	firing	0.913	change	0.946	61	planets	0.885	solar	0.855	trillions	0.895
12	outweigh	0.937	efficiency	0.911	technologies	0.945	62	eco	0.883	modification	0.855	scenarios	0.893
13	warmer	0.937	streamlined	0.911	coal	0.943	63	cities	0.882	sciences	0.854	earths	0.893
14	plants	0.934	generating	0.908	global	0.942	64	automobiles	0.882	automobiles	0.853	abundance	0.891
15	drought	0.933	altering	0.906	hydroelectric	0.941	65	greenhouse	0.880	regulatory	0.852	attribute	0.890
16	manipulation	0.929	coal	0.906	emissions	0.940	66	notoriously	0.879	trapping	0.851	greenhouse	0.888
17	emissions	0.929	consumption	0.900	firing	0.937	67	strict	0.878	surprises	0.850	enhancement	0.887
18	global	0.929	adapt	0.898	outweigh	0.936	68	porous	0.878	earths	0.849	doom	0.886
19	imperative	0.927	sparked	0.895	generating	0.933	69	groundwater	0.878	measured	0.848	funneling	0.885
20	arizona	0.924	dimming	0.894	carbon	0.930	70	consumption	0.877	mover	0.847	groundwater	0.885
21	attribute	0.923	georgetown	0.892	arizona	0.930	71	modification	0.876	waterkeeper	0.846	hotter	0.884
22	scientists	0.923	carbon	0.889	editorials	0.929	72	hotter	0.876	change	0.845	copenhagen	0.884
23	planet	0.920	masonry	0.888	plants	0.927	73	earths	0.875	deniers	0.843	oceans	0.883
24	pollution	0.919	global	0.886	humanitys	0.926	74	markedly	0.875	sub	0.843	windstorms	0.883
25	curbing	0.918	erratic	0.885	altering	0.926	75	retaining	0.875	blackouts	0.843	planets	0.881
26	coal	0.917	searchable	0.884	manipulation	0.924	76	attests	0.875	manipulation	0.842	emission	0.881
27	editorials	0.915	faster	0.882	pollution	0.923	77	dimming	0.875	depleting	0.841	munich	0.881
28	targets	0.914	emissions	0.881	employing	0.923	78	employing	0.874	funneling	0.841	rogue	0.880
29	oceans	0.912	skeptics	0.880	drought	0.922	79	proportion	0.873	curbing	0.841	markedly	0.878
30	vigil	0.912	proportion	0.877	extracted	0.921	80	efficiency	0.873	plants	0.841	pollute	0.878
31	scenarios	0.911	trillions	0.876	foretaste	0.920	81	depleted	0.873	sources	0.841	ozone	0.878
32	extracted	0.911	foretaste	0.876	skeptics	0.919	82	exemplified	0.872	frequent	0.841	depleting	0.877
33	humanitys	0.911	warming	0.875	lowering	0.919	83	murky	0.872	oil	0.841	epa	0.877
34	distraction	0.910	reduce	0.875	dioxide	0.918	84	sparked	0.870	planet	0.839	overheated	0.877
35	pentagon	0.910	editorials	0.875	efficiency	0.918	85	essay	0.870	emission	0.838	contiguous	0.877
36	contiguous	0.909	humanitys	0.875	planet	0.917	86	atmospheric	0.869	ozone	0.837	frequent	0.876
37	controlling	0.908	eco	0.875	curbing	0.917	87	overheated	0.869	pentagon	0.836	sensible	0.874
38	carbon	0.907	ton	0.874	consumption	0.915	88	copenhagen	0.869	windstorms	0.836	freely	0.872
39	dioxide	0.906	efficient	0.872	expands	0.914	89	fahrenheit	0.868	acceptance	0.836	kerry	0.871
40	extremes	0.905	cities	0.872	subtler	0.913	90	energy	0.868	buildup	0.835	ton	0.870
41	munich	0.903	doom	0.870	dimming	0.912	91	adapt	0.868	copenhagen	0.835	fahrenheit	0.870
42	firing	0.902	compounding	0.869	talks	0.911	92	windstorms	0.867	focuses	0.835	exemplified	0.869
43	subtler	0.902	mentioning	0.868	sparked	0.910	93	funneling	0.867	vein	0.834	persistence	0.869
44	foretaste	0.900	climate	0.868	pentagon	0.909	94	illustrative	0.866	epa	0.834	atmospheric	0.869
45	generating	0.899	reducing	0.867	eco	0.909	95	vapor	0.866	drought	0.833	environmental	0.866
46	environmental	0.899	pressures	0.866	adapt	0.909	96	abundance	0.864	harvard	0.832	increasing	0.865
47	fossil	0.899	arizona	0.864	imperative	0.908	97	prosperity	0.864	redundant	0.832	levi	0.865
48	deniers	0.898	candlelit	0.863	trapping	0.908	98	freely	0.863	greenhouse	0.829	meaningfully	0.864
49	trapping	0.897	dioxide	0.862	cities	0.907	99	emission	0.862	temperature	0.827	porous	0.863
50	skeptics	0.896	degrees	0.862	regulating	0.906	100	scientific	0.862	iron	0.826	essay	0.862

Table A.2: Results of LSA for Hurricane Sandy for 3 different queries. Words are ordered based on their cosine distance from the query vector. Includes the 100 words most similar to the query.

Hurricane Katrina LDA									
topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
quinn	leve	job	hous	billion	polic	bush	bodi	price	school
team	corp	hous	water	tax	casino	presid	death	oil	student
season	engin	krt	home	feder	offic	democrat	offici	percent	univers
time	flood	st	street	hous	peopl	republican	state	energi	tulan
player	water	home	time	senat	street	hous	home	gas	colleg
play	canal	antoin	day	congress	day	polit	die	gasolin	educ
game	protect	back	peopl	cut	depart	white	victim	rate	back
coach	wall	school	back	republican	fire	administr	peopl	market	campus
start	system	restaur	tree	spend	citi	senat	famili	week	return
point	louisiana	peopl	live	bill	biloxi	respons	parish	product	high
open	armi	work	boat	budget	hotel	govern	st	month	district
make	offici	month	resid	govern	crime	nation	louisiana	consum	enrol
made	surg	worker	work	money	store	american	identifi	report	public
day	feet	louisiana	build	program	reddick	time	morgu	economi	class
sign	project	day	neighborhood	state	water	leader	relat	compani	warm
top	level	live	damag	propos	time	critic	coron	increas	research
week	pump	chitrib	roof	cost	back	peopl	dr	gulf	time
score	lake	rebuild	photograph	bush	gambl	iraq	dead	fuel	hurrican
world	design	end	flood	plan	mississippi	parti	found	expect	teacher
lead	environment	return	photo	million	hous	effort	remain	gallon	institut
topic 10	topic 11	topic 12	topic 13	topic 14	topic 15	topic 16	topic 17	topic 18	topic 19
peopl	leve	red	famili	gras	game	music	town	peopl	ship
black	hous	cross	home	mardi	team	jazz	plan	time	airlin
king	flood	donat	children	french	saint	band	build	american	show
time	protect	relief	day	restaur	play	musician	develop	disast	news
west	rebuild	organ	live	parad	season	art	school	news	time
mayor	home	volunt	back	street	home	cultur	hous	report	northrop
day	system	victim	school	back	footbal	museum	state	world	network
presid	feder	fund	mother	peopl	player	perform	design	stori	travel
polit	work	peopl	friend	quarter	coach	play	communiti	book	air
bloomberg	offici	million	peopl	time	state	festiv	resid	nation	nbc
democrat	peopl	chariti	im	home	time	artist	architect	thing	million
campaign	hotel	disast	call	day	leagu	song	board	public	broadcast
franklin	state	american	hous	citi	stadium	work	meet	word	report
candid	neighborhood	money	stay	make	giant	show	public	natur	abc
ferrer	engin	group	time	club	san	time	local	day	cruis
poll	corp	rais	dont	louisiana	back	concert	street	govern	program
hop	powel	effort	work	cook	bowl	includ	architectur	media	film
made	billion	food	life	krew	louisiana	orchestra	project	great	channel
hip	busi	org	son	hotel	field	event	urban	make	televis
dont	krt	shelter	left	celebr	win	record	peopl	histori	navi
topic 20	topic 21	topic 22	topic 23	topic 24	topic 25	topic 26	topic 27	topic 28	topic 29
compani	insur	peopl	hospit	guard	nagin	evacu	state	hous	fema
busi	flood	church	patient	nation	neighborhood	water	car	evacue	respons
work	damag	black	health	state	resid	peopl	charg	fema	feder
employe	billion	massachusett	medic	militari	black	offici	law	peopl	agenc
million	state	state	nurs	troop	mayor	resid	court	offici	brown
contract	compani	poverti	care	offici	citi	louisiana	vehicl	home	disast
servic	loss	work	dr	bush	rebuild	rita	investig	houston	govern
worker	mississippi	romney	center	unit	white	area	offic	feder	emerg
bank	home	poor	doctor	forc	peopl	flood	attorney	agenc	secur
custom	homeown	american	peopl	feder	elect	coast	lawyer	hotel	offici
week	pay	evacue	evacu	louisiana	home	state	louisiana	trailer	homeland
oper	claim	servic	state	equip	vote	texa	case	famili	hous
port	cost	communiti	flu	day	hous	center	judg	state	depart
system	allstat	job	emerg	effort	area	wind	report	shelter	report
execut	area	base	staff	relief	flood	emerg	fraud	emerg	manag
line	properti	day	home	presid	return	gulf	station	live	chertoff
area	louisiana	time	day	respons	plan	home	feder	month	white
damag	industri	live	die	blanco	percent	day	offici	apart	bush
small	feder	worker	diseas	disast	lower	houston	file	govern	plan
call	polici	nation	univers	rescu	landrieu	mile	system	assist	investig

Table A.3: A 30 topic LDA model for Hurricane Katrina. Each topic contains the 20 most probable (stemmed) words in its distribution. We stem words according to a Porter stemmer.

Hurricane Sandy LDA									
topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
power	obama	climat	hous	school	broadway	park	train	hospit	insur
util	romney	flood	home	time	street	tree	author	home	compani
servic	presid	chang	water	fund	theater	boardwalk	station	patient	percent
compani	campaign	protect	beach	peopl	time	jersey	line	health	sale
author	elect	build	car	day	work	damag	servic	medic	month
electr	state	rise	live	student	open	fire	tunnel	nurs	market
island	republican	sea	flood	children	perform	seasid	jersey	evacu	busi
custom	vote	water	peopl	public	peopl	shore	gas	emerg	increas
state	polit	risk	point	famili	day	busi	transport	center	million
system	governor	level	fire	american	show	height	power	dr	loss
grid	voter	energi	street	red	week	summer	damag	peopl	industri
long	day	natur	rockaway	donat	power	town	subway	citi	home
verizon	poll	power	back	case	run	time	street	offici	report
nation	democrat	weather	day	work	danc	beach	manhattan	resid	expect
work	peopl	develop	insur	cross	play	work	offici	island	billion
phone	debat	make	damag	govern	light	pier	transit	day	rate
commiss	candid	cost	resid	live	night	island	long	care	week
network	presidenti	state	work	disast	cancel	stand	system	bird	retail
con	time	plan	famili	parent	halloween	back	day	mayor	consum
edison	nation	surg	neighborhood	relief	close	visit	island	mold	claim
topic 10	topic 11	topic 12	topic 13	topic 14	topic 15	topic 16	topic 17	topic 18	topic 19
museum	hous	wind	show	peopl	concert	feder	water	beach	build
art	water	power	time	home	perform	billion	system	sand	street
work	peopl	day	stewart	live	ticket	state	state	island	develop
galleri	build	close	peopl	hous	music	hous	million	park	apart
water	resid	weather	make	water	show	aid	flood	dune	properti
street	home	coast	photo	hotel	million	disast	plant	long	million
damag	volunt	expect	live	day	money	money	cost	offici	floor
flood	food	servic	twitter	polic	benefit	program	car	rockaway	estat
space	day	travel	call	work	hall	damag	occupi	corp	water
center	work	area	work	resid	rais	govern	sewag	project	manhattan
compani	power	offici	news	famili	song	republican	river	debri	flood
build	live	peopl	stori	apart	peopl	jersey	peopl	town	resid
seaport	red	state	includ	time	night	million	dutch	home	real
includ	island	damag	inform	island	work	congress	project	resid	owner
insur	apart	flood	magazin	door	relief	cuomo	build	communiti	squar
offic	week	nation	photograph	evacu	refund	senat	geotherm	sea	damag
artist	street	massachusett	design	call	springsteen	insur	work	day	tenant
aquarium	heat	center	post	worker	jersey	cost	park	boardwalk	month
site	brooklyn	report	print	staten	sale	offici	area	public	feet
research	hook	hour	page	report	band	homeown	engin	work	move

Table A.4: A 20 topic LDA model for Hurricane Sandy. Each topic contains the 20 most probable words in its distribution. We stem words according to a Porter stemmer.

Sandy Topic 0				Sandy Topic 2				Katrina Topic 8			
1	power	51	generat	1	climat	51	coastal	1	price	51	drop
2	util	52	solar	2	flood	52	bloomberg	2	oil	52	reserv
3	servic	53	spokesman	3	chang	53	public	3	percent	53	close
4	compani	54	voic	4	protect	54	barrier	4	energi	54	inflat
5	author	55	energi	5	build	55	part	5	gas	55	spend
6	electr	56	manag	6	rise	56	elev	6	gasolin	56	refin
7	island	57	emerg	7	sea	57	presid	7	rate	57	august
8	custom	58	liberti	8	water	58	system	8	market	58	depart
9	state	59	local	9	risk	59	map	9	week	59	natur
10	system	60	respons	10	level	60	gas	10	product	60	chief
11	grid	61	governor	11	energi	61	vulner	11	month	61	job
12	long	62	prepar	12	natur	62	disast	12	consum	62	end
13	verizon	63	feder	13	power	63	peopl	13	report	63	septemb
14	nation	64	copper	14	weather	64	fuel	14	economi	64	profit
15	work	65	govern	15	develop	65	event	15	compani	65	feder
16	phone	66	rate	16	make	66	polici	16	increas	66	gain
17	commiss	67	problem	17	cost	67	step	17	gulf	67	retail
18	network	68	regul	18	state	68	zone	18	fuel	68	record
19	con	69	link	19	plan	69	damag	19	expect	69	interest
20	edison	70	cost	20	surg	70	live	20	gallon	70	damag
21	day	71	percent	21	nation	71	effect	21	cent	71	share
22	restor	72	backup	22	warm	72	unit	22	barrel	72	rais
23	public	73	report	23	infrastructur	73	coast	23	higher	73	term
24	communic	74	investig	24	global	74	research	24	stock	74	demand
25	million	75	damag	25	increas	75	agenc	25	economist	75	futur
26	worker	76	chief	26	citi	76	recent	26	quarter	76	billion
27	execut	77	elli	27	reduc	77	long	27	econom	77	level
28	cuomo	78	ed	28	carbon	78	generat	28	high	78	declin
29	offici	79	wireless	29	environment	79	heat	29	cost	79	hit
30	employe	80	carrier	30	scientist	80	effort	30	suppli	80	investor
31	batteri	81	presid	31	billion	81	rais	31	day	81	survey
32	offic	82	hit	32	engin	82	pollut	32	analyst	82	state
33	oper	83	counti	33	studi	83	industri	33	refineri	83	remain
34	week	84	general	34	time	84	project	34	nation	84	effect
35	call	85	consum	35	resili	85	standard	35	industri	85	hurrican
36	time	86	consolid	36	futur	86	code	36	time	86	impact
37	charg	87	equip	37	emiss	87	hit	37	rose	87	heat
38	provid	88	director	38	area	88	issu	38	point	88	credit
39	plan	89	issu	39	govern	89	ocean	39	rise	89	labor
40	includ	90	cabl	40	feet	90	oyster	40	fell	90	servic
41	pay	91	critic	41	requir	91	design	41	fed	91	american
42	wire	92	cellphon	42	higher	92	larg	42	averag	92	continu
43	statu	93	technolog	43	mayor	93	offici	43	trade	93	unit
44	failur	94	run	44	extrem	94	warn	44	million	94	show
45	home	95	caus	45	propos	95	face	45	growth	95	produc
46	line	96	telephon	46	high	96	east	46	index	96	note
47	board	97	substat	47	plant	97	sever	47	yesterday	97	petroleum
48	panel	98	guard	48	includ	98	univers	48	sale	98	earn
49	hour	99	place	49	insur	99	decad	49	crude	99	concern
50	area	100	chairman	50	world	100	solut	50	coast	100	import

Table A.5: A 100 word extension of selected topics from the Sandy and Katrina LDA models.

Appendix B

Supplementary Materials for Chapter 4

B.1 Anomaly Correlation

We use anomaly correlation (Pearson Correlation) to determine the relationship between the Twitter happiness time series and the polling data. When doing so, we subtract the mean of the series, m, from each data point, h_i, to determine anomalies, and then calculate the cosine of the angle between the series of anomalies, i.e.,

$$H_{an} = \{h_i - m\}_{i=1}^{L} \tag{B.1}$$

$$Corr_{\text{an}}(H, P) = \frac{H_{an} \cdot P_{an}}{||H_{an}|| \cdot ||P_{an}||} \tag{B.2}$$

The variables H and P represent happiness time series and polling time series respectively, and L is the length of the time series.

B.2 Additional Figures and Tables

Each word in our data set was previously assigned a happiness score through Amazon's Mechanical Turk (labMT scores). We investigate the relationship between surveyed scores and ambient happiness scores in Fig. B.1. We find a positive slope, indicating that ambient happiness rises with surveyed happiness, however we see a much smaller range of scores, which can be attributed to averaging a large amount of words. We give the top 10 and bottom 10 words sorted by ambient happiness in Table B.1. Top words included birthday wishes and prize giveaways, and bottom words suggest legal news stories.

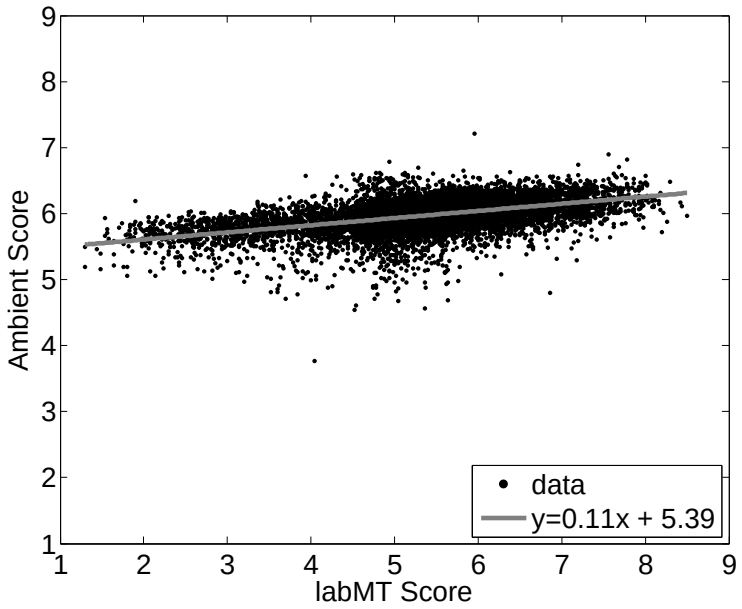

Figure B.1: Surveyed happiness versus ambient happiness for all words in the labMT dataset. The small positive slope indicates that ambient happiness increases with surveyed happiness, however ambient happiness covers a smaller range of values. An interactive version is available in the online Appendix.

Top 10				Bottom 10			
Rank	Word	Ambient	labMT	Rank	Word	Ambient	labMT
1.	collected	7.21	5.96	9780.	defendants	4.87	4.26
2.	merry	6.90	7.56	9781.	prosecutors	4.87	4.20
3.	birthday	6.82	7.78	9782.	suspects	4.86	3.60
4.	iya	6.79	4.94	9783.	suspected	4.81	3.52
5.	prizes	6.73	7.20	9784.	indicted	4.81	3.60
6.	b-day	6.71	7.68	9785.	seas	4.80	6.84
7.	2-bath	6.69	5.28	9786.	pleaded	4.78	3.84
8.	entered	6.65	5.82	9787.	sentenced	4.71	3.70
9.	giveaway	6.62	6.38	9788.	civilians	4.68	5.84
10.	shipping	6.61	5.46	9789.	welt	3.77	4.04

Table B.1: The top 10 and bottom 10 words sorted by ambient happiness. Ambient happiness is calculated using word frequencies from September 2008 through November 2015. Non-English words and words with frequencies under 1000 are removed, leaving 9789 remaining in our ambient dataset.

Top 10				Bottom 10			
Rank	Word	Ambient	labMT	Rank	Word	Ambient	labMT
1.	birthday	6.82	7.78	9780.	seas	4.80	6.84
2.	b-day	6.71	7.68	9781.	civilians	4.86	5.84
3.	merry	6.90	7.56	9782.	defendants	4.87	4.26
4.	prizes	6.73	7.20	9783.	prosecutors	4.87	4.20
5.	giveaway	6.62	6.38	9784.	welt	3.77	4.04
6.	collected	7.21	5.96	9785.	pleaded	4.78	3.84
7.	entered	6.65	5.82	9786.	sentenced	4.71	3.70
8.	shipping	6.61	5.46	9787.	indicted	4.81	3.60
9.	2-bath	6.69	5.28	9788.	suspects	4.86	3.60
10.	iya	6.79	4.94	9789.	suspected	4.81	3.52

Table B.2: The top 10 and bottom 10 words according to ambient happiness, sorted by labMT score.

Top 10				Bottom 10			
Rank	Word	Ambient	labMT	Rank	Word	Ambient	labMT
1.	laughter	5.96	8.50	9780.	died	5.76	1.56
2.	happiness	6.11	8.44	9781.	kill	5.71	1.56
3.	love	6.17	8.42	9782.	killed	5.56	1.56
4.	happy	6.48	8.30	9783.	cancer	5.93	1.54
5.	laughed	5.87	8.26	9784.	death	5.66	1.54
6.	laugh	6.01	8.22	9785.	murder	5.39	1.48
7.	laughing	5.71	8.20	9786.	terrorism	5.16	1.48
8.	excellent	6.31	8.18	9787.	rape	5.46	1.44
9.	laughs	6.06	8.18	9788.	suicide	5.49	1.30
10.	joy	6.19	8.16	9789.	terrorist	5.19	1.30

Table B.3: The top 10 and bottom 10 words according to labMT score.

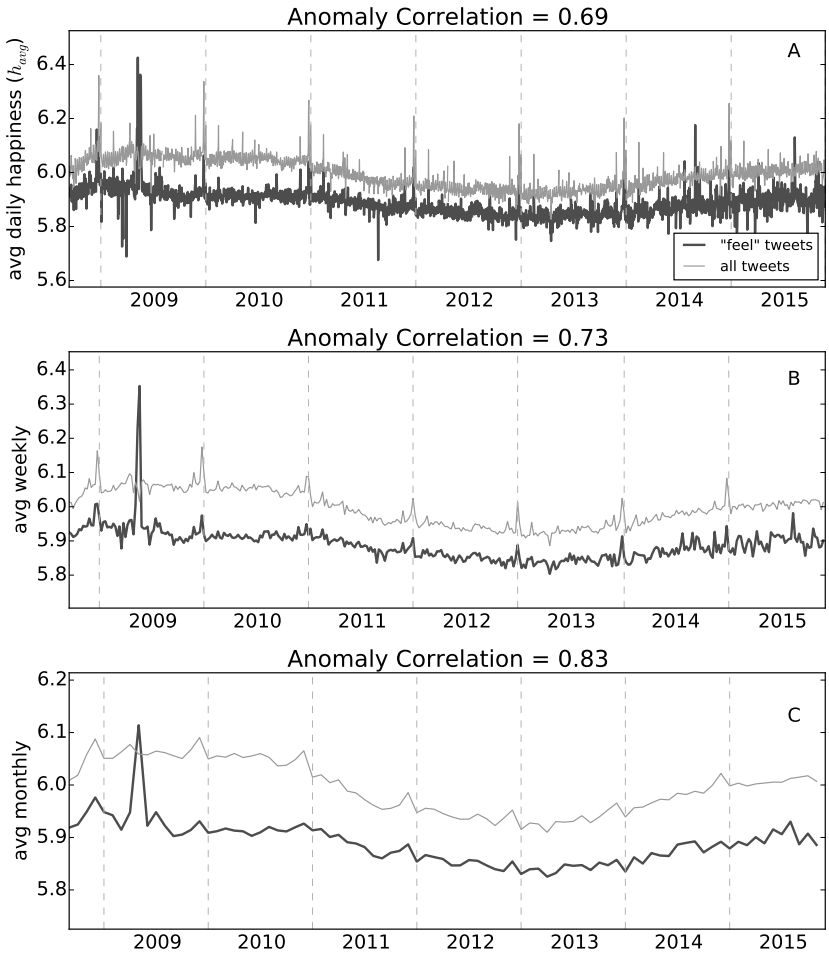

Figure B.2: Ambient happiness of "feel" compared to overall happiness by (A) day, (B) week, and (C) month. The ambient happiness of the word "feel" correlates strongly with the average happiness of tweets that do not contain "feel", and the correlation grows stronger as we decrease the temporal resolution. This indicates that the shape of overall happiness remains the same whether a user is directly or indirectly expressing an emotion on Twitter. An interactive version of the overall signal can be found at `hedonometer.org`.

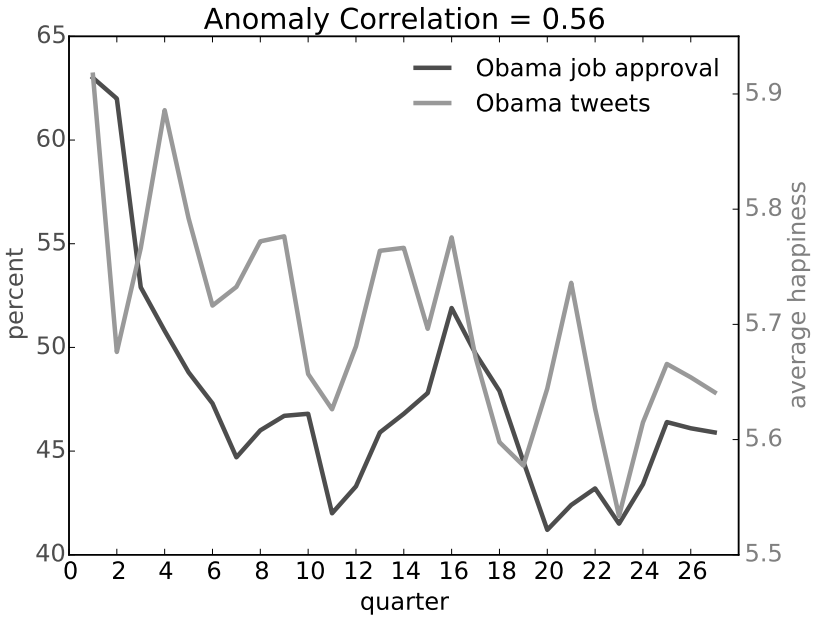

Figure B.3: Average quarterly happiness of tweets containing "Obama" with Obama's quarterly job approval rating from Gallup. We find a relatively high correlation with solicited polling data.

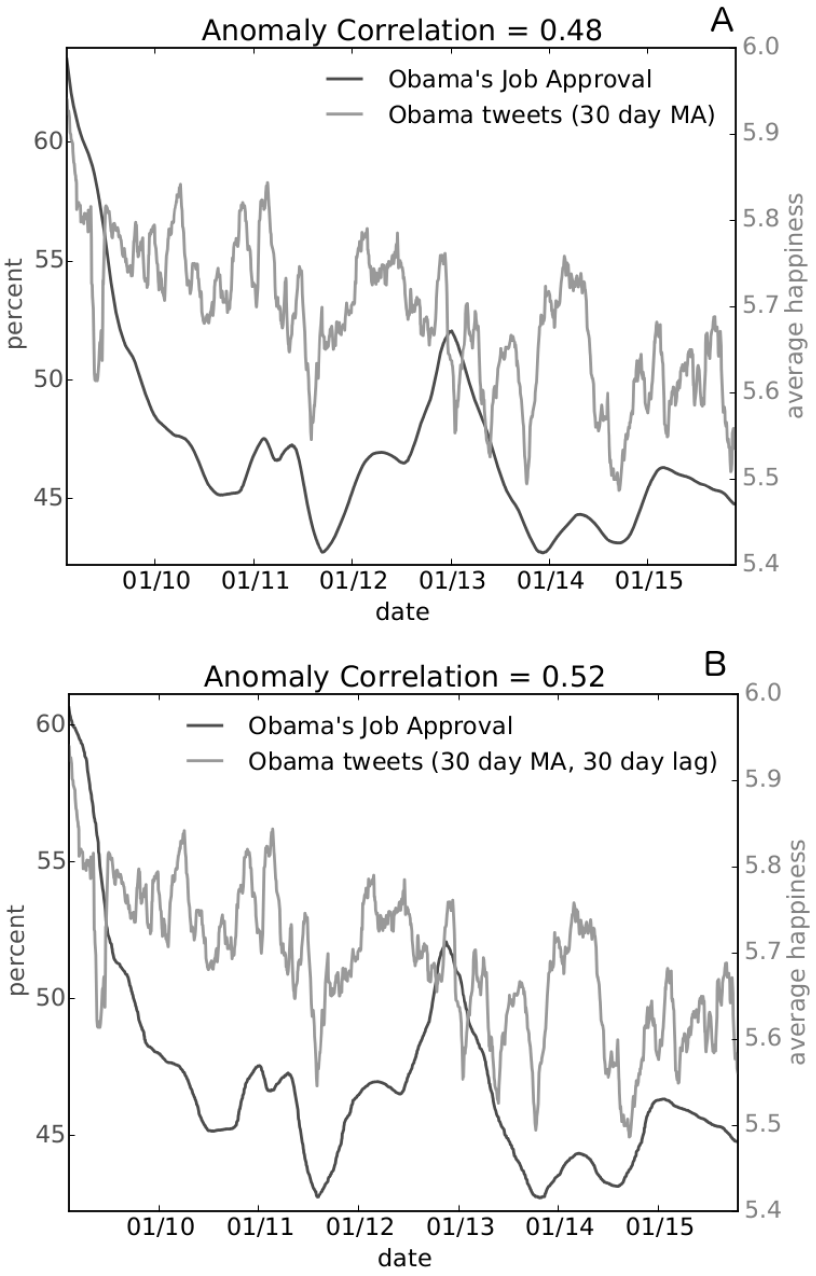

Figure B.4: (A) Average daily happiness of tweets containing "Obama" with Obama's daily job approval rating from Pollster. (B) 30 day lag. We find a relatively high correlation with solicited polling data.

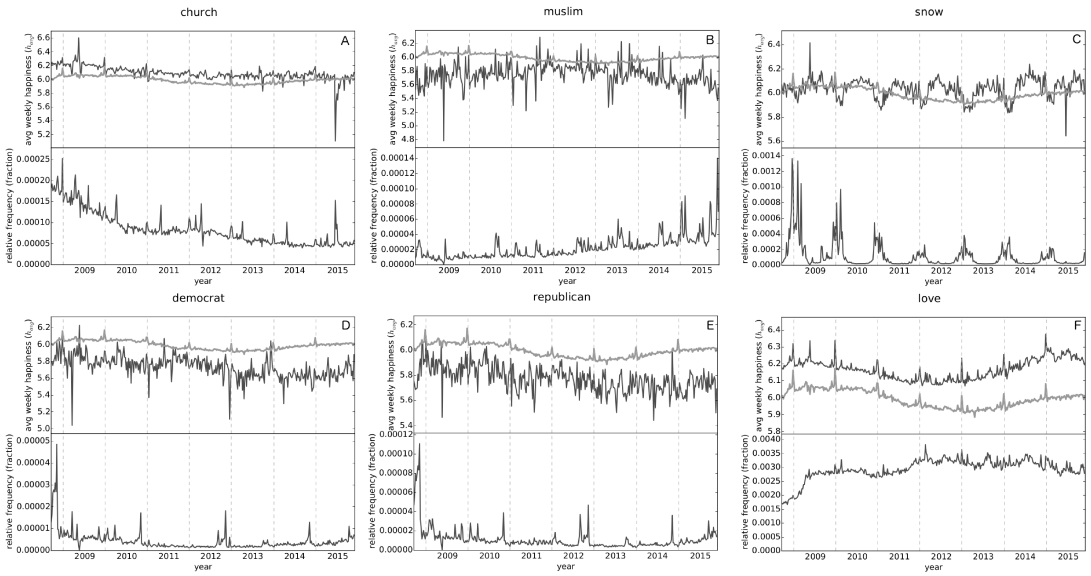

Figure B.5: Six examples of weekly ambient happiness time series (top) with the weekly relative frequency for the word (bottom). Relative frequency is calculated by dividing the total frequency of the word by the total frequency of all words on a given week. (A) "church" (B) "mulsim" (C) "snow" (D) "democrat" (E) "republican" (F) "love"

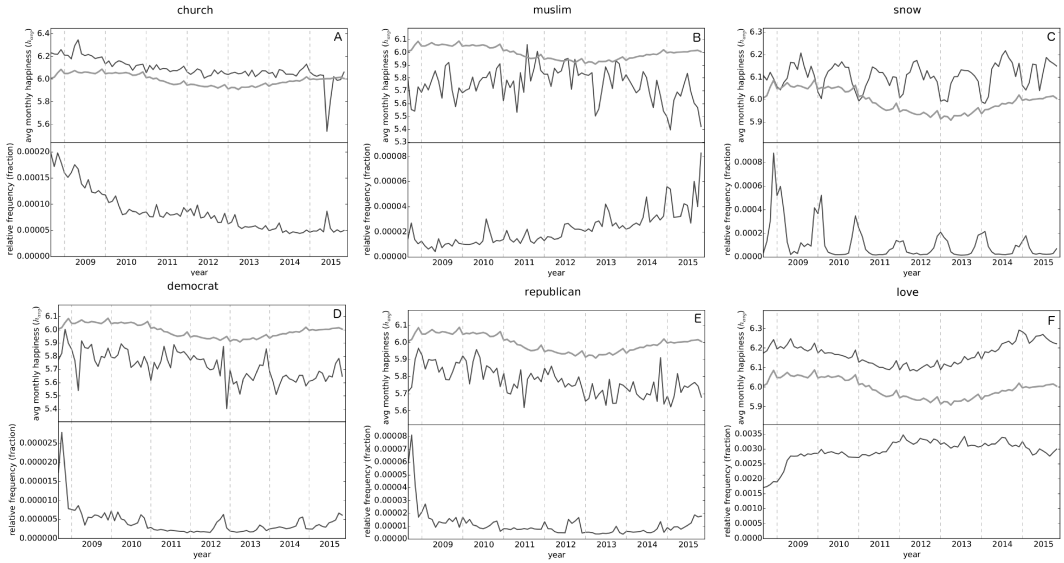

Figure B.6: Six examples of monthly ambient happiness time series (top) with the monthly relative frequency for the word (bottom). Relative frequency is calculated by dividing the total frequency of the word by the total frequency of all words on a given month. (A) "church" (B) "mulsim" (C) "snow" (D) "democrat" (E) "republican" (F) "love"

B.3 Gallup Yearly Polling

Gallup trends provide yearly polling data on many topics without a subscription. The Gallup survey questions can be found in Table B.4. These polls, however, take place only once a year in the same month over several days. This presents a challenge as to the amount of Twitter data we should include in our correlations, as opinions may change. For each Gallup datapoint, we use the current year's worth of tweets from 2009 through 2015 for various subjects of national or global interest. Fig. B.7 shows several topics that correlate quite well with ambient happiness on Twitter. We find

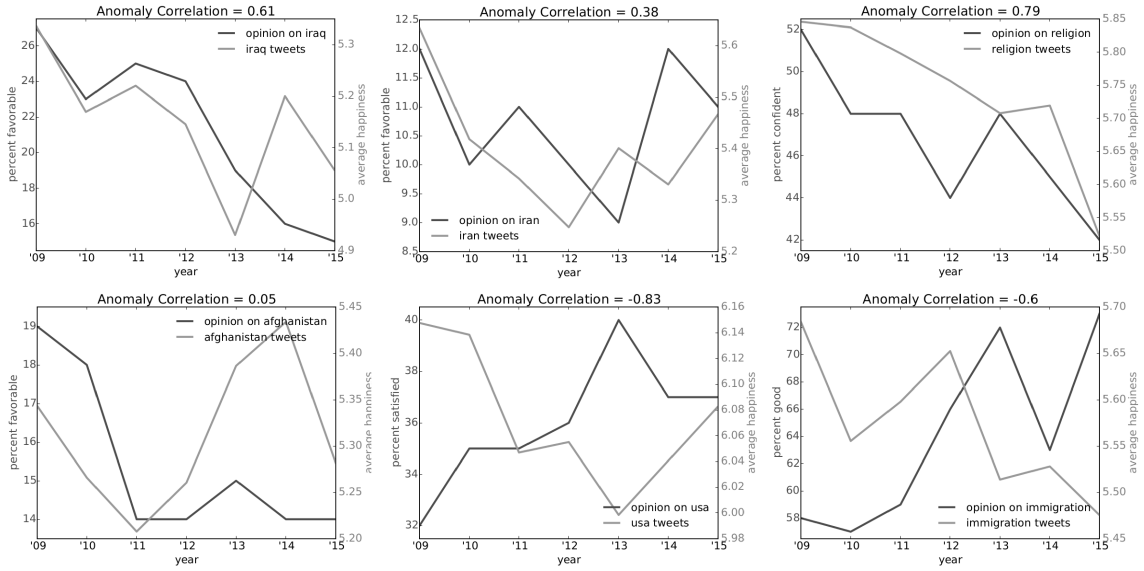

Figure B.7: Correlations between average ambient happiness and opinion polls on various global subjects. We obtain varying levels of correlation between the topics due the limited availability of traditional polling data. For example, Twitter sentiment tracks public opinion surrounding Iraq and religion quite well, but performs poorly on Afghanistan. The specific questions can be found in Table B.4.

that the favorability of two major countries, Iran and Iraq, has a positive correlation with the ambient happiness of "Iran" and "Iraq". We also find that the United States

opinion on religion has a strong positive correlation with yearly ambient happiness of “religion". Other topics, including the United States opinion on Afghanistan and immigration show no significant correlation to Twitter data. There is a strong negative correlation between the satisfaction of the position of the United States in the world, indicating there may be some sarcasm associated with “usa" tweets.

Topic	Survey Question	Frequency	Source
Iraq	What is your overall opinion of Iraq? Is it very favorable, mostly unfavorable, mostly unfavorable, or very unfavorable?	Yearly	Gallup
Iran	What is your overall opinion of Iran? Is it very favorable, mostly unfavorable, mostly unfavorable, or very unfavorable?	Yearly	Gallup
Afghanistan	What is your overall opinion of Afghanistan? Is it very favorable, mostly unfavorable, mostly unfavorable, or very unfavorable?	Yearly	Gallup
USA	On the whole, would you say you are satisfied or dissatisfied with the position of the United States int he world today?	Yearly	Gallup
Religion	Please tell me how much confidence you , yourself, have in the church or organized religion – a great deal, quite a lot, some, or very little?	Yearly	Gallup
Immigration	On the whole, do you think immigration is a good thing or a bad thing for this country today?	Yearly	Gallup
Obama	Do you approve of disapprove of the way Barak Obama is handling his job as president?	Quarterly	Gallup
Obama	Average of latest opinion polls on Obama’s job approval	Daily	Pollster

Table B.4: Survey questions for polling data from various resources used in our analysis.

Bibliography

[1] Gallup trends. http://www.gallup.com/poll/trends.aspx. Accessed: 2016-03-08.

[2] Google correlate. https://www.google.com/trends/correlate,.

[3] Pollster API. http://elections.huffingtonpost.com/pollster/api. Accessed: 2016-03-08.

[4] University of Michigan index of consumer sentiment. http://www.sca.isr.umich.edu/tables.html. Accessed: 2016-03-08.

[5] Daniel R. Abbasi. *Americans and Climate Change: Closing the Gap between Science and Action*. Yale school of forestry & environmental studies publication series, 2006.

[6] Xiaoran An, Auroop R Ganguly, Yi Fang, Steven B Scyphers, Ann M Hunter, and Jennifer G Dy. Tracking climate change opinions from Twitter data. *Workshop on Data Science for Social Good*, 2014.

[7] William RL Anderegg, James W Prall, Jacob Harold, and Stephen H Schneider. Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27):12107–12109, 2010.

[8] Dolan Antenucci, Michael R. Andwerson, Penghua Zhao, and Michael Cafaerlla. A query system for social media signals. 2015.

[9] Liisa Antilla. Climate of scepticism: Us newspaper coverage of the science of climate change. *Global environmental change*, 15(4):338–352, 2005.

[10] Liisa Antilla. Self-censorship and science: a geographical review of media coverage of climate tipping points. *Public Understanding of Science*, 2008.

[11] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[12] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[13] Pablo Barberá. Less is more? how demographic sample weights can improve public opinion estimates based on twitter data.

[14] Ralf Barkemeyer, Suraje Dessai, Beatriz Monge-Sanz, Barbara Gabriella Renzi, and Giulio Napolitano. Linguistic analysis of IPCC summaries for policymakers and associated coverage. *Nature Climate Change*, page 10.1038/nclimate2824, 2015.

[15] Allan Bell. Media (mis) communication on the science of climate change. *Public understanding of science*, 3(3):259–275, 1994.

[16] Michael W Berry and Murray Browne. *Understanding search engines: mathematical modeling and text retrieval*, volume 17. Siam, 2005.

[17] Yves Bestgen. Improving text segmentation using latent semantic analysis: A reanalysis of choi, wiemer-hastings, and moore (2001). *Computational Linguistics*, 32(1):5–12, 2006.

[18] John L Beven, Lixion A Avila, Eric S Blake, Daniel P Brown, James L Franklin, Richard D Knabb, Richard J Pasch, Jamie R Rhome, and Stacy R Stewart. Atlantic hurricane season of 2005. *Monthly Weather Review*, 136(3):1109–1173, 2008.

[19] Eirc S. Blake, Tom B. Kimberlian, Robert J. Berg, John P. Cangialosi, and John L. Beven. Tropical cyclone report, hurricane sandy. *National Hurricane Center*, 2013.

[20] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[21] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[22] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit

assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.

[23] Maxwell T Boykoff. *Who speaks for the climate?: Making sense of media reporting on climate change*. Cambridge University Press, 2011.

[24] Maxwell T Boykoff and Jules M Boykoff. Climate change and journalistic norms: A case-study of US mass-media coverage. *Geoforum*, 38(6):1190–1204, 2007.

[25] Ronald Brownstein. Hard choices blow in the winds of katrina, and now rita. *The Los Angeles Times*, Sep 26 2005.

[26] Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, 2012.

[27] Don Carli. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/dcarli/status/953288121.

[28] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[29] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9(6):811–824, 2012.

[30] Robert B Cialdini and Nathalie Garde. *Influence*, volume 3. A. Michel, 1987.

[31] Eric M Clark, Jake Ryland Williams, Chris A Jones, Richard A Galbraith, Christopher M Danforth, and Peter Sheridan Dodds. Sifting robotic from organic text: A natural language approach for detecting automation on twitter. *Journal of Computational Science*, 2015.

[32] Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. *PLoS ONE*, 10(8), 2015.

[33] Julia B Corbett and Jessica L Durfee. Testing public (un) certainty of science media representations of global warming. *Science Communication*, 26(2):129–151, 2004.

[34] James W Dearing and Everett M Rogers. *Agenda-setting*, volume 6. Sage Publications, 1996.

[35] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[36] John P Dickerson, Vadim Kagan, and VS Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 620–627. IEEE, 2014.

[37] Peter U Diehl, Bruno U Pedroni, Andrew Cassidy, Paul Merolla, Emre Neftci, and Guido Zarrella. Truehappiness: Neuromorphic emotion recognition on truenorth. *arXiv preprint arXiv:1601.04183*, 2016.

[38] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449, 2013.

[39] Cory Doctorow. *Twitter*, 2009 (accessed March 19, 2015). https://twitter.com/doctorow/status/1482803994.

[40] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752, 2011.

[41] Peter T Doran and Maggie Kendall Zimmerman. Examining the scientific consensus on climate change. *Eos, Transactions American Geophysical Union*, 90(3):22–23, 2009.

[42] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *arXiv preprint arXiv:1407.5225*, 2014.

[43] Christopher Field and Maarten Van Aalst. *Climate change 2014: Impacts, adaptation, and vulnerability*, volume 1. IPCC, 2014.

[44] Christopher B Field. *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[45] E. M. Fischer and R. Knutti. Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nature Climate Change*, advance online publication, April 2015.

[46] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, 3, 2013.

[47] William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, pages 1–37, 1989.

[48] Open View Gardens. *Twitter*, 2011 (accessed March 19, 2015). https://twitter.com/openviewgardens/status/99975488293978112.

[49] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, page 0894439313493979, 2013.

[50] Doris A Graber. *Mass media and American politics*. SAGE, 2009.

[51] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[52] David J Griggs and Maria Noguer. Climate change 2001: the scientific basis. contribution of working group i to the third assessment report of the intergovernmental panel on climate change. *Weather*, 57(8):267–269, 2002.

[53] David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics, 2008.

[54] Lawrence C Hamilton and Mary D Stampone. Blowin' in the wind: Short-term weather and belief in anthropogenic climate change. *Weather, Climate, and Society*, 5(2):112–119, 2013.

[55] Rob Heidrick. Hurricane season could bring higher energy prices. *Texas Enterprise*, 2013.

[56] David K. Henry, Sandra Cooke-Hull, Jacqueline Savukinas, Fenwick Yu, Nicholas Elo, and Bradford Vac Arnum. Economic impact of Hurriane Sandy: Potential economic activity lost and gained in New Jersey and New York. Technical report, U.S. Department of Commerce, 09 2013.

[57] David J Hess. Transitions in energy systems: The mitigation–adaptation relationship. *Science as Culture*, 22(2):197–203, 2013.

[58] Peter D Howe, Hilary Boudet, Anthony Leiserowitz, and Edward W Maibach. Mapping the shadow of experience of extreme weather events. *Climatic Change*, 127(2):381–389, 2014.

[59] Peter D Howe, Matto Mildenberger, Jennifer R Marlon, and Anthony Leiserowitz. Geographic variation in opinions on climate change at state and local scales in the usa. *Nature Climate Change*, 2015.

[60] Daniel G. Huber and Jay Gulledge. *Extreme weather and climate change: Understanding the link, managing the risk*. Pew Center on Global Climate Change Arlington, 2011.

[61] IPCC. Climate change 2014 mitigation of climate change, intergovernmental panel on climate change. 2014.

[62] Sep Kamvar and Jonathan Harris. *We feel fine: An almanac of human emotion*. Simon and Schuster, 2009.

[63] Thomas Kaplan. Experts advise Cuomo on disaster measures. *The New York Times*, January 4 2013.

[64] Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics, 2006.

[65] Andrei P Kirilenko, Tatiana Molodtsova, and Svetlana O Stepchenkova. People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Global Environmental Change*, 30:92–100, 2015.

[66] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PLoS ONE*, 7(1):e29484, 2012.

[67] RD Knabb, JR Rhome, and DP Brown. Tropical cyclone report?hurricane katrina. national hurricane center. *Miami, FL*, 2006.

[68] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Nowcasting disaster damage. *arXiv preprint arXiv:1504.06827*, 2015.

[69] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3):e1500779, 2016.

[70] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[71] Thomas K Landauer and Michael L Littman. Computerized cross-language document retrieval using latent semantic indexing, April 5 1994. US Patent 5,301,109.

[72] A Leiserowitz, E Maibach, C Roser-Renouf, and JD Hmielowski. Extreme weather, climate & preparedness in the american mind. *Yale University and George Mason University. New Haven, CT.)(Report)*, 2012.

[73] Anthony Leiserowitz, Edward Maibach, Connie Roser-Renouf, Geoff Feinberg, and Peter Howe. Climate change in the american mind: Americans' global warming beliefs and attitudes in April, 2013. *Yale University and George Mason University. New Haven, CT: Yale Project on Climate Change Communication*, 2013.

[74] Anthony Leiserowitz, Edward Maibach, Connie Roser-Renouf, and Nicholas Smith. Global warming's six americas, march 2012 and november 2011. *Yale University and George Mason University*, 2012.

[75] Anthony A Leiserowitz. American risk perceptions: Is climate change dangerous? *Risk analysis*, 25(6):1433–1442, 2005.

[76] Ye Li, Eric J Johnson, and Lisa Zaval. Local warming daily temperature change influences belief in global warming. *Psychological Science*, 2011.

[77] Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PloS one*, 9(5):e94093, 2014.

[78] Yu-Ru Lin, Drew Margolin, Brian Keegan, and David Lazer. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*, pages 737–748. International World Wide Web Conferences Steering Committee, 2013.

[79] David MacKay. *Sustainable Energy-without the hot air*. UIT Cambridge, 2008.

[80] Edward W Maibach, Anthony Leiserowitz, Connie Roser-Renouf, and CK Mertz. Identifying like-minded audiences for global warming public engagement campaigns: An audience segmentation analysis and tool development. *PloS one*, 6(3):e17571, 2011.

[81] Michael E. Mann and Kerry A. Emanuel. Atlantic hurricane trends linked to climate change. *Eos, Transactions American Geophysical Union*, 87(24):233–241, 2006.

[82] Philip M McCarthy, Stephen W Briner, Vasile Rus, and Danielle S McNamara. Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In *Natural language processing and text mining*, pages 107–122. Springer, 2007.

[83] Media Insight Project. How Americans get their news. *The Personal News Cycle*, 2014.

[84] Yelena Mejova, Ingmar Weber, and Michael W Macy. *Twitter: a digital socioscope*. Cambridge University Press, 2015.

[85] Bert Metz. *Controlling climate change*. Cambridge University Press, 2009.

[86] William K Michener, Elizabeth R Blood, Keith L Bildstein, Mark M Brinson, and Leonard R Gardner. Climate change, hurricanes and tropical storms, and rising sea level in coastal wetlands. *Ecological Applications*, 7(3):770–801, 1997.

[87] Greg Miller. Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815, 2011.

[88] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[89] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 2012.

[90] David Murray, Joel B Schwartz, and S Robert Lichter. *It ain't necessarily so: How media make and unmake the scientific picture of reality*. Rowman & Littlefield, 2001.

[91] Teresa A. Myers, Edward W. Maibach, Connie Roser-Renouf, Karen Akerlof, and Anthony A. Leiserowitz. The relationship between personal experience and belief in the reality of global warming. *Nature Climate Change*, 3(4):343–347, 2013.

[92] Robert K Nelson. Mining the dispatch. *Mining the dispatch*, 2010.

[93] NewGreenStuff. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/NewGreenStuff/status/953099924.

[94] Humanitarian News. *Twitter*, 2010 (accessed March 19, 2015). https://twitter.com/HumanityNews/status/11612292989.

[95] OneWorld News. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/OneWorld_News/status/1083004712.

[96] OneWorld News. *Twitter*, 2008 (accessed March 19, 2015). https://twitter.com/OneWorld_News/status/953758970.

[97] NewsOnGreen. *Twitter*, 2010 (accessed March 19, 2015). https://twitter.com/NewsOnGreen/status/11608867076.

[98] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[99] U.S. Department of Energy. Hurricane Katrina situation report #11. *Office of Electricity Delivery and Energy Reliability (OE)*, 2005.

[100] U.S. Department of Energy. Comparing the impacts of northeast hurricanes on energy infrastructure. *Office of Electricity Delivery and Energy Reliability (OE)*, 2013.

[101] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *ICWSM*, pages 265–272, 2011.

[102] TR Peterson and JL Thompson. Environmental risk communication: responding to challenges of complexity and uncertainty. *Handbook of risk and crisis communication (pp. 591Á606). New York: Routledge*, 2009.

[103] Tuan Q Phan and Edoardo M Airoldi. A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences*, 112(21):6595–6600, 2015.

[104] Tuan Q. Phan and Edoardo M. Airoldi. A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences*, 2015.

[105] Roger A Pielke Jr, Chris Landsea, Max Mayfield, Jim Laver, and Richard Pasch. Hurricanes and global warming. *Bulletin of the American Meteorological Society*, 86(11):1571–1575, 2005.

[106] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[107] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3, 2013.

[108] Susanna Hornig Priest. *Doing media research: an introduction*. Sage, 2009.

[109] Joseph T Ripberger, Hank C Jenkins-Smith, Carol L Silva, Deven E Carlson, and Matthew Henderson. Social media and severe weather: Do tweets provide a valid indicator of public attention to severe weather risk communication? *Weather, Climate, and Society*, 6(4):520–530, 2014.

[110] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, pages 9–17. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), 2009.

[111] Laurie A. Rudman, Meghan C. McLean, and Martin Bunzl. When truth is personally inconvenient, attitudes change the impact of extreme weather on implicit support for green politicians and explicit climate-change beliefs. *Psychological science*, 2013.

[112] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[113] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[114] Jonathon P Schuldt, Sara H Konrath, and Norbert Schwarz. "global warming" or "climate change"? whether the planet is warming depends on question wording. *Public Opinion Quarterly*, page nfq073, 2011.

[115] Shifting Solutions. *Twitter*, 2009 (accessed March 19, 2015). https://twitter.com/ShiftSolutions/status/1485975759.

[116] Jennie C Stephens, Gabriel M Rand, and Leah L Melnick. Wind energy in us media: a comparative state-level analysis of a critical climate change mitigation technology. *Environmental Communication*, 3(2):168–190, 2009.

[117] Jennie C Stephens, Elizabeth J Wilson, Tarla R Peterson, and James Meadowcroft. Getting smart? climate change and the electric grid. *Challenges*, 4(2):201–216, 2013.

[118] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[119] Mark Twain, Michael Barry Frank, Robert Pack Browning, Lin Salamo, Frederick Anderson, and Mark Twain. *Mark Twain's Notebooks & Journals, Volume III:(1883-1891)*, volume 8. Univ of California Press, 1980.

[120] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T Jost, Jonathan Nagler, and Joshua Tucker. Social media and political communication: a survey of twitter users during the 2013 italian general election. *Rivista italiana di scienza politica*, 43(3):381–410, 2013.

[121] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[122] Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[123] Elizabeth J Wilson, Jennie C Stephens, Tarla Rai Peterson, and Miriam Fischlein. Carbon capture and storage in context: The importance of state policy and discourse in deploying emerging energy technologies. *Energy Procedia*, 1(1):4519–4526, 2009.

[124] Kris M Wilson. Mass media as sources of global warming knowledge. *Mass Comm Review*, 22:75–89, 1995.

[125] Kris M Wilson. Drought, debate, and uncertainty: measuring reporters' knowledge and ignorance about climate change. *Public Understanding of Science*, 9(1):1–13, 2000.

[126] WWF. *Twitter*, 2012 (accessed March 19, 2015). https://twitter.com/WWF/status/196902312797671424.

[127] Tze-I Yang, Andrew J Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics, 2011.

[128] Lisa Zaval, Elizabeth A Keenan, Eric J Johnson, and Elke U Weber. How warm days increase belief in global warming. *Nature Climate Change*, 4(2):143–147, 2014.