Online Collective Attention: Standardizing the Measurement of Sociotechnical Systems.

A Dissertation Presented

by

Michael V. Arnold

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 Complex Systems and Data Science

May, 2024

Defense Date: March 18, 2023
Dissertation Examination Committee:

Christopher M. Danforth, Ph.D., Advisor
Peter Sheridan Dodds, Ph.D., Advisor
Matthew Price, Ph.D., Chairperson
Jeremiah Onaolapo, Ph.D.
Jean-Gabriel Young, Ph.D.
Holger Hoock, DPhil, Dean of Graduate College

Abstract

Unprecedented growth in digital information has transformed data in the social sciences from scarce to abundant. Social media in particular has created a new algorithmically mediated sociotechnical system, one where billions of daily communications influence our perspective on reality in poorly understood ways. Boosted by advances in high performance computing, natural language processing, and machine learning, the digital traces left behind by these electronic breadcrumbs hold immense promise for measuring collective attention and sentiment at the societal scale.

In one study, using hurricane name mentions as a proxy for awareness. We find that the exogenous temporal dynamics are remarkably similar across storms, but that overall collective attention varies widely even among storms causing comparable deaths and damage. We construct 'hurricane attention maps' and observe that hurricanes causing deaths on (or economic damage to) the continental United States generate substantially more attention in English language tweets than those that do not. We find that a hurricane's Saffir-Simpson wind scale category assignment is strongly associated with the amount of attention it receives. Higher category storms receive higher proportional increases of attention per proportional increases in number of deaths or dollars of damage, than lower category storms. The most damaging and deadly storms of the 2010s, Hurricanes Harvey and Maria, generated the most attention and were remembered the longest, respectively. On average, a category 5 storm receives 4.6 times more attention than a category 1 storm causing the same number of deaths and economic damage.

In a second study, we explore using well curated, large-scale corpora of social media posts containing broad public opinion as an alternative data source to complement traditional surveys. While surveys are effective at collecting representative samples and are capable of achieving high accuracy, they can be both expensive to run and lag public opinion by up to a month. Both of these drawbacks could be overcome with a real-time, high volume data stream and fast analysis pipeline. A central challenge in orchestrating such a data pipeline is devising an effective method for rapidly selecting the best corpus of relevant documents for analysis. Querying with keywords alone often includes irrelevant documents that are not easily disambiguated with bag-of-words natural language processing methods. Here, we explore methods of corpus curation to filter irrelevant tweets using pre-trained transformer-based models, fine-tuned for our binary classification task on hand-labeled tweets. We are able to achieve F1 scores of up to 0.95. The low cost and high performance of fine-tuning such a model suggests that our approach could be of broad benefit as a pre-processing step for social media datasets with uncertain corpus boundaries.

In a third chapter, I describe my contributions to nearly 2 dozen studies leveraging Twitter data to explore collective attention, sentiment, and language. These cover a range of topics including the COVID-19 pandemic, politicians and K-pop stars, public health, and social movements, demonstrating the broad value of social media data in interdisciplinary research.

Citations

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

Arnold, M. V., Dewhurst, D. R., Alshaabi, T., Minot, J. R., Adams, J. L., Danforth, C. M., & Dodds, P. S.. (2021). Hurricanes and hashtags: Characterizing online collective attention for natural disasters. *PloS one.* 16(5), e0251762.

AND

Arnold, M. V., Dodds, P. S., & Danforth, C. M.. (2023). Curating corpora with classifiers: A case study of clean energy sentiment online. *In Review.*

Acknowledgements

It's difficult, maybe impossible, to adequately express my gratitude to the many people who have been instrumental in helping me to and through this journey. I feel incredibly fortunate to be a part of so many vibrant communities, full of curiosity, kindness and passion. First, to my partner Summer, whose advice and companionship have been a bedrock. I couldn't be defending my dissertation now without her patient support. To my parents, who nurtured curiosity and a wry sense of humor early on and provided so much advice and reassurance over the course of my studies. To my brother, Joe, for teaching me that so much rests on the decision of where to position yourself and focus. I'm still absorbing this.

To my advisors Professor Peter Dodds and Professor Chris Danforth, whose classes inspired me as an undergrad to continue to grad school so many years ago and have continued to challenge and encourage me ever since. Thanks to Peter for generating and entertaining so many ideas, creating space to broadly explore and create, and for much advice on writing. Thanks to Chris for connecting me with so many people, papers, and opportunities, for patiently guiding me to narrow scope and prioritize what matters most. This dissertation wouldn't have been possible without their kindness and support.

Thanks to Jeremiah Onaolapo, Jean-Gabriel Young, and Matthew Price for serving on my dissertation committee. To my other professors at the University of Vermont, Laurent Hébert-Dufresne, James Bagrow, Safwan Wshah, for designing so many rigorous, engaging courses that built up the foundational knowledge and skills that this work relies on.

I've been incredibly privileged to work with many brilliant co-authors over the years; thanks to Kesley Linnell, Mikaela Fudolig, Anne Marie Stupinski, Colin Van Oort, Morgan Weaving, Julia Zimmerman, Sean Rogers, Henry Wu, Ryan Gallagher, Brooke Foucault Welles, Randall Harp, Matthew Price, Ellen McGinnis, Ryan McGinnis, Bryn Loftness, Juniper Lovato, Morgan Frank, and Andrew Reagan. I'm excited to get back to work with many of you as this dissertation period ends. To Thayer Alshaabi for his inspiring diligence

and attention to detail, and for much late night debugging and commiseration. To Yoshi Bird for boundless encouragement and passion for all things housing. Thanks to Josh Minot, Jane Adams, and David Dewhurst for our work together on Storywrangler.

And many more thanks to Amelia Tarren, Mariah Boudreau, Bradform Demarest, Parisa Suchdev, Justin Paluba, Carter Ward, Will Thompson, Erik Weis, Shun Zhang, Kathryn Cramer, Sam Rosenblatt, Avi Chawla, and especially Jonathan St-Onge for the vibrant community and conversations that have inspired so many ideas.

There are many prior mentors to thank, for their encouragement and their ideas that have continually reshaped how I see the world. To Mark Graham, for sharing his love of the humanities, polemics, and critical thinking. To Lisa Schnell, whose kindness first made me feel at home at UVM and who often reminded me that the pursuit of knowledge never ends. To Matthew White, my undergraduate thesis adviser, for his warm advice and for a host of opportunities and unforgettable experiences.

To Zach Miller for sharing his infectious curiosity and many adventures, whether we made it to a trailhead or not. To Adam Pritchard for his authenticity and friendship; a trip home can never be complete without a visit. To Steven Van Wyck for many enjoyable mountaintop conversations. To Xiaozhi Zhang for sharing so many memorable moments, though life's ups and downs. To Bin Du and Xixi Liang for years of camaraderie and a furry addition to our family. To so many friends from VPOP, who continually inspire me with their dedication to improving the world, one conversation at a time.

Finally, I'm grateful to the UVM Computing staff past and present, including Mike Austin, Jim Lawson, Sean Blackerby, Rachel Dotey, and Katy Czar who have worked so diligently to enable research. To Andrea Elledge for years of work behind the scenes, making everything run smoothly. To Melissa Parr for years of assistance, advice, and fond memories. Thanks to the Graduate College, MassMutual, and the Vermont Advanced Computing Center for funding and support for my research.

Table of Contents

Citations ii
Acknowledgements iii
List of Figures xvi
List of Tables xvii

1 Introduction 1
1.1 Background 1
1.2 Data Infrastructure 4
1.3 Outline 8

2 Hurricanes and hashtags: Characterizing online collective attention for natural disasters 11
2.1 Abstract 11
2.2 Introduction 12
2.3 Materials and methods 15
2.3.1 n-gram usage rates 15
2.3.2 Deaths, damages, and locations 16
2.4 Results 17
2.4.1 Hurricane Attention Maps 17
2.4.2 Hurricane Attention Comparison 19
2.4.3 Attention and Impact Regressions by Category 23
2.4.4 Regression Models for Impacts, Impact Interactions and Hurricane Category 28
2.5 Concluding Remarks 31

3 Curating corpora with classifiers: A case study of clean energy sentiment online 35
3.1 Abstract 35
3.2 Introduction 36
3.3 Materials and methods 40
3.3.1 Description of data sets 40
3.3.2 Sentence embeddings 41
3.3.3 Relevance classification 41
3.4 Results 43
3.4.1 Interpretations of sentence embeddings 43
3.4.2 Ambient time series plots 45
3.4.3 Lexical calculus: Word shift plots 45
3.4.4 Allotaxonometry 46
3.4.5 Solar Energy Case Study 47
3.4.6 Wind Energy Case Study 49
3.4.7 Nuclear Energy Case Study 51

3.5 Concluding remarks 53

4 Selected contributions to published work 59

4.1 Collective Attention and *n*-gram Usage Rate Studies 60

4.1.1 Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter 61

4.1.2 How the worlds collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter . 64

4.1.3 Divergent modes of online collective attention to the COVID-19 pandemic are associated with future caseload variance 68

4.1.4 Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter 69

4.1.5 Twitter misogyny associated with Hillary Clinton increased throughout the 2016 U.S. election campaign 71

4.1.6 Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop. 74

4.1.7 Say their names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd . 75

4.1.8 The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020 76

4.2 Location-based Social Media Studies 78

4.2.1 Expecting the Unexpected: Predicting Panic Attacks from Mood and Twitter 78

4.2.2 The sleep loss insult of Spring Daylight Savings in the US is observable in Twitter activity 79

4.2.3 An assessment of measuring local levels of homelessness through proxy social media signals 81

4.3 Novel Methods for Social Media Data 82

4.3.1 The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series 83

4.3.2 Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning 84

5 Conclusions 88

5.1 Collective Attention Prediction 88

5.2 Sub-population Culturomics 89

5.3 Opinion Polling 90

5.4 GPU Accelerated NLP 91

5.5 Past, Present, and Future Hedonometer 91

5.6 Tweets with Locations 92

5.7 Ousiometric Lexicon Generation 93

5.8 BERTopic Weighting Scheme 93
5.9 Additional Corpora 95

6 Supporting information for Hurricanes and Hashtags 97
6.1 Summary Tables for Regressions 97
6.2 2-gram Attention Proportion of “`hurricane`” Usage Rate 98
6.3 Bi-exponential Decays 99
6.4 Hurricane Attention Maps 105

List of Figures

2.1 **Hashtag attention map and usage rate time series** for 1-grams matching the case-insensitive pattern “#hurricane*” for all four hurricanes reaching at least category 4 in the 2017 hurricane season. Markers along the hurricane trajectory indicate the National Oceanic and Atmospheric Administration (NOAA) reported position for every day at noon UTC. On the map, the smoothed rate of hashtag usage is wrapped in an envelope around the hurricane trajectory in panel A, showing the spatial dependence of attention on Twitter. In the lower two plots, panels B and C, we show the usage rates for hashtags and 2-grams matching `hurricane*` in English language tweets on linear and logarithmic scales. Usage rates within all tweets are indicated with a solid line, while usage rates in ‘organic’ tweets (tweets that are not retweets), are represented by a dashed line. The day of maximum attention on Twitter is marked with a star or a diamond for hashtags or 2-grams, respectively. Generally, hurricanes making landfall on the continental United States received greater attention than those not making landfall. The hashtag usage rate for Hurricanes Harvey and Irma at their maximum were approximately an order of magnitude larger than the maximum hashtag usage corresponding to hurricane Maria, and two orders of magnitude larger than Hurricane Jose. 18

2.2 **Radar plots comparing the eight most monetarily damaging hurricanes in the North Atlantic basin from 2009 to 2018.** For each plot, starting at the top position and rotating clockwise the measures are: the sum of usage rate of the hashtag, the number of days to reach 90% and 50% of the total attention received during that season, the total cost in dollars attributed to damage caused by the hurricane (in its year), the number of deaths attributed to the hurricane, and maximum usage rate of the hashtag during the year of interest. All measurements are normalized to the maximum value achieved by any hurricane. Hurricane Harvey was the most talked about hurricane, as well as the most damaging. Hurricane Irma was the most talked about on any single day. Hurricane Maria caused the most deaths, and had the longest attention half-life of all measured hurricanes. Raw values for this figure are shown in Section 6.1. Hashtag usage rate spark lines above each radar plot are normalized to show the common decay shape, and can not be compared to evaluate *relative* volume, and are shown on a log scale. 21

2.3 **Scatter plots for integrated hashtag usage rate versus the deaths and damages caused by each storm.** There is a clear positive association between the total attention represented by hashtags and the impacts of these storms. We reported Spearman's rho, ρ_s, in the top left corner of each plot. While for some categories, there is little evidence for a positive association, for the entire dataset $\rho_s \sim 0.54$. We perform a Bayesian linear regression for each category storm between the $\log I$ and log impacts. We show the mean model, along with the credible interval within a standard deviation of the mean model. We use hybrid axis with logarithmic scaling for most horizontal and vertical values and linear scaling near zero, in order to show storms that caused zero deaths or damages, as well as storms for which we measured a hashtag usage rate of zero. Changes in axis scaling occur at the blue dashed lines. Generally, more powerful storms received more attention, higher category storms received more attention even when causing minimal damage, and high category storms had a higher regression slope. These results suggest that for powerful storms, a given increase in impact was associated with a larger increase in attention. While for category one storms a 10-fold increase in deaths is associated with a two-fold increase in attention, for category five hurricanes, this same 10-fold increase in attention is associated with a 27-fold increase in attention. 24

2.4 **Posterior distributions of regression parameters** for the model $\log_{10} I \sim a_0 + a_1 X_i$,where X_i is either the log number of deaths (A and C) or log damages in dollars associated with the storm (B and D), and $\log_{10} I$ is the log integrated hashtag usage rate. The trend in regression coefficients for association between the log attention and log deaths suggests that higher category storms receive more attention per unit impact, while the trend of intercepts shows increasing baseline attention for a hypothetical minimally disruptive storm causing exactly \$1 in damages or one death. For regression coefficients relating log attention to log damages, Category 4 and 5 storms receive more attention per unit increase in log damages than lower category storms. However, the coefficients are smaller in magnitude due to damages varying across 7 orders of magnitude, as compared to deaths varying over 4 orders of magnitude. There is a larger uncertainty for the category 5 intercept values, as only 6 storms of this intensity formed between 2009 and 2019 in the Atlantic basin. At the right of each plot, we show the coefficients for the model fit for all hurricanes (blue violin), excluding tropical storms. Above each category, we show the value of the mean posterior distribution for each parameter. For a table of mean parameter values, see Table 6.1. 27

2.5 **Parameter distributions for models 1, 2 and 3**. Plots A–C show posterior distributions for regression 1, plots D–G show distributions for regression 2, which includes the addition of an interaction term, and plots H–O showing distribution for regression 3, which includes indicators variables for hurricane categories two through five. The addition of the interaction term, $a_{d,D}$ increases posterior variance for a_{deaths} as well as reducing its mean from $a_{\text{deaths}} = 0.49$ in regression 1 to $a_{\text{deaths}} = 0.05$ in regression 2 and $a_{\text{deaths}} = 0.12$ in regression 3, suggesting that while the number of deaths is associated with increased attention, attention response is primed by destruction. Additionally, the hurricane category indicator variables in regression 3 show the progressive increase in attention given to higher category storms compared to category 1 hurricanes. 32

3.1 **Embedded tweet distribution plot for the combined datasets.** Using a pre-trained model for semantically meaningful sentence embeddings based on MPNet, we plot the distribution of tweets within this semantic space. In both plots, points are tweets projected into 2D using UMAP for dimensionality reduction [108]. In panel A, we perform density based, hierarchical clustering using HDBSCAN and color by cluster. In panel B, we color by both the keyword used to query and the classification as relevant or non-relevant to the topic of clean energy. Relevant tweets containing the keywords ‘`wind`’, ‘`solar`’, and, to a lesser extent, ‘`nuclear`’ are relatively close together on the right in the embeddings, while non-relevant tweets are more dispersed. 42

3.2 **Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, containing the keyword ‘`solar`’.** In the top panel, we show the number of tokens with LabMT [42] sentiment scores in each corpus on each day. ‘Relevant’ tweets, in blue, have more scored tokens early on, but the number tokens in ‘non-relevant’ tweets increase in relative proportion over time. The center panel shows the average sentiment for each corpus, including a measurement of English language tweets as a whole in gray for comparison. Before 2019, the measured sentiment for both corpora are comparable, but subsequently the mean sentiment of ‘non-relevant’ tweets drops. In the bottom panel we plot the standard deviation of the sentiment measurement, which captures a broader distribution of sentiment scores for ‘non-relevant’ tweets. Without classification filtering, the ambient sentiment measurement would be entirely misleading, appearing as though the sentiment contained in tweets containing the word ‘`solar`’ dropped dramatically in 2019, when in fact sentiment has only modestly declined. 48

3.3 **Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, all containing the keyword ‘wind’.** In the top panel, we show the number of tokens with LabMT sentiment scores for each corpus during each two week period [42]. R tweets, in blue, have more than an order of magnitude fewer tokens per time window over the entire study period. The center panel shows the average sentiment for each corpus, including measurement of English language tweets as a whole in gray for comparison. R ‘wind’ tweets are more positive than Twitter on average early on, but this difference is reduced over time. Because most ‘wind’ tweets are non-relevant, sentiment of the combined corpus closely follows the NR sentiment. In the bottom panel we plot the standard deviation of the sentiment measurement, which captures a broader distribution of sentiment scores for ‘non-relevant’ tweets, as was the case for all case-studies we examined. Without classification filtering, the ambient sentiment measurement would have been dominated by NR tweets. 50

3.4 **Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, all containing the keyword ‘nuclear’.** In the top panel, we show the number of tokens with LabMT [42] sentiment scores for each corpus in each two week period. The number of relevant n-grams, in blue, is consistently lower than non-relevant n-grams. The center panel shows the average sentiment for each corpus, including measurement of English language tweets as a whole in gray. We found that R tweets had higher sentiment than NR tweets containing ‘nuclear’, but had much lower sentiment than Twitter as a whole. Sentiment appears relatively stable for both corpora with periods of higher sentiment around 2017 and 2020-2022 for the R corpus. In the bottom panel, we plot the standard deviation of the sentiment measurement, which shows a broader distribution of sentiment scores for NR tweets, as well as sentiment for both corpora trending down slightly. 52

3.5 **Sentiment shift plots comparing the classified relevant (R) and non-relevant (NR) tweet corpora for tweets containing the keywords '`solar`', '`wind`', and '`nuclear`'.** We show the top 20 words contributing to the difference in LabMT sentiment between the corpora. **A.** Relevant tweets that are related to clean energy are more positive on average for all keywords when compared to non-relevant tweets. Sad words that are less common in relevant '`solar`' tweets are 'radiation', 'pressure', and 'humidity', which largely refer to the weather. Happy words like 'energy' and 'power' are more common in relevant tweets compared to tweets non-relevant to solar energy. **B.** For '`wind`', relatively sad terms like 'humidity' and 'pressure' are less common in relevant tweets (these appear in clearly non-related tweets about the weather), while happy terms like 'energy', 'power', and 'solar' are more common in tweets relevant to wind as a renewable energy source. **C.** For '`nuclear`', relevant tweets are on average more positive due to sad words like 'war', 'weapons', and 'bomb' being less common in relevant tweets, while happy words like 'power' and 'energy' are more common. The two prominent sad words 'nuclear' and 'waste' go against the positive difference in moving from non-relevant to relevant tweets as they both occur more frequently in relevant tweets. 54

3.6 **Allotaxonograph comparing the rank divergence of words classified as relevant to solar energy discourse to those containing the keyword '`solar`' but classified as non-relevant.** On the main 2D rank-rank histogram panel, words appearing on the right have a higher rank in the 'relevant' subset than in 'non-relevant', while phrases on the left appeared more frequently in the 'non-relevant' tweets. The panel on the right shows the words which contribute most to the rank divergence between each corpus. We observe that many words associated with weather bots, such as 'mph,' 'uv,' and 'pressure,' are more frequently used in non-relevant posts, while words like 'panels,' 'energy,' and 'power,' used more in tweets relevant to solar energy. Notably, commonly used function words, such as 'the,' 'and,' and 'are,' are off-center in the rank-rank histogram, a further indication that many of the 'non-relevant' tweets are from automated accounts publishing weather data rather than using conversational English. The balance of the words in these two subsets is noted in the bottom right corner of the histogram, showing the percentage of total counts, all words, and exclusive words. For this example the two subsets are nearly balanced, indicating that the filtered corpus contains less than 50% of word tokens from the raw query. See Dodds *et al.* [45] for a full description of the allotaxonometric instrument. 55

4.1 An example document showing a ‘storyon’ counter objected represented within the database. For each language, we store a collection of n-gram counters. By querying for `word` we can assemble word usage rate time series. Querying for `time`, we can assemble a daily Zipf distribution. More complex queries are also enabled; future studies could query by `rank` to study the emergence of slang or other new types entering a language. 62

4.2 Reprint of Figure 1 from [8], with caption as follows: “For each n-gram, we display daily rank in gray overlaid by a centered monthly rolling average (colored lines), and highlight the n-gram’s overall highest rank with a solid disk. **A.** Anticipation and memory of calendar years for all of Twitter. **B.** Annual and periodic events: Christmas in English (blue), Easter in Italian (orange), election in Portuguese (green), and summer in Swedish (red). **C.** Attention around international sports in English: Olympics (blue), FIFA world cup (orange), and Super Bowl (red). **D.** Major scientific discoveries and technological innovations in English. **E.** Three famous individuals in relevant languages: Ronaldo (Portuguese), Trump (English), and Pope Francis (Italian). **F.** Major infectious disease outbreaks. **G.** Conflicts: Gaza in Arabic (blue), Libya in French (orange), Syria in Turkish (green), and Russia in Ukrainian (red). **H.** Protest and movements: Arab Spring (Arabic word for ‘revolution’, blue), Occupy movement (English, orange), Brexit campaign (English, green), #MeToo movement (English, brown), and Black Lives Matter protests (English, red).” 63

4.3 Screenshot of the realtime, 15 minute resolution n-gram viewer available on the storywrangler website. The database is no longer realtime, due to the end of our data sharing agreement with Twitter, but the final two weeks of 15 minute resolution data, from May 20th, 2023 to June 1st, 2023, remains publicly available. 64

4.4 Reprint of Figure 6 from [9], with caption as follows: “Time series for daily reported case loads and death compared with a list of 10 salient 1-grams for the top language spoken in each country. For each n-gram, we display a weekly rolling average of usage ranks at the day scale in gray overlaid by an average of all these 1-grams in black marking their corresponding ranks using the left vertical axis. Similarly, we use the right vertical axis to display a weekly rolling average of daily new cases (red solid-line), and reported new deaths (orange dashed-line).” 66

4.5 Reprint of Figure 2 from [9], with caption as follows: "Top 20 (of 1,000) 1-grams for our top 12 languages for the first three weeks of April 2020 relative to a year earlier. Our intent is to capture 1-grams that are topically and culturally important during the COVID-19 pandemic. While overall, we see pandemic-related words dominate the lists across languages, we also find considerable specific variation. Words for virus, quarantine, protective equipment, and testing show different orderings (note that we do not employ stemming). Unrelated 1-grams but important to the time of April 2020 are in evidence; the balance of these are important for our understanding of how much the pandemic is being talked about." 67

4.6 Figure 2 from [36], with caption as follows: "We display the mean normalized log rank timeseries of the top 20 words closest to each of $E[C_1]$ and $E[C_2]$ in dashed curves and the single word closest to each of $E[C_1]$ and $E[C_2]$ in thin solid curves for each of the first 12 of 24 languages. The divergent modes of dynamic behavior are consistent across most languages, with some languages (English, French, German, and Indonesian) displaying prominently larger peaks in words closest to $E[C_2]$ during late January through early February 2020. Other languages, such as Korean and Tagalog, do not display this behavior." 69

4.7 Figure from preliminary class project, showing average happiness values of reply threads averaged by user and represented with markers related to account type. We found that tweets with a lower ratio of replies to likes were more likely to have higher happiness scores as measured by the LabMT sentiment lexicon. 71

4.8 Reprint of Figure 2 from [115], with caption as follows: Ternary histograms and $N_{retweets}/N_{replies}$ ratio time series for the **@BarackObama** (A–D) and **@realDonaldTrump** (E–H) Twitter accounts. The ternary histograms (A–C and E–H) represent the count of retweet, favorite, and reply activities normalized by the sum of all activities. White regions indicate no observations over the given time period. See Fig 4 for examples of full time series for response activity for example tweets. Heatmap time series (D and H) consist of monthly bins representing the density of tweets with a given ratio value. Single observations (bin counts <2) are represented by grey points. The two dates annotated correspond to the date of Trump's declaration of candidacy (2015–05–16) and the 2016 general election (2016–11–09). We show the tendency for Trump tweets to have ternary ratio values with a greater reply component—with pre-candidacy tweets having higher variability and pre-election tweets having a higher $N_{retweets}/N_{replies}$ ratio value. Post-election Obama tweets have ternary ratio values with more likes than other periods for both Obama and Trump. 72

4.9 State sentiment maps for two selected days in the study period. May 23, 2022 was the day prior to the Uvalde, TX school shooting, and May 25, the day following the shooting. 80

4.10 Measures of collective attention and sentiment for US tweets containing homelessness from the paper. 83
4.11 Reprint of Figure 3 from [37], with capition as follows: A comparison between the standard discrete wavelet transform (DWT) and our discrete shocklet transform (DST) of a sociotechnical time series. Panel (B) displays the daily time series of the rank of the word "trump" on Twitter. As a comparison with the DST, we computed the DWT of using the Ricker wavelet and display it in panel (A). Panel (C) shows the DST of the time series using a symmetric power shock. 85
4.12 Screenshot of one of the paper's associated interactive online supplementary figure showing ranked spike indicators. 86

5.1 Multi-language comparison of collective attention for the 1-gram "Russia", translated into 20 languages. Notably, all languages see a spike in mentions corresponding to the invasion of Ukraine in Feburary 2022, but Ukranian and Russian see the smallest increase. 90
5.2 Spatially aggregated counts of tweets containing anchors, '`Farm`' and '`Traffic`'. While both counts are highly correlated with population density, '`Traffic`' counts seem to be more intensely peaked in urban areas, while '`Farm`' counts appear to be more broadly distributed. There seems to be potential for further studies of spatial variation of language usage. 94

6.1 Word usage rate proportions of "`hurricane *`" in English tweets 106
6.2 Attention proportions of "`Huracàn *`" in Spanish. We can see that the word usage rate surrounding "`Hurricane Maria`" captures a similar amount of the total attention for the 1-gram hurricane as "`Huracàn Marìa`" captures. Additionally, hurricane Harvey's 2-gram usage rate is lower in Spanish than in English, while Hurricane Katrina is talked about considerably in English but does not rise about the 50000th most used 2-gram in Spanish. As always, usage rates are case-insensitive. 106
6.3 **Bi-exponential Hurricane decay half-lives:** Distributions of fitted half-lifes for the populations u and v. The mean half-lives for $\tau_1 = 1.3$ days and $\tau_2 = 156$ days for hashtags and $\tau_1 = 1.1$ days and $\tau_2 = 241$ days for 2-grams. For τ_2 the median half-lives are also interesting since we suspect the longest half-lives are due to poor fits. For hashtags $\tau_2 = 23$ days, and for 2-grams $\tau_2 = 24$ days. 107

6.4 **Simple Exponential Hurricane decay half-lives:** Distributions of fitted half-lives for a single population. The median half-lives for $\tau = 5.3$ days a for hashtags and $\tau = 5.2$ days for 2-grams. The simple exponential model fails to explain the break in attention decay for larger storms, receiving more attention. The bi-modal distribution of half-lives for 2-grams suggests that there are two categories of storms, ones with larger half-lives have more data, and thus the longer decay increases the fitted half-life. Meanwhile, smaller storms receive so little attention, that we don't measure any after a week or so, leading to a much smaller half-live, which corresponds to τ_1 in our bi-exponential fit. 107
6.5 **Decay Model Comparision:** Distributions of Mean Squared Error (MSE). The bi-exponential model has the lowest average MSE, followed by the simple exponential decay. The power law decay fails to capture the dynamics of attention decay, when the fit is compared to the data visually, and is reflected in the higher average MSE. 108
6.6 Hurricane bi-exponential decay fits for hashtag usage rates and 2-gram usage rates for "`hurricane *`" 109
6.7 Hurricane decays fits for all hurricanes for which we have at least 10 days of 2-gram usage rate data. Fits are performed for the function $y = \frac{N}{p+r-q}[(p-q)e^{-(p+r)t} + re^{-qt}]$, a simple two population decay model as proposed by Candia et al. [23]. Here p would be interpreted as rate of decay from population 1, r would be the transfer rate from population 1 to population 2, and r would be the rate of decay from population 2. Population 1 might be thought of as bystandards with a shorter attention span, while population two are those living with the ramifications, or working on the recovery who lose attention more slowly. Reported on the graph are the half lives associated with fitting this model for both the hashtag usage rate and 2-gram usage rate, $\tau_1 = \frac{\ln 2}{p+r}$ and $\tau_2 = \frac{\ln 2}{q}$ 110
6.8 Hurricane Attention Map and time series for 2009 111
6.9 Hurricane Attention Map and time series for 2010 112
6.10 Hurricane Attention Map and time series for 2011 113
6.11 Hurricane Attention Map and time series for 2012 114
6.12 Hurricane Attention Map and time series for 2015 115
6.13 Hurricane Attention Map and time series for 2016 116
6.14 Hurricane Attention Map and time series Map and time series for 2018 . . . 117

List of Tables

3.1 **Paraphrased example tweets for relevant (R) and non-relevant (NR) examples in each case study.** 57

3.2 **Summary statistics and model performance for each of the three case studies.** 58

6.1 Mean Regression Parameters fit for storms of each category. See Fig. 2.4 for full parameter distributions. 97

6.2 Priors for Regression 1: Linear in deaths and damages 98

6.3 Results for Regression 1: Linear in deaths and damages 98

6.4 Priors for Regression 2: Additional interaction term 98

6.5 Results for Regression 2: Additional interaction term 98

6.6 Priors for Regression 3: Additional categorical term for hurricane category . 99

6.7 Results for Regression 3: Additional categorical term for hurricane category 99

6.8 The unnormalized values associated with radar plots in Section 2.4, sorted by estimated damage. 100

6.9 Fitted half-lives τ_1 and τ_2 for all storms with at least 10 days of hashtag usage.103

6.10 Fitted half-lives τ_1 and τ_2 for all storms with at least 10 days of 2-gram usage.104

Chapter 1

Introduction

1.1 Background

Human behavior is being digitally observed and continuously recorded at a previously unfathomable rate. As the world continues transitioning into the internet age, researchers studying social systems have leveraged the data richness and complexity offered by publicly available user generated content. But the research value of social media data in particular has yet to be fully realized for legal, economic, and ethical reasons.

Historically, written corpora distilled from large collections of books, newspapers, academic articles, etc., reflected the perspective of a few wealthy and non-representative authors, and were edited to conform to linguistic norms. During the last 20 years, the growing adoption of social media facilitated by mobile devices has brought forth a golden era for Computational Social Science [89]. Indeed, never before in human history has there been such a detailed record of everyday speech and attention, notably offering indications of consumption and popularity among average people.

For social science research, Twitter data in particular offers some notable advantages over other large-scale texts. Twitter enables studies with high temporal resolution, not only because tweets contain timestamps accurate to the second, but because the potential time lag between events occurring in the world and citizens publishing their thoughts can be under a minute. No other readily available platform has such a high resolution timescale for activity, such that surprisingly small slices of time contain word distributions with measurable associations to events in the offline world. Finally and importantly, Twitter encodes popularity through retweets, a critical signal for characterizing cultural amplification that is missing from many other text data sources [110, 126].

So what can we do with one hundred billion tweets? Researchers have (almost) as many ideas; as of February 2023 over eight million studies of Twitter data have found their way to Google Scholar. Within our modest contribution to this body of work, we transform unstructured text into a few standardized data products, measurements, and observations about the world. Our main aim is to render important cultural phenomenon in the social world more quantifiable, not unlike efforts to construct instruments such as the microscope and the telescope transformed our understanding of the natural and physical world.

Put into practice, first we tokenize human expressions found in text, breaking up sentences into words and phrases, referred to as n-grams, and counting the number of occurrences of each. For any period in time, we can compute a usage rate and rank of n-grams. Then, to study the dynamics of collective attention, we compute n-gram usage timeseries [8] to characterize changes in the popularity of words and phrases over time.

We're often interested in more than raw quantities of attention. Using the tools described in this thesis, we can measure the sentiment of a text, either using dictionaries such as labMT, or more black-box machine learning models [98, 129]. We often find there is value in the interpretability of dictionary-based methods, which allow us to compute word level contributions to differences in sentiment between corpora [55]. Comparing distributions of text using tools like rank-turbulence divergence, we uncover changes in word usage at all scales, from changes in frequently used function words to the rarest hashtags [45].

The computational methods that enable work with text as data, inspired by existing tools often used in statistics and physics, have also recently improved in quality and speed [167]. Tokenization, the process of breaking up unstructured text into discrete tokens that can be measured, has accelerated rapidly through the use of Graphics Processing Units (GPUs). Advanced tokenizers, such as WordPiece, have improved the effective vocabulary size of models by breaking up unknown words into substrings that can still be represented in a semantic space [175].

Within the past decade, text representations have improved dramatically, enabling a range of natural language processing (NLP) classification tasks. In the 2010s, global word embeddings like word2vec, gloVe, and FastText emerged to represent words in semantically meaningful vector spaces [81,112,127]. Contextual embeddings based on transformers, such as BERT, enabled distinct embeddings for words with multiple meanings that improve performance on downstream tasks [35,156]. These contextual word embeddings were precursors to today's Large Language Models (LLMs) like ChatGPT.

The new data availability and rapid innovation in NLP methods has attracted a broad range of interdisciplinary interest [109]. Within economics for example, there's an interest in how narrative shapes the economic decisions of groups, leading to irrational exuberance or sustained collective pessimism [142]. Communications researchers studying social movements such as the Arab Spring, #MeToo, and Black Lives Matter have used tweets to explore and quantify digital activism and protest activity [78,102,173].

Prior work has explored the potential of social media to supplement gold standard representative surveys of public opinion [121]. Measuring text in tweets containing the keyword '`Climate`', researchers found polarizing words [29,30].

Having described the background areas associated with the present thesis, we now move into discussion of the technical challenges posed by doing serious computational work at large-scale.

1.2 Data Infrastructure

A number of conceptual and technical challenges stand in the way of unlocking the potential for using social media data to do sound computational social science. Conceptually, researchers need to select a defensibly relevant corpus from the universe of tweets, and make a measurement on the text or metadata contained within the corpus that reflects real-world activity. Enabling these two steps requires solving a number of technical challenges and evaluating each alternative. Qualities of scientific interest include:

- the relevance of a given corpus to the field of interest,
- the quality of the corpus filtering, to increase relevance of the selected content,
- the success rate in sorting algorithmically generated content from human generated content (and hybrids in between), and
- practical considerations such as execution time, cost, and accessibility.

While each of the above issues could merit its own PhD thesis, in the following paragraphs, I address some challenges I overcame in the design of the computational pipeline before turning my attention to some of the former issues in the bulk of the thesis.

Minimizing query execution time is an important practical and technical consideration given the massive size of the Twitter dataset. Response times for human/machine interactions have been studied for decades to better understand the impacts on usability [58, 140]. To enable interactive image view, query times should not exceed roughly two seconds, around which user satisfaction switches from positive to negative for image loading [50]. For collaborative exploratory work, keeping the response time under two minutes is a good approximate target; too much slower and researchers' ability to iterate is diminished. Finally, for a full analysis, being able to execute queries within around eight hours enables running code overnight. These execution time thresholds are helpful in considering what kinds of research behavior is possible at different latencies.

Our database infrastructure attempts to enable queries at each of these scales. Storywrangler, our research group's web-based interface to explore n-gram timeseries, indexes millions of tokens daily for nearly 15 years. It runs on a back-end capable of executing single timeseries queries within tens of milliseconds, enabling a smooth interactive user experience.

DataMountain, the 64TB RAM database cluster I designed with UVM's Enterprise Technology Services to store tweets, enables query execution at both (a) the two-minute response scale and (b) the eight hours response scale, by storing multiple corpora at different sampling resolutions. We curate the following collections for the Computational Story Lab:

- `decahose` - a roughly 10% sample of all public Twitter messages between 2008 and 2023, containing 110B tweets,
- `gardenhose` - a 1% sample of Twitter, and
- `driphose` - a 0.1% sample of Twitter each derived from the decahose.

Having separate collections allows researchers to rapidly prototype with a small sample of tweets that can be returned within a few minutes. It also enables seamlessly scaling up the sample size to reduce variance.

For many kinds of queries, databases can reduce the number of required read operations by orders of magnitude. Keyword queries benefit from text indexes, especially for infrequently used terms. For example, the term '`Homelessness`' is used roughly two times per every million words on Twitter. Prior to DataMountain, to search for tweets containing this keyword in our collection, we'd need to scan 150,000 times more tweets to find matches. This query could take hours at best, but was highly variable depending on shared compute resource allocation, and could take days at worst. With DataMountain, a search for this keyword takes under 20 seconds without caching, and a mere 700 milliseconds when cached.

In designing any major hardware system, there are trade-offs between costs and benefits.

For DataMountain, a number of compromises were made to fit the specific constraints of the University research funding environment. Money for high-performance computing is typically awarded in the form of hardware or cloud computing grants. The low upfront costs of cloud computing, which is appealing to businesses with long-term cash-flow, are less attractive for a University due to the lower overall long-term costs of on-premises hardware. We made a decision to trade latency for capacity by substituting some random access memory (RAM) for 3D-XPoint persistent memory (PMEM). PMEM technology provides significantly more capacity per dollar than RAM at the cost of increased latency compared to RAM, but was still a vast improvement over reading from disk.

In building our DataMountain server, we made some non-standard design decisions around redundancy and storage capacity. Within industry, where database uptime is mission-critical, and any interruption in service both halts revenue streams and potentially damages trust in a company's brand long term, redundancy is built into database systems to ensure nearly 100% uptime. However, high availability has a cost, with production systems keeping at least three separate database replica sets.

For a University with limited funding, tripling the cost of the machine in order to enable replication within the system was simply infeasible. Fortunately, for a research cluster on campus, the risk associated with downtime is smaller. If a disk or memory module fails on a research cluster, scientists can simply wait a few days for replacement parts or for data to insert and index. We judged that a system with three times the memory capacity, but only 90% the uptime, would be more valuable to researchers than a smaller memory system with high availability. Additionally, we planned for all data to be backed-up on flat files outside of the database cluster, so risk of data loss remains minimal.

Perhaps the most important criteria is accessibility. Without designing for user experience, resources will be underutilized, negatively impacting the research potential of the investment. For example, if a software package is restricted to on-campus users and takes an hour to install, and even longer to learn to use, it will necessarily only be seen by a small handful of participants. To address these challenges, our team secured a large grant from the National Science Foundation to broaden access to the DataMountain infrastructure. The Science of Online Corpora, Knowledge, and Stories (SOCKS) project, a five year $20M Experimental Program to Stimulate Competitive Research (EPSCoR) grant from NSF, will support development work aimed at giving academic disciplines traditionally underrepresented in computing access to hi-performance tools to answer questions, e.g., in Political Science, Cultural Anthropology, Psychology, Public Health, and many other fields.

To enable standardized, repeatable text measurements in service to hypothesis testing in these domains, we deploy a host of lexical instruments. Our research group, the Computational Story Lab, has built up a wealth of tools over the years, such as sentiment dictionaries like labMT [86], word-shift graphs to compare texts [55], allotaxonographs to measure rank-turbulence divergences between texts [45], and contagiograms measuring dynamics of social amplification [10] to name a few. Ensuring that these tools are maintained and packaged to enable rapid exploration of text datasets is a long-term goal.

Concurrently, we have to be aware of and up front about the limitations of these tools to answer questions about the world. While a collection of 100 billion tweets is almost incomprehensibly large, there are still many topics of potential interest that receive relatively little attention online. The word usage frequencies are heavy-tailed, and this limits our ability to reduce the variance of measurements. Indeed, roughly half of the hashtags seen each day have never been seen before, making them challenging to index. This problem is further compounded when we're interested in high resolution time-series, or fine-grained spatial structures.

For research topics where the population of interest is not 'people who tweet', but rather the speakers of a language or likely voters in an election, we have to heavily discount the usefulness of Twitter data. Twitter's users are not a representative sample of any other populations [116]. Changing demographics of users and the creation of algorithmically voiced accounts should make us wary of claims that trends between these two distinct populations would be correlated [15, 16]. As response rates for traditional polling have declined, researchers have investigated using non-representative samples by using respondents' demographic information to reweigh responses using multilevel regression and poststratification [159]. Each chapter in the thesis addresses the representativeness issue in bespoke fashion.

While the systems engineering to enable research at scale has been a substantial component of my work, the driving motivation is to enable interdisciplinary experimentation, powered by new, interpretable measurements. The collaborative process of instrument building provides feedback both to improve next iteration of measurement tools and to shape a new generation of research questions.

1.3 Outline

With the computational infrastructure in place, we now describe several specific applications of our work toward understanding real-world phenomenon through our lexical lenses.

In Chapter 2, we demonstrate how social media n-gram usage rate timeseries can be used to study collective attention paid towards hurricanes [14]. We measure associations between our collective attention proxies and hurricane impacts, like estimated damages and deaths. We also validate our proxy of choice, hashtags following the form, '`#hurricane*`', by replicating our regression analysis on attention proxies based on 2-gram usage rate timeseries.

For case studies examining topics blessed with unambiguous search terms, such as a consistently used hashtag or a name without polysemy, a simple keyword search could be sufficient to curate a corpus of tweets. Unfortunately, there are often no keywords capable of filtering for selecting relevant tweets. Looking to examine sentiment related to car brands for example, we found tweets containing '`Ford`' were a mix of topics from the American auto-maker to allegations against Supreme Court nominee Brett Kavanaugh. Studying attention paid to top causes of mortality, tweets containing '`Cancer`' were dominated by horoscope readings. Framing these questions as a supervised classification task requires researchers to label training data, but improves our understanding of corpus quality by enabling estimates of precision and recall. In Chapter 3, we propose addressing this corpus curation problem with contextual sentence embedding powered text classifiers [15], and use a case-study related to clean energy policy to demonstrate the value and performance of our approach.

In Chapter 4, I discuss my individual contributions to a subset of the 23 scientific manuscripts that I have had the opportunity to be a co-author on during my PhD. Storywrangler, on which many later projects rely, is an instrument we created to parse, store, and serve word (and higher n-gram) usage timeseries, which capture changes in language usage due to human attention and algorithmic influences on Twitter [8]. Along the way, I helped establish incremental results as we worked towards a final version of Storywrangler. For example, we identified inconsistencies with Twitter's langauage identification software that corrupted language based n-gram usage timeseries [40]. To address this corruption, we performed language classification on nearly 100TB of tweets and described the changing user base by language and sharing behavior [10].

We used Storywrangler data to study the dynamics of language around, and attention paid to, an emerging global pandemic [9]. More broadly applicable, we developed techniques to identify emergent words during periods of substantial language shifts. Further research along these lines used unsupervised timeseries clustering to identify distinct phases of language use during first months of the COVID-19 pandemic [36].

Attempting to quantify levels of lexical fame, we studied U.S. presidential candidates, and the K-pop sensation BTS [41]. We found that the 1-gram'`BTS`' was briefly used more frequently than the word '`the`'. Using the transition to Daylight Savings Time along with Twitter activity patterns, we estimated the delay in the onset of sleep across the United States. In the process, I built a user location database storing city and state labels for 5% of our tweet collection [96]. We explored the public's social media engagement with US presidents, finding increased controversy in the replies to President Donald Trump's tweets throughout his presidency as measured by the ratio of replies to retweets or favorites [115].

In the process of creating and exploring these new sociotechnical datasets, we often found that existing measurements were poorly suited to answering research questions of interest. We created measurements to detect patterns in timeseries [37], to compare language distributions (or any heavy-tailed distribution) using rank-turbulence divergence [45] and probability-turbulence divergence [47]. While sentiment or valence is likely the most commonly measured dimension of meaning, other semantic differentials convey distinct meanings. We introduced ousiometrics to measure meaning within a novel power-danger framework [39].Recognizing the need to update semantic lexicons, as well as the expense associated with human ratings, we explored using word embeddings to score additional words to augment existing lexicons [11].

Collaborations with researchers outside of the core Storywrangler team were also fruitful. We created a lexicon to measure misogynistic language, and used Storywrangler's parser on a sample of tweets mentioning female politicians to track trends in hateful language [163]. We explored the viability of using mentions of homelessness on Twitter as proxy for state-level homelessness rates [17]. I also contributed state-level community sentiment measurements to a study on factors associated with the onset of panic attacks [106]. Finally, we measured the dynamics of attention paid to Black victims of fatal police violence on Twitter [173].

Following these three chapters, I conclude this dissertation with a reflection on outcomes, along with plans for future work.

Chapter 2

Hurricanes and hashtags: Characterizing online collective attention for natural disasters

2.1 Abstract

We study collective attention paid towards hurricanes through the lens of n-grams on Twitter, a social media platform with global reach. Using hurricane name mentions as a proxy for awareness, we find that the exogenous temporal dynamics are remarkably similar across storms, but that overall collective attention varies widely even among storms causing comparable deaths and damage. We construct 'hurricane attention maps' and observe that hurricanes causing deaths on (or economic damage to) the continental United States generate substantially more attention in English language tweets than those that do not. We find that a hurricane's Saffir-Simpson wind scale category assignment is strongly associated with the amount of attention it receives. Higher category storms receive higher proportional increases of attention per proportional increases in number of deaths or dollars of damage, than lower category storms. The most damaging and deadly storms of the 2010s, Hurricanes Harvey and Maria, generated the most attention and were remembered the longest, respec-

tively. On average, a category 5 storm receives 4.6 times more attention than a category 1 storm causing the same number of deaths and economic damage.

2.2 Introduction

The collective understanding and memory of historic events shapes the common world views of societies. In a narrative economy, attention is a finite resource generating intense competition [28, 52, 53, 76, 92, 120, 142, 143, 152]. As commerce and communication shift to online platforms, so too has the narrative economy moved to the digital realm. In 2018, over $100 billion dollars were spent on internet advertising in the United States, nearly overtaking the $110 billion spent on traditional media advertising—about 1% of the US GDP [77]. Today, social media both facilitates and records an extraordinary percentage of the world's public communication [118, 128]. For computational social scientists, the migration of parts of the narrative economy to the web continues to present an immense opportunity, as the discipline becomes data-rich [111, 125].

Academics have become interested in narrative spreading around newsworthy events on social media platforms such as Twitter, as increasingly political fights for influence or narrative control are fought by actors as wide ranging from activists and police departments [56], to state censors suppressing discourse internally and state supported troll factories spreading divisive narratives internationally [16, 20, 32, 69, 123, 150]. In 2019, the social media platform Twitter boasted over 145 million daily active users [135].

Quantifying the spread of narratives and the total attention commanded by them is a daunting task. Recent work has made progress in tracking the spread of quoted and modified phrases through the news cycle, and others have worked to identify actant-relationships and compile contextual story graphs from social media posts [92, 141]. In comparison, quantifying attention directed towards a topic, person or event is a somewhat easier task. Rather than identifying actors and identifying what they act on, as is the case for narrative attention, we can simply count mentions of an entity. Since increasing raw attention or number of mentions is often the zeroth order activity in public relations campaigns, quantifying the volume of attention, irrespective of the sentiment or narrative within which the attention is embedded, seems a natural first step [46].

An understanding of attention has typically focused on time dynamics as measured by the number of mentions in a given corpus, explaining either temporal decay of interest or heavy-tailed allocation of attention given to a spectrum of topics through some preferential attachment mechanism [23,49,63,72,73,101,155,157]. Another group of studies have worked to classify attention time series from social media as either exogenous or endogenous to the system, modeling the functional form of collective attention decay, or determining if spreading crosses a critical threshold [33, 83, 91, 172]. While these studies have typically focused on scientific works, patents, or cultural products such as movies, the rise of large social media datasets have enabled the investigation of a wider range of topics in online public discourse [87].

In this study we examine the collective attention focused on hurricanes, using Twitter, which allows us to capture more natural speech intended for human readers as opposed to search terms. Twitter data has been used to measure shifts in collective attention surrounding exogenous events like earthquakes by looking for jumps in the Jensen-Shannon divergence between tweet rate distributions between days, or creating real-time earthquake detection using keyword based methods [134, 137].

Here, we use collective attention in a more narrow sense. Instead of looking for anomalous tweet rates, we study n-gram usage rates for hashtags and 2-grams associated with individual events. Specifically, we examine the usage rates of hashtags and 2-grams matching the case-insensitive pattern “#hurricane*” and “hurricane *”, respectively. Natural disasters provide an ideal case study, since they are generally unexpected, producing the signature of an exogenous event. However, the volume of attention given to any particular hurricane varies widely across several orders of magnitude, as does the severity of the storm in terms of the lives lost and damages caused.

Prior efforts have examined the attention received by disasters by type and location, as measured by time devoted on American television news network coverage, and striking discrepancies: for example, to have the same estimated probability of news coverage as a disaster in Europe, a disaster in Africa would need to cause 45 times as many deaths [51]. The same study found that in order to receive equivalent coverage to a deadly volcano, a flood would need to cause 674 times as many deaths, a drought 2,395 times as many, and a famine 38,920 times as many casualties.

Strong hurricanes are more likely to capture attention than weak hurricanes, and hurricanes impacting the continental United States capture much more attention than those failing to make landfall. To what degree does attention shrink when hurricanes make landfall outside of the continental US? The 2017 hurricane season is a particularly stark example, showing that for comparably powerful storms above category 4, those projected to make landfall over the continental United States were talked about nearly an order of magnitude more than Hurricane Maria, which impacted Puerto Rico, and two orders of magnitude more than Hurricane Jose, which never made landfall.

Given the attention received by some hurricanes so unbalanced, we must ask the question: Do government or humanitarian relief resources get dispersed with greater generosity for storms that capture public attention, or are these organizations insulated from popular attention? For the 2017 hurricane season, more money was spent more quickly to aid the victims of hurricanes Harvey and Irma than victims of Hurricane Maria, contributing to the significantly higher death toll and adverse public health outcomes in Puerto Rico [168]. While the attention and policies of government agencies are not usually dictated from Twitter, public attention certainly has some effect on the focus of agencies and allocation of government resources, and recently more attention has been focused on understanding the discourse on social media before, during, and after natural disasters [2, 7, 31, 104, 119, 162]

We structure our paper as follows. In Section 2.4, we examine the spatial associations between hurricanes and the attention they receive, we compute and compare measures of total attention, maximum daily attention, and non-parametic measures of the rate of attention decay for the most damaging hurricanes in the past decade. We present conclusions in Section 2.5. Finally, we outline our methods and data sources, covering the collection of n-gram usage rate data in English tweets as well as data sources for hurricane locations and impacts.

2.3 Materials and methods

2.3.1 n-gram usage rates

We query the daily usage rate of hashtags referencing hurricanes are queried from a corpus of 1-gram—words or other single word-like constructs—usage rate time series, computed from approximately 10% of all posts ("tweets") from 2009 to 2019 collected from Twitter's "decahose" [93]. We define usage rate, f, as

$$f(t) = c_{\tau}(t) \Big/ \sum_{\tau' \in \mathcal{D}_t} c_{\tau'}(t),$$

with count, c_{τ}, of a particular 1-gram divided is by the count of all 1-grams occurring on a given day, $\mathcal{D}_t$. The usage rates are based only on the usage rate of 1-grams observed in tweets classified as English by FastText, a language classification tool [10, 81]. Our usage rate data set includes separate usage rates for 1-grams in "organic" tweets, tweets that are originally authored, as well as usage rates of 1-grams in all tweets (including retweets and quote tweets). More details about the parsing of the Twitter n-gram data set are available in [8].

For the purpose of studying attention, our usage rates are derived from the corpus with all tweets, including retweeted text, to better reflect not only the number of people tagging a storm, but also the number of people who decide the information contained therein was worth sharing.

We studied the usage rate of 1-grams exactly matching the form "`#hurricane*`", where $*$ represents a storm's name. We also measured the usage rate of 2-grams matching the pattern "`hurricane *`" for each storm name. All string matching is case-insensitive.

For the ten years covered by the HURDAT2 dataset (described in Section 2.3.2) overlapping with our Twitter dataset, there have been 75 storms reaching at least category 1 in the North Atlantic Basin. Within our 10% sample of tweets, we count over all storms a total of 1,824,842 hashtag usages within a year of each storm, and 3,643,411 instances of the matching 2-gram.

2.3.2 Deaths, damages, and locations

To augment our usage rate data set, we downloaded data associated with all hurricanes in the North Atlantic basin from 2008 to 2019 from Wikipedia [166]. Included in the Wikipedia data are the damage estimates (US$) and deaths caused by each storm, as well as the dates of activity and areas effected. We also used the HURDAT2 data set containing the positions and various meteorological attributes of all North Atlantic hurricanes from 1900 to 2018 for the spatial component of this work [164]. HURDAT2 is compiled by the National Hurricane Center including updated and revised data, which reflects the official record of each cyclone's history. For the time range overlapping with the Twitter derived data set, HURDAT2 has 3 hour resolution.

2.4 Results

2.4.1 Hurricane Attention Maps

In Fig. 2.1, we show hurricane positions as well as their hashtag usage rate timeseries with a time series indicating the usage rate of the hashtag of the form `#hurricane*`.

We plot the same hashtag usage rate time series below on both linear and logarithmic axes, as well as 2-gram usage rates. For clarity, we only include hurricanes reaching at least category 4.

The hurricane map tracks are meant to show the spatial dependence of attention given to hurricanes, while giving enough visual cues to connect locations along the path to the time the attention was observed. We generated the map shown in Fig. 2.1 by filling in the polygon defined by the set of points lying at the end of a line segment of length proportional to the smoothed usage rate of the related hashtag, along the vector normal to the current velocity of the hurricane, and centered at the hurricane position at the given time.

Our hashtag usage rate is at the day scale, while HURDAT has 3 hour resolution, so the wrapped attention volume is smoothed with a moving average with a window size of one day to avoid discontinuous jumps. This method obscures any sub-day scale resolution on the map, which could be related to the daily fluctuation of tweet volume as well as varying interest in the hurricanes. While we lose some granularity using daily usage rates, the decays in attention are spread out over days and weeks for smaller storms, and months for larger storms. Daily resolution is sufficient to capture the longer decays in attention, which are our primary interest.

Examining the map, we can see the minimal attention paid to Hurricane Harvey as it traveled across the Caribbean sea and made landfall in Mexico. It is only after crossing the Gulf of Mexico that the hashtag registered on our instrument, and only when it was about to make landfall over Texas did the hashtag usage rate approach its maximum rate,

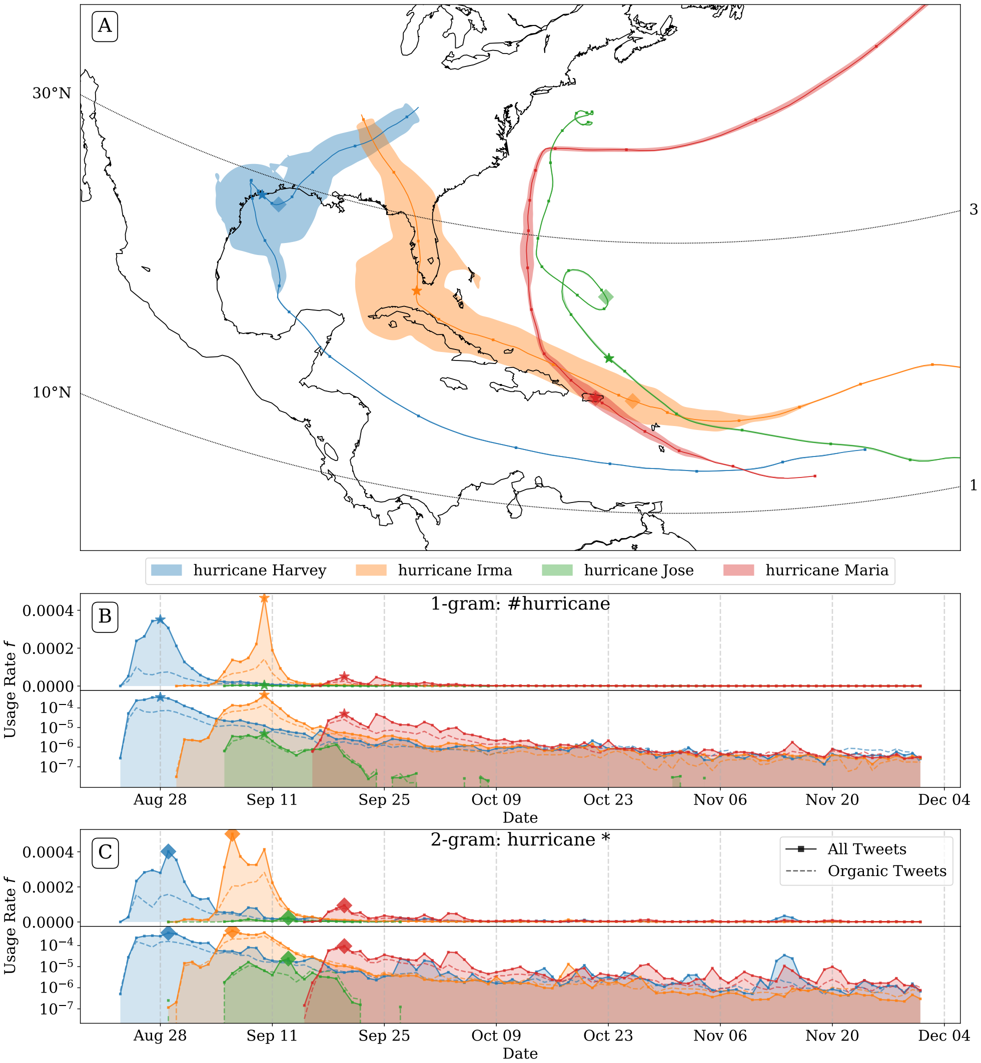


Figure 2.1: ***Hashtag attention map and usage rate time series*** *for* 1*-grams matching the case-insensitive pattern "*`#hurricane*`*" for all four hurricanes reaching at least category 4 in the 2017 hurricane season. Markers along the hurricane trajectory indicate the National Oceanic and Atmospheric Administration (NOAA) reported position for every day at noon UTC. On the map, the smoothed rate of hashtag usage is wrapped in an envelope around the hurricane trajectory in panel A, showing the spatial dependence of attention on Twitter. In the lower two plots, panels B and C, we show the usage rates for hashtags and* 2*-grams matching* `hurricane*` *in English language tweets on linear and logarithmic scales. Usage rates within all tweets are indicated with a solid line, while usage rates in 'organic' tweets (tweets that are not retweets), are represented by a dashed line. The day of maximum attention on Twitter is marked with a star or a diamond for hashtags or 2-grams, respectively. Generally, hurricanes making landfall on the continental United States received greater attention than those not making landfall. The hashtag usage rate for Hurricanes Harvey and Irma at their maximum were approximately an order of magnitude larger than the maximum hashtag usage corresponding to hurricane Maria, and two orders of magnitude larger than Hurricane Jose.*

approximately 3 of every 10,000 1-grams in English tweets. It appears that the devastation wrought by Harvey primed hurricane-related conversation, as the next hurricane, Irma was talked about long before it made landfall. While Irma was talked about with a similar usage rate as Harvey as it impacted Puerto Rico, Hispaniola, and Cuba, it spiked while making landfall in the Florida keys.

Comparing the attention generated by the previous two storms, Hurricane Maria generated substantially less hashtag usage. The peak of its attention gathered as it made landfall over Puerto Rico as a category 4 storm, with less than a fifth of the attention as the hurricanes making landfall on the US. Part of the reason may be due the affected area being Spanish speaking, while our hashtag usage measurement only counts occurrences in English tweets. We find that usage rates of the 2-gram "`Huracán Maria`" in Spanish tweets were also lower than the usage rates for "`Huracán Irma`", but comparable to those for "`Huracán Harvey`." See Fig. 6.1 to compare top hurricane related 2-gram time series for the 2017 hurricane season in English and Spanish.

Another potential contributing factor for the low volume of Hurricane Maria tweets could be that Puerto Rico's electric grid was destroyed and 95% of cell towers were down in the aftermath of the storm, making it impossible for those directly affected to communicate about the storm [139]. Unfortunately, due to Twitter's usage norms in this time period, we do not have locations for the vast majority of tweets. The number of people affected by the storms could also help explain the different levels of attention, as both Hurricane Harvey and Irma affected 19 million people, while Maria affected about 4 million [21].

2.4.2 Hurricane Attention Comparison

To compare the variation in attention received by different storms, we combined measurements of the hashtag usage rate with deaths and damages caused by each storm from 2009 to 2019. The supplimentary materials, Section 6.1, shows these raw measured values for the most damaging hurricanes in this period.

In Fig. 2.2, we show radar plots (radial, categorical charts) comparing six measurements of impact and attention for each of the eight most damaging hurricanes in the time period of study [165].

Included measurements are:

- Max Usage Rate—peak attention on any single day
- Integrated Usage Rate—total attention over the entire hurricane season
- Quantile 0.9: $Q_{0.9}$—days to 90% attention
- Quantile 0.99: $Q_{0.99}$—days to 99% attention
- Damage—total damage caused by the storm in US dollars
- Deaths—total deaths associated with the storm (both direct and indirect)

The relative magnitude of each quantity is shown as a fraction of the maximum value for any storm in the study. The quantile values are non-parametric measurements of the attention time scale—comparable to half-lives but without the assumption of an exponential decay. Some storms receive significant interest months after they pass, usually related to the recovery efforts. Spark lines above each plot show the attention time series for the year after each storm, as measured by the log usage rate, but do not convey relative scale.

The three most damaging storms, Hurricanes Harvey, Maria, and Irma, all destroyed tens of billions of dollars of property. Storms in Fig. 2.2 are ordered by damage, with the least damaging being Hurricane Irene in 2011, which still destroyed an estimated $14 billion in property.

The most deadly North Atlantic hurricane in the past decade was Hurricane Maria, killing over 3000 people over the course of the extended disaster. The next most deadly storms were Hurricanes Matthew, Sandy, Irma, and Harvey, all killing at least 100 people.

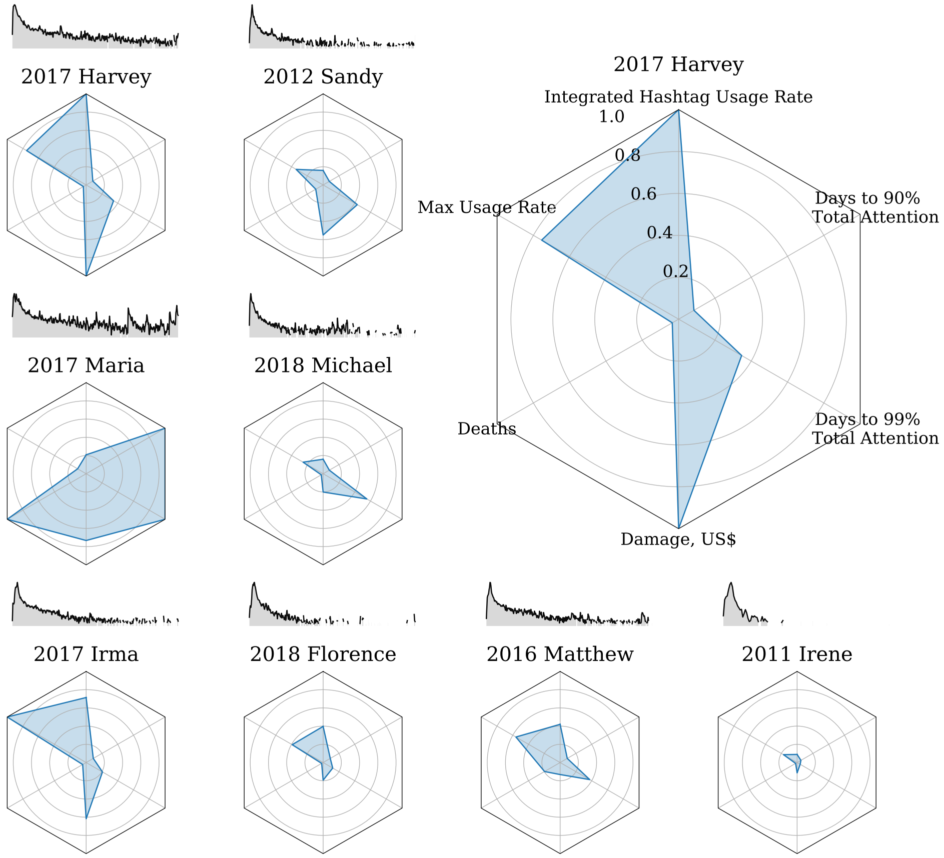


Figure 2.2: ***Radar plots comparing the eight most monetarily damaging hurricanes in the North Atlantic basin from 2009 to 2018.*** *For each plot, starting at the top position and rotating clockwise the measures are: the sum of usage rate of the hashtag, the number of days to reach 90% and 50% of the total attention received during that season, the total cost in dollars attributed to damage caused by the hurricane (in its year), the number of deaths attributed to the hurricane, and maximum usage rate of the hashtag during the year of interest. All measurements are normalized to the maximum value achieved by any hurricane. Hurricane Harvey was the most talked about hurricane, as well as the most damaging. Hurricane Irma was the most talked about on any single day. Hurricane Maria caused the most deaths, and had the longest attention half-life of all measured hurricanes. Raw values for this figure are shown in Section 6.1. Hashtag usage rate spark lines above each radar plot are normalized to show the common decay shape, and can not be compared to evaluate relative volume, and are shown on a log scale.*

Among the storms shown in the Fig. 2.2, Hurricanes Florence and Irene were the least deadly, causing 58 and 57 deaths, respectively.

The highest hashtag usage rate on a single day was associated with Hurricane Irma, reaching $\max f_{\tau} = 4.6 \times 10^{-4}$, or 4.6 of every 10,000 1-grams, as the storm made landfall over the Florida Keys. Other storms reached comparable single day usage rates, such as Hurricanes Harvey and Matthew, reaching $\max f = 3.5 \times 10^{-4}$ and $\max f = 2.6 \times 10^{-4}$, respectively. Within the top eight most damaging storms, the hashtag associated with Hurricane Maria had the lowest maximum usage rate. The hashtag "#hurricanemaria" appeared only five times for every 100,000 1-grams as Maria made landfall in Puerto Rico.

The highest integrated hashtag usage rate was associated with Hurricane Harvey, followed by Hurricanes Irma, Matthew, and Florence. The integrated hashtag usage rate for "#hurricaneharvey", $I = 2.3 \times 10^{-3}$. Hashtags associated with Hurricanes Sandy and Irene had the total attention, with $I = 3.7 \times 10^{-4}$ and $I = 2.0 \times 10^{-4}$, respectively.

Due to the extended crisis in the aftermath of Hurricane Maria, the hashtag continued to be used at relatively high volumes even a year after the storm had passed, leading to much larger value for $Q_{0.9}$ of 175 days [132, 181]. Typical values for $Q_{0.9}$ were around 1–4 days, with more prolonged and damaging storms like Harvey in 2017 taking 15 days to reach 90% total attention. In comparison no other storm took longer than 100 days to reach this benchmark. We chose the longer term attention timescale benchmark, $Q_{0.99}$, to describe how long until nearly all storm focused attention has passed. We observe the hashtag associated with Hurricane Maria is the largest for this measurement as well, with $Q_{0.99}$ of 363 days, which should be interpreted as attention not dying away within a year, since we truncate the timeseries after one year. Hurricane Michael, Sandy, and Harvey also have triple digit values for $Q_{0.99}$, as they continued to be talked about, albeit at much lower levels than their peak. Other storms quickly lose attention, such as Hurricane Irene, which took only 12 days to reach 99% total attention.

We observed variation in the overall radar plot shape. More recent storms have been more damaging and deadly, and we find higher measures of total attention and attention decay. A number of storms like Sandy, Michael, and Matthew have relatively higher values for both maximum usage rate and number of days to reach 99% total attention. While there is significant variation in the magnitude of these measurements, the essential exogenous shape of the hashtag usage rate timeseries, f, is consistent.

2.4.3 Attention and Impact Regressions by Category

We next explore the associations between damage, deaths, and attention given to hurricanes. In Fig. 2.3, we show the scaling relationship between attention and impacts for each category storm on the Saffir-Simpson wind scale [151]. Each sub-panel plots the integrated usage rate, $I = \sum_t f(t)$ for hashtag or 2-gram τ, against a measure of storm impact, where t runs over an index of the 365 days after each storm began. I is chosen as a measure of total attention given to the storm during its respective hurricane season, which can be compared across years since it is already normalized to the total volume of conversation on Twitter. Color represents the maximum category storm reached, and the smaller subplots are breakout panels for each category. We include Spearman's ρ, a non-parametric measure of rank correlation, in each panel.

We perform linear regressions on storms in each category separately, a choice that models the attention received by different category storms as separate processes. With models in Section 2.4.4, we separately consider attention as a singular process where we account for the hurricane's maximum category rating using an explicit indicator variable.

Model Choice and Fitting Procedure For each category and each impact, we model total attention as

$$\log_{10} I = a_0 + a_{\text{impact}} X_{\text{impact}} + \varepsilon_\tau, \tag{2.1}$$

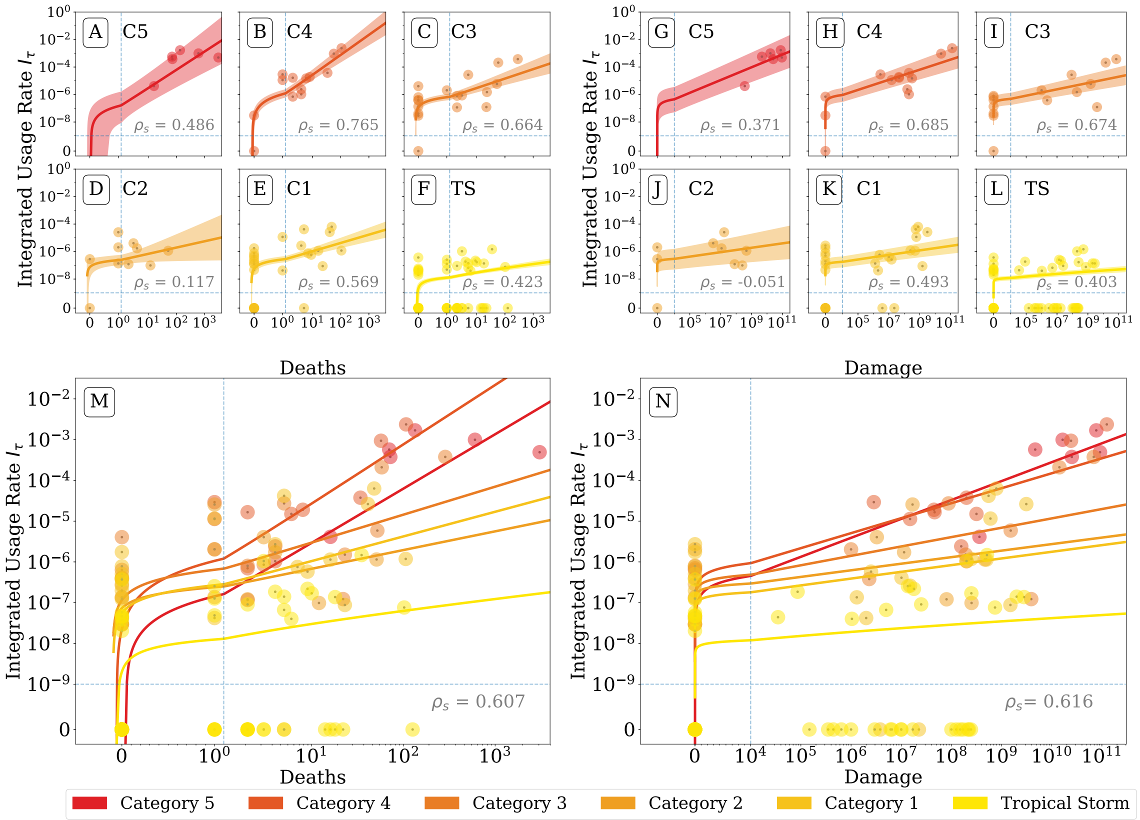


Figure 2.3: ***Scatter plots for integrated hashtag usage rate versus the deaths and damages caused by each storm.*** *There is a clear positive association between the total attention represented by hashtags and the impacts of these storms. We reported Spearman's rho, ρ_s, in the top left corner of each plot. While for some categories, there is little evidence for a positive association, for the entire dataset $\rho_s \sim 0.54$. We perform a Bayesian linear regression for each category storm between the $\log I$ and log impacts. We show the mean model, along with the credible interval within a standard deviation of the mean model. We use hybrid axis with logarithmic scaling for most horizontal and vertical values and linear scaling near zero, in order to show storms that caused zero deaths or damages, as well as storms for which we measured a hashtag usage rate of zero. Changes in axis scaling occur at the blue dashed lines. Generally, more powerful storms received more attention, higher category storms received more attention even when causing minimal damage, and high category storms had a higher regression slope. These results suggest that for powerful storms, a given increase in impact was associated with a larger increase in attention. While for category one storms a 10-fold increase in deaths is associated with a two-fold increase in attention, for category five hurricanes, this same 10-fold increase in attention is associated with a 27-fold increase in attention.*

where X_{impact} is either $\log_{10}$ deaths or $\log_{10}$ damages caused by each storm. We use a logarithmic model both to capture the scaling relationships between impacts and attention and to inform on the relative changes in attention associated with storm impacts. We offset

I by 10^{-8} and the log impacts, X_{impact} by $\$10,000$ and 0.1 deaths, respectively to avoid divergent log data where observed values are equal to zero.

We set a zero-centered normal prior on the slope of the regression model as $a_1 \sim$ $\mathbf{normal}(0, 1)$. We set a normal prior on the intercept of the model with mean equal to $\log_{10} I = -8$, the minimum value of the offset added to I. We did not have strong beliefs about the likely precision of a_0 since it was not *a priori* clear how much attention would be paid to hurricanes with very little associated monetary damage or few deaths. We thus set a weak hyper-prior on the precision of a_0, $\tau \sim \mathbf{gamma}(3, 1)$; the intercept of the regression is distributed as $a_0 \sim \mathbf{normal}(-8, \tau^{-1})$.

We found regression coefficients by sampling with the No-U-Turn-Sampler (NUTS), using 8 chains with 2000 draws each after 1000 steps of burn-in [74]. Our models converged, with the Gelman-Rubin statistic, $\hat{R}$, never exceeding 1.004 for any parameter in the 12 models fit.

Model Posteriors and Discussion In Fig. 2.3, we show the fitted regressions for each category. The size of the impact and attention variables vary over many orders of magnitude, but also include zero values, corresponding to storms that cause no deaths or damage, or had zero usage of the hashtag associated with their name during the year the storm was active. Note that it should not be surprising that tropical storms appear to receive less attention via our hashtag usage rate measurement, since they never officially become hurricanes, and thus many of the tropical storm hashtags have an integrated usage rate, $I = 0$.

To display all data, we use symmetric log axes: logarithmic for large values and linear for small values. We indicate the switch point from linear to log space axis as blue dotted lines. This choice of axes causes the linear regressions on the log transformed data to appear curved for small values.

In each of the small subplots of Fig. 2.3, we show the 1σ credible interval for the model as a band around the mean regression model. The credible interval is noticeably wider for category five storms, which is reasonable given there are only seven storms reaching this category. Generally the mean regression lines are ordered such that higher category storms are receiving more attention than lower category storms. The slopes of the regressions are also higher for higher category storms. However, to better understand the models, we need to compare the model parameters individually.

In Fig. 2.4 we provide posterior distributions for model parameters, which show that, as expected, more intense storms receive more attention per unit of log impact than weaker storms. For category five storms, we find a mean regression co-efficient of $a_{\text{deaths}} = 1.35 \pm 0.39$, using the format $\mu \pm \sigma$ where μ is the mean and σ is the standard deviation, while for category one storms we find a mean regression co-efficient of $a_{\text{deaths}} = 0.61 \pm 0.18$.

Looking at associations between log damages and log attention we find $a_{\text{damage}} = 0.46 \pm 0.07$ for category 5 storms, while for category one storms we find $a_{\text{damage}} = 0.17 \pm 0.05$.

To interpret the regression coefficients, a_{impact}, as representing proportional increases in attention per proportional increase in impact, we exponentiate the coefficient. Thus, our model shows a 10-fold increase in deaths for a category 5 storm is associated with a 22-fold increase in attention, while for a category 1 storm the same 10-fold increase in deaths is associated with a 4-fold increase in attention.

The intercepts, a_0, for higher category storms tend to be larger, meaning that for a theoretical minimally disruptive storm causing exactly \$1 of damages or one death, a powerful storm would be talked about more, as shown in Fig. 2.4. We believe this trend could continue for category 5 storms, but we have observed only $n = 6$ such storms for the duration of our attention dataset. We interpret the intercepts as indications of how much attention low-impact storms receive on average.

In Fig. 2.4, we fit another regression model on all hurricanes examining log deaths and log attention. We find a 10-fold increase in deaths is associated with a 14-fold increase in attention, since the mean value of $\bar{a}_{\text{deaths}} = 1.16 \pm 0.15$ For damages, coefficients tend to be

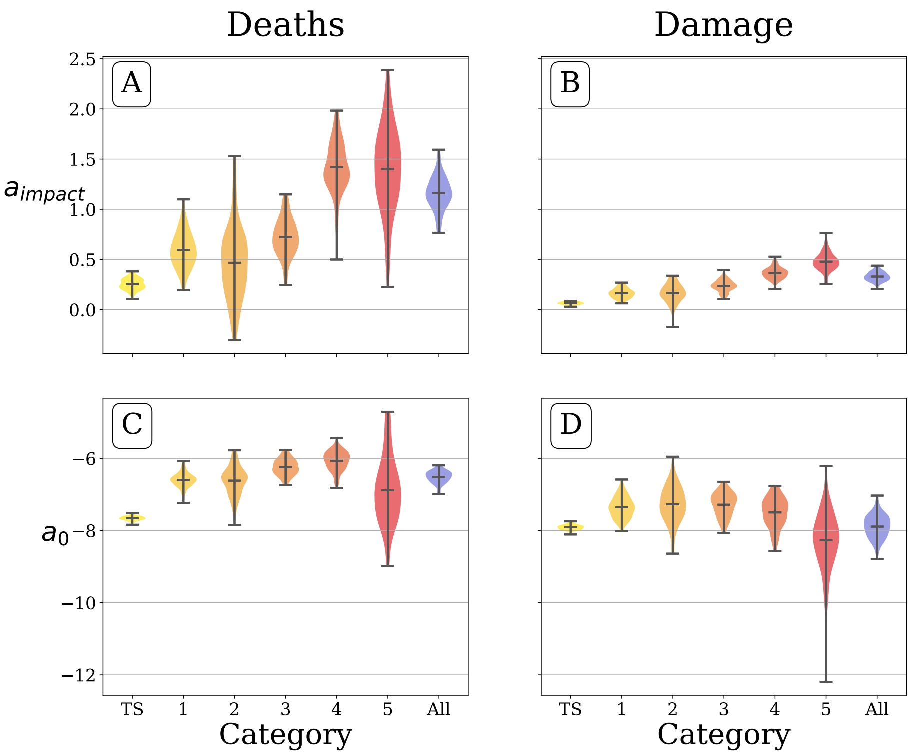


Figure 2.4: ***Posterior distributions of regression parameters*** *for the model* $\log_{10} I \sim a_0 + a_1 X_i$*,where* X_i *is either the log number of deaths (A and C) or log damages in dollars associated with the storm (B and D), and* $\log_{10} I$ *is the log integrated hashtag usage rate. The trend in regression coefficients for association between the log attention and log deaths suggests that higher category storms receive more attention per unit impact, while the trend of intercepts shows increasing baseline attention for a hypothetical minimally disruptive storm causing exactly $1 in damages or one death. For regression coefficients relating log attention to log damages, Category 4 and 5 storms receive more attention per unit increase in log damages than lower category storms. However, the coefficients are smaller in magnitude due to damages varying across 7 orders of magnitude, as compared to deaths varying over 4 orders of magnitude. There is a larger uncertainty for the category 5 intercept values, as only 6 storms of this intensity formed between 2009 and 2019 in the Atlantic basin. At the right of each plot, we show the coefficients for the model fit for all hurricanes (blue violin), excluding tropical storms. Above each category, we show the value of the mean posterior distribution for each parameter. For a table of mean parameter values, see Table 6.1.*

lower than those for deaths: $\bar{a}_{\text{damage}} = 0.31 \pm 0.05$. We intepret this coefficient as a 10-fold increase in damage being associated with no more than a 2-fold increase in attention.

2.4.4 Regression Models for Impacts, Impact Interactions and Hurricane Category

In order to better understand the scaling of attention with hurricane impacts, we fit a number of models on the log transformed data. We applied the same offsets as in the previous section to avoid non-finite log transformed data. We exclude tropical storms, since their attention is not captured in same way as our string matching for hurricanes.

Regression 1 We fit the regression model,

$$\log_{10} I = a_0 + a_{\text{death}} X_{\text{death}} + a_{\text{damage}} X_{\text{damage}} + \varepsilon, \tag{2.2}$$

where both predictors X are log impacts, which we be referred to as regression 1. The regression coefficients can be interpreted as the increase in log attention received for every unit increase in log impact. Likewise, the intercept can be interpreted as the expected attention for a minimally damaging storm causing one death and \$1 of damage. This model is distinguished from the previous section by including both log impacts in a single model, while not including an interaction term as later models will.

We set priors for the model as shown in Section 6.1. We chose the intercept, $a_0 \sim$ $\mathbf{normal}(-8, 3)$, to be centered around -8, approximately the lowest usage rate captured in our data, as we guess storms causing 1 death and \$1 worth of damage are talked about relatively little, but wish to allow a wide range of uncertainty spanning a few orders of magnitude. We chose the priors for the regression coefficients, $a_{\text{death}} \sim \mathbf{normal}(0, 1)$ and $a_{\text{damage}} \sim \mathbf{normal}(0, 1)$, to be weakly informative and centered around zero, as to not bias towards any association. We sampled the coefficients' posterior distributions using NUTS, using 8 chains with 2000 draws each, after 500 steps of burn-in [74]. We found the model converged, with the maximum value of $\hat{R} = 1.000$.

We show the posterior distributions of model parameters for regression one in Panel A of

Fig. 2.5, which have a positive scaling between both deaths and damages, and the amount of attention commanded by the storm, as measured by the log hashtag usage rate. We intepret the mean value of $a_0 = -7.57 \pm 0.5$ for the regression constant as the expected log hashtag usage rate for a minimally destructive storm, i.e., that in English tweets, the hashtag usage rate would integrate to $10^{-7.57}$ over the season. We provide summary statistics in Table 6.3.

At first glance, this level of attention seems remarkably low: if occurring all in a single day, this is little more than 1 usage for every 100 million 1-grams. The most devastating storms can have integrated usage rates of $I = 2.3 \times 10^{-3}$, five orders of magnitude more attention than our regression constant. However, the least impactful storms affect relatively few people, while the most destructive storms significantly disrupt the lives of tens of millions, so the differences in the scale of total hashtag usage rate are not unreasonable. See Section 6.1 for measured values corresponding to each storm.

We find $a_{\text{death}} \simeq 0.49$ and $a_{\text{damage}} \simeq 0.24$. Because $10^{0.24} \simeq 1.7$, considering the results in linear space, a 10-fold increase in damages is associated with a 1.7-fold increase in hashtag usage rates, while a 10-fold increase in deaths is associated with a 3-fold increase.

Regression 2 For the second regression, an interaction term was introduced between the log number of deaths and the log damages,

$$\log_{10} I = a_0 + a_{\text{death}} X_{\text{death}} + a_{\text{damage}} X_{\text{damage}} + \\ a_{d,D} X_{\text{death}} X_{\text{damage}} + \varepsilon. \quad (2.3)$$

Prior distributions for the intercept and main effect coefficients are unchanged from regression 1, and we set the prior distribution for the interaction coefficient to be $a_{d,D} \sim$ $\mathbf{normal}(0, 1)$, a standard weakly informative prior for regression coefficients. We used identical fitting procedures as above, and found the models converged with a maximum value of $\hat{R} = 1.0001$.

Here, the intercept is largely the same as the simplest regression model. Interpreting a_{death} as the conditional relationship between log usage rate and log deaths when total damage is \$1, the $a_{\text{death}} = 0.05$ implies that for a 10-fold increase in deaths is associated with a 1.12-fold increase in hashtag usage rate, though the standard error includes zero. Similarly, $a_{\text{damage}} = 0.22$ implies a 10-fold increase in damage is associated with a 1.6-fold increase in hashtag usage rate. Finally, the interaction coefficient $a_{\text{d,D}}$ is small, but positive: a 10-fold increase in $X_{\text{death}}X_{\text{damage}}$ is associated with a 1.14-fold increase in hashtag usage rate. Notably, the inclusion of the interaction term significantly reduces the regression coefficient associated with deaths, while the coefficient associated with damage is largely unchanged. This provides evidence that storms that cause a large number of deaths and damages are associated with higher volumes of attention, while a storm causing a large number of deaths but relatively less damage will attract much less attention for Twitter users. This leads us to believe that damages could act as a priming factor for human attention, in part explaining why deadly disasters in capital-poor countries often receive less attention than when similarly deadly storms occur in wealthy areas.

Regression 3 To better understand the effect of hurricane category on attention, we performed a regression including this categorical variable, modeled as

$$\begin{aligned}\log_{10} I = a_0 + a_{\text{death}}X_{\text{death}} + a_{\text{damage}}X_{\text{damage}} + \\ a_{d,D}X_{\text{death}}X_{\text{damage}} + \sum_j a_{C_j}X_{C_j} + \varepsilon, \quad (2.4)\end{aligned}$$

where the index j runs from 2 to 5. We did not include a variable for category 1 hurricanes to avoid issues of multi-colinearity. Fitting procedures were identical to above, and we found the model converged with the max value of $\hat{R} = 1.0003$.

We did not change priors for the model coefficients from above for existing parameters, and we set the coefficients for category indicator variables to a weakly informative prior,

$a_{C_i} \sim \mathbf{normal}(0, 1)$. Since we have included our hurricane categories, the interpretation of the intercept a_0 is now the expected log integrated hashtag usage rate I for a category one hurricane, which causes one death and \$1 of damage. The value is similar to the other regression models. Effect sizes for a_{damage} and $a_{\text{d,D}}$ are reduced in magnitude slightly compared to the preceding regression.

As measured by the integrated hashtag usage rate, compared to a category 1 storm causing the same deaths and damages, hurricanes in:

- category 2 receive 1.14 times more attention,
- category 3 receive 1.5 times more attention,
- category 4 receive 5.6 times more attention,
- and category 5 receive 4.6 times more attention.

We show the posterior distributions for regression three in Panel C of Fig. 2.5.

2.5 Concluding Remarks

We have explored the attention given to hurricanes as measured by the hashtag and 2-gram usage rate. We quantify the relative volume of attention time series for major storms. We find evidence that not only are more powerful—higher maximum category rating—storms talked about more than weaker storms, but they are talked about more when they inflict the same amount of damage or take the same number of lives. Further, different attention scaling relationships exist for different category storms. For the most destructive storms, we demonstrate that a 10-fold increase in deaths is associated with a 27-fold increase in attention, while for weaker storms the same proportional increase in deaths would lead to only a 3-fold increase in attention on average.

How people outside of the government agencies and non-governmental organizations (NGOs) tasked with responding to natural disasters perceive the importance of disasters

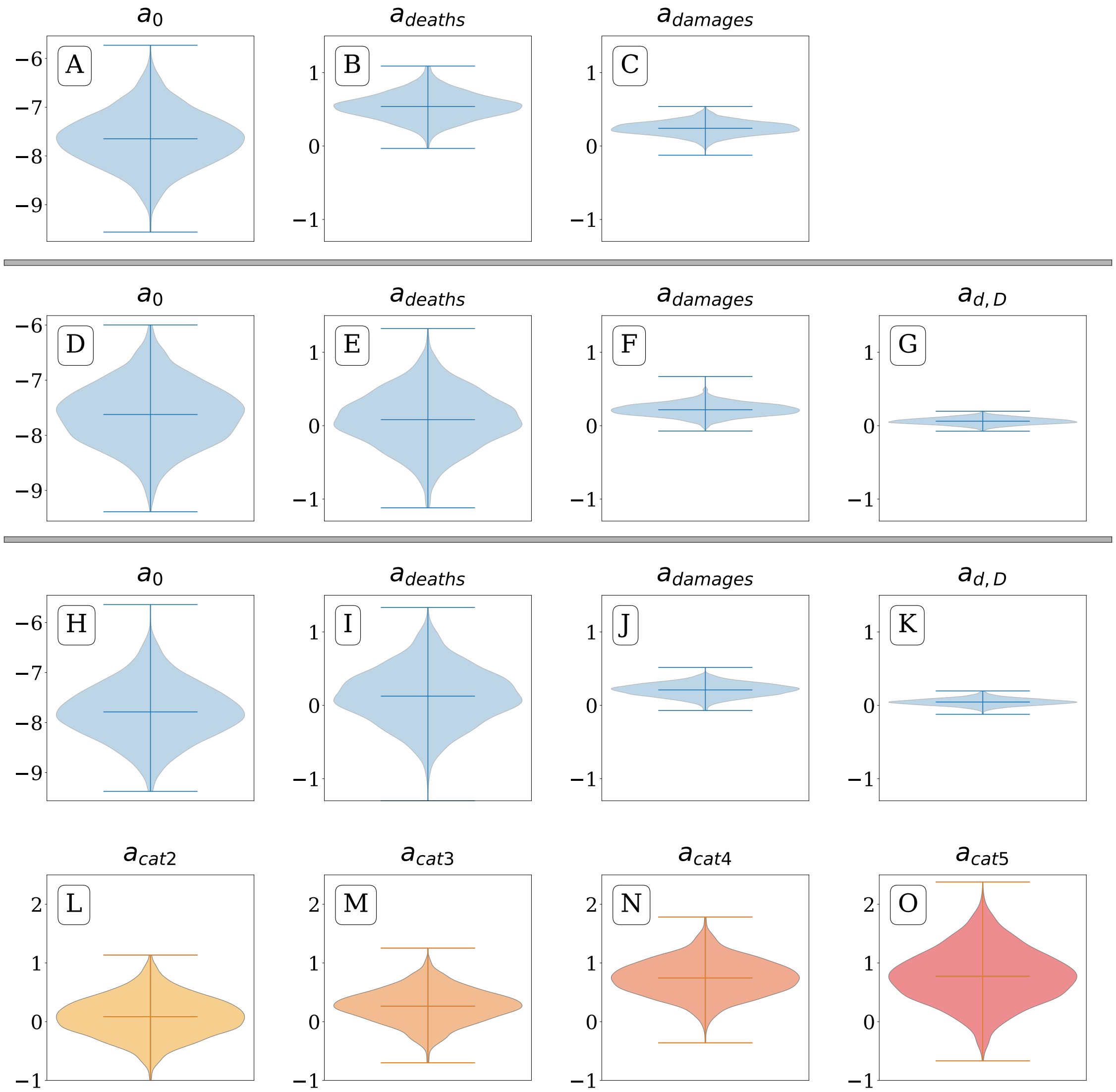


Figure 2.5: ***Parameter distributions for models 1, 2 and 3****. Plots A–C show posterior distributions for regression 1, plots D–G show distributions for regression 2, which includes the addition of an interaction term, and plots H–O showing distribution for regression 3, which includes indicators variables for hurricane categories two through five. The addition of the interaction term,* $a_{d,D}$ *increases posterior variance for* a_{deaths} *as well as reducing its mean from* $a_{\text{deaths}} = 0.49$ *in regression 1 to* $a_{\text{deaths}} = 0.05$ *in regression 2 and* $a_{\text{deaths}} = 0.12$ *in regression 3, suggesting that while the number of deaths is associated with increased attention, attention response is primed by destruction. Additionally, the hurricane category indicator variables in regression 3 show the progressive increase in attention given to higher category storms compared to category 1 hurricanes.*

have real-world consequences [22, 113]. We hypothesize that monetary donations to NGOs that assist with hurricane disaster relief efforts are strongly associated with the amount of attention attracted by the hurricane. If this is true, it could be advantageous for NGOs to prospect for financial contributions while collective attention is focused most strongly on a storm [70]. It is also possible that the speed and scale of governmental relief programs are influenced by popular attention paid to storms, and previous work has shown that relief has been inequitable in the past [168]. Future work could compare the quantities of non-profit and governmental assistance with attention volume.

While the users of Twitter are certainly not representative of the world, or even English speakers, measuring the text they generate approaches measurement of the population at large, at least more-so than published books or edited newspaper columns [75, 80, 117, 145, 169]. The digital signatures left behind by our collective online presence offers rich data for observational studies of everyday language with unprecedented time resolution. Of course, many tweets referencing hurricanes are authored by journalists or news organizations and future efforts could attempt to disentangle the various motivations contributing to the overall usage rate of hashtags and other n-grams.

Another limitation of our work, particularly relevant to any geospatial findings, is that we only consider tweets classified as English. While the density of English speakers closely mirrors the population density for much of the United States, we observe much lower usage rates for the English language hashtags and 2-grams over predominately Spanish speaking areas. While different populations may use different n-grams to reference the same storm, for the purposes of our study we have focused only on the English-speaking population of Twitter.

Future work could consider how to better quantify the total fraction of conversation of Twitter focused on a storm or event of interest. Our current method only includes counts for individual n-grams, which we believe acts as a proxy of total attention, but almost certainly underestimates the total fraction of text devoted to discussing a topic. Hashtag co-occurrence network-based methods could help to identify the most prominent hashtags associated with a given storm, or any event of interest, and to classify tweets as relevant. Examining properties of this network changing in time, such at the integrated usage rate of all significant hashtags within one degree could give a more unbiased view of the total attention surrounding the hurricane than our current method. Other dynamics of hurricanes could be explored in this way, perhaps by encoding Jenson-Shannon Divergence shifts between hashtags as a node attribute [45], or more simply how the most frequently used hashtags in this ego network change in rank over time, as different phases of the storm occur. Authors of previous works studying the effectiveness of NGO hashtag usage following natural disasters could exploit these network based methods [176].

Chapter 3

Curating corpora with classifiers: A case study of clean energy sentiment online

3.1 Abstract

Well curated, large-scale corpora of social media posts containing broad public opinion offer a supplemental data source to complement traditional surveys. While surveys are the gold standard for collecting representative samples and are capable of achieving high accuracy, they can be both expensive to run and lag public opinion by days or weeks. Both of these drawbacks could be addressed with a real-time, high volume data stream and fast analysis pipeline, and provide valuable insights so long as the limitations of using these non-representative populations are understood and acknowledged. A central challenge in orchestrating such a data pipeline is devising an effective method for rapidly selecting the best corpus of relevant documents for analysis. Querying with keywords alone often includes irrelevant documents that are not easily disambiguated with bag-of-words natural language processing methods. Here, we explore methods of corpus curation to filter irrelevant tweets using pre-trained transformer-based models, fine-tuned for our binary classification task on hand-labeled tweets. We are able to achieve F1 scores of up to 0.95. The low cost and high performance of fine-tuning such a model suggests that our approach could be of broad

benefit as a pre-processing step for social media datasets with uncertain corpus boundaries.

3.2 Introduction

The wide-spread availability of social media data has resulted in an explosion of social science studies as researchers adjust from data scarcity to abundance in the digital age [88, 89]. The potential for large scale digitized text to help understand human behavior remains immense. Researchers have attempted to quantify myriad social phenomena through changes in language use of societies over time, typically through the now massive collections of digitized books and texts [110] or natively digital large-scale social media datasets [8].

Analysis of social media data promises to supplement traditional polling methods by allowing for rapid, near real-time measurements of public opinion, and for historical studies of public language [29, 30, 121, 173]. Polling remains the gold standard for measuring public opinion where precision matters, such as predicting the outcomes of elections. Where trends in attention or sentiment suffice, social media data can provide insights at dramatically lower costs [122]. However, for targeted studies using social media data, researchers need a principled way to define the potentially arbitrary boundaries of their corpus [144].

When researchers characterize online discourse around a specific topic, a few approaches are available. Each comes with trade-offs, both in the costs of researchers' time, as well as the resulting precision and recall of the corpus.

For some studies a corpus is best defined by a set of relevant users, such as a set of politicians' social media accounts or the set of users following a notable account [82]. Studies that observe the behavior of networked publics often take this user-focused approach [19]. For studies of social media advertising, a list of relevant buyers can be used to define the boundaries, whether politicians or companies [3, 90].

To curate a topic-focused corpus limited keyword filters can be an effective strategy. Keywords can be used to match a broad cross-section of relevant posts with high precision, but often have low recall [100]. Relevant hashtags can signal a user's intent to join a specific online conversation beyond their immediate social network. Hashtag based queries have been used by researchers to construct focused corpora of tweets ranging from sports and music [18, 27], to public health, natural disasters, political activism, and protests [14, 54, 56, 57, 64, 78, 95, 102, 148].

Alternatively, researchers can query for posts with an expansive set of keywords to increase recall at the expense of precision. Researchers can generate such a set of keywords algorithmically, or by asking experts with domain knowledge, or via a combination of the two. Expert-crafted keyword lists have been used by researchers to study topics such as social movements and responses to the COVID-19 pandemic [26, 67, 78, 144]. Other researchers have generated lists of keywords algorithmically, e.g., using Term Frequency - Inverse Document Frequency (TF-IDF) [4] and word embeddings [105], or by comparing the distribution of words in a corpus of interest to a reference corpus and selecting words with high rank-divergence contributions [5, 9, 45, 114, 149]. Regardless of the methods used to choose keywords, continued expansion beyond the most relevant ones necessarily reduces precision. Researchers can further refine the set of relevant keywords to balance precision and recall, and add complexity to their queries with exclusion terms or Boolean operators to require multiple keywords. The possibilities are endless [138] and reviewers receive little information available to decide if the choices made were appropriate.

While some topic-focused social media datasets can be well curated with simple heuristics or rules-based classifiers, others could benefit from an alternative paradigm. Here, we argue for a two step pre-processing pipeline that combines broad, high recall keyword queries with fine-tuned, transformer-based classifiers to increase precision. Our approach can trade the labor costs associated with building rules-based filters, for the cost of labeling social media data, which could potentially be further reduced using few-shot learning [161], while still achieving high precision.

The tools available for text classification have improved significantly over the past decade. Since the introduction of Word2Vec in 2013 and GloVe in 2014, the natural language processing community has had access to high quality, global word embeddings [112, 127]. These embeddings are trained vector representations of words from a given corpus of text, enabling word comparisons with distance metrics. However, global embeddings average the representations of words, making them unsuitable for document classification where key terms have multiple meanings. The subsequent development of large pre-trained language models enabled high performance on downstream tasks with relatively little additional computational cost to fine-tune [35, 99]. Such models provide contextual, rather than global, word embeddings.

Since 2019, pre-trained language models have become less resource intensive while improving performance. Knowledge distillation has enabled models like DistilBert and MiniLM, which retain the performance of full sized models while requiring significantly less memory and performing inference more rapidly [136, 160]. Smaller, faster models enable researchers with limited resources to adopt these tools for NLP tasks, requiring only a laptop for state-of-the-art performance. Improved pre-training, introduced with MPNet, combines the benefits of masked language modeling (MLM) and permuted language modeling (PLM), better making use of available token and position information [147].

While transformer-based language models provide state of the art performance on natural language processing tasks, they can be difficult to understand and visualize. Using twin and triplet network structures, pre-trained models can be trained to generate semantically meaningful sentence embeddings that can be compared using cosign distances [130]. Through pre-training with contrastive learning on high quality datasets, general purpose sentence embeddings like E5 have become the new state-of-the-art [158].

Text classification still remains a difficult task. Existing models are less successful with longer texts [59], and text classification with a large number of classes remains challenging [25]. However, for the specific task of classifying tweets [13] as ‘relevant’ (R) or ‘non-relevant’ (NR) to a specific topic—an instance of binary classification—we feel existing models are sufficiently capable. Sophisticated, pre-trained language models are readily accessible to researchers from Hugging Face [171] and can be easily fine-tuned with a limited amount of labeled data [161, 177]. Do *et al.* found that fine-tuned models trained with expert labeled data can outperform crowdworkers and match the performance of trained research assistants [38]. Tools like ChatGPT have been shown to outperform untrained human crowd-workers for zero-shot text classification, while costing an order of magnitude less [61].

As a case study, we examine online language around emission-free energy technologies. In democratic societies the social perception of technologies affects the willingness of governments to extend subsidies, expedite permitting, or regulate competing energy sources, ultimately effecting the energy mix of the grid. Quantifying public attitudes is useful for policy makers to be responsive to public preferences and for science communicators to respond when public opinion does not reflect expert consensus.

To quantify public perceptions of energy on social media sites, researchers have use a variety of methods to curate tweet corpora. This could be as simple as querying for a single hashtag. Jain *et al.* choose ‘#RenewableEnergy’ to generate a corpus for a renewable energy classification study [79]. Zhang *et al.* query for tweets containing a list of hashtags, before quantifying overall attention trends and sentiment by energy source [180]. Li *et al.* use a two-phase approach, querying for relevant hashtags, before filtering non-relevant tweets with keywords, such as those containing both ‘#solar’ and ‘eclipse’, with filter keywords built on a trial-and-error approach [94]. Alternatively, Kim *et al.* use keyword phrases, such as ‘solar energy’ and ‘solar panel’, to search for relevant tweets, before using RoBERTa to classify sentiment [85]. Vågerö *et al.* use a contextual language model to classify sentiment of tweets towards wind power in Norway [154]. Using Reddit, Kim *et al.* study renewable energy discourse by collecting all messages from a particular subreddit, a page devoted to a topic, before analyzing a word co-occurrence network [84].

Published studies use a wide range of corpus curation techniques and provide varying levels of justification for each choice. Although we focus on the topic of renewable energy, we hope our methods are broadly applicable to any text-based social media dataset.

We structure the remainder of this paper as follows. In the Methods and Data section we present a description of our dataset and discuss the task of relevance classification as it relates to corpus curation. In the Results section, we present case studies for the keywords '`solar`', '`wind`', and '`nuclear`'. We examine the ambient sentiment time series for each corpus, and compare measurements between the unfiltered, relevant and non-relevant text. To show the differences in language between these corpora, we present sentiment shift plots [55] and allotaxonographs [45]. Finally, we share concluding remarks and potential future research.

3.3 Materials and methods

We explore the performance of text classifiers powered by contextual sentence embeddings for social media corpus curation through a selection of case studies related to clean energy.

3.3.1 Description of data sets

In this study, we examine ambient tweet datasets, collections of tweets that are anchored by a single keyword or set of keywords. From Twitter's Decahose API, a random 10% sample of all public tweets, we select tweets containing user-provided locations [153]. We extracted these locations from a free text location field in each user's bio, if the text matched a valid `'city, state'` string in the United States [66,96]. From this selection, we query for tweets that both contain keywords of choice and are classified as being written in the language English by FastText [81]. We define the results of this query as the unfiltered ambient corpus.

To illustrate the utility of our methods, we chose three keywords related to non-fossil fuel energy generating technologies, `‘wind’`, `‘solar’`, and `‘nuclear’`. Over the study period from 2016 to 2022, these keywords matched 3.43M, 1.39M, and 1.29M tweets in our subsample, respectively. In Tab. 3.1, we show example tweets from each corpus. We binned tweets into windows of two weeks, balancing the desire for large sample sizes for each bin with the need for higher resolution to show short term dynamics. While the terms of our service agreement with Twitter do not allow us to publish raw tweets, we provide relevant tweet IDs for rehydration.

3.3.2 Sentence embeddings

To better visualize the results of our classification algorithms, we chose pre-trained language models which had been fine-tuned to perform sentence embeddings. In Fig. 3.1, we can see the resulting distribution of tweets, colored by keyword and predicted class. We also considered that vector representations for sentences would better align with our desired abstraction level for the relevance classification task.

3.3.3 Relevance classification

Our task of interest is classifying if a post, in its entirety, is relevant to the researcher’s chosen topic of interest. Conceptually, this task is related to semantic textual similarity, for which sentence embeddings have achieved state of the art performance [24, 71]. Rather than finding nearest neighbors in a semantic space, we are training a classifier to partition the semantic space into relevant and non-relevant regions.

For training, we hand-label a random sample of 1000 matching tweets for each keyword as either ‘Relevant’ (R) or ‘Non-Relevant’ (NR) to energy production. We have made tweet IDs and corresponding labels available for both the training data as well as predicted labels

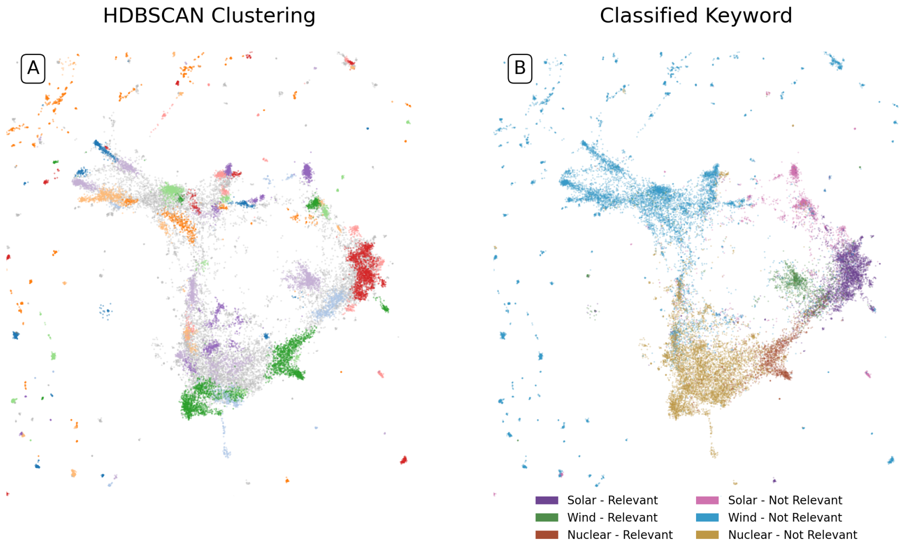


Figure 3.1: ***Embedded tweet distribution plot for the combined datasets.*** *Using a pre-trained model for semantically meaningful sentence embeddings based on MPNet, we plot the distribution of tweets within this semantic space. In both plots, points are tweets projected into 2D using UMAP for dimensionality reduction [108]. In panel A, we perform density based, hierarchical clustering using HDBSCAN and color by cluster. In panel B, we color by both the keyword used to query and the classification as relevant or non-relevant to the topic of clean energy. Relevant tweets containing the keywords '*`wind`*', '*`solar`*', and, to a lesser extent, '*`nuclear`*' are relatively close together on the right in the embeddings, while non-relevant tweets are more dispersed.*

for the full data set.

We then fine-tune nine models for comparison, based on pre-trained contextual sentence embeddings [147,160]. We list the performance of these models in Table 3.2. For each model we labeled a random sample of one thousand (1,000) tweets. We choose a train-test split of 67% and 33%. Tweets are limited to a max of 280 characters for the duration of our study period, shorter than the minimum truncation length of 256 word pieces for the models we tested.

3.4 Results

3.4.1 Interpretations of sentence embeddings

We first examine our corpus within a semantically meaningful sentence embedding, shown in Fig. 3.1. For each tweet, we compute embeddings using `all-mpnet-base-v2`, a high performing, general-purpose sentence embedding model based on MPNet. The model is pre-trained to minimize cosign distance between a corpus of 1 billion paired texts and accessed using the sentence transformers python package [130].

We include embeddings of all three corpora, anchored by the keywords '`solar`', '`wind`', and '`nuclear`', and project onto two dimensions for visualization using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction [108]. In the 2D projection, semantic distances between words are distorted. Local relationships are preserved, but global position and structure is not.

In Fig. 3.1A, we perform unsupervised clustering using HDBSCAN and color by cluster [107]. Although we cannot share the interactive version of these plots, which allow the individual tweet texts to be read, we can summarize as follows. On the right side, a large red cluster contains tweets that are primarily about solar energy. To the left in light blue, we identify a dense cluster of wind and solar tweets. Nearby in light purple, we find a cluster of wind energy related tweets. The close green cluster contains nuclear energy tweets, with those being closer to the solar and wind tweets more likely to mention renewable energy source, while those further away only discuss nuclear in isolation.

We found the performance of the semantic embedding impressive, but clustering within this embedding was unsuitable for corpus curation. For example, tweets arguing the relative merits of multiple technologies fell into a lower density location in the embedding space, and were classified as outliers by HDBSCAN, though they would clearly be classified as relevant by human raters.

In Fig. 3.1B, we show the results of our three supervised text classifiers, based on MPNet trained for sentence embeddings and fine-tuned on a dataset of 1000 labeled tweets for each keyword. The local positioning of tweets within the embedding reflects similarity in the sentence embedding space. Tweets classified as relevant to clean energy technologies are clustered on the right-hand side, and overlap where they are mentioned together. For paraphrased example tweets within each classification, refer to Tab. 3.1.

On the bottom third of the embedding, relevant '`nuclear`' tweets smoothly transition into non-relevant tweets, reflective of the occasionally blurry line between nuclear energy and weapons programs.

'`Solar`' tweets, by contrast, are easily separable. Phrases like 'solar system', 'solar eclipse', and 'solar opposites' (a television sitcom) are common example usages. These are entirely unrelated to solar energy and the sentence embedding model places them in distinct regions of the semantic space.

Relevant '`wind`' tweets are also clearly separable from non-relevant tweets, which often contain phrases related to the weather, such as 'wind storm' or 'wind speed', or more rhetorical expressions like 'wind up' or 'second wind'. A number of weather bots regularly report wind speed measurements with a template format changing only speed and location. These tweets become close neighbors in the semantic embedding and, when projected onto two dimensions by UMAP, are split off from the larger connected component and pushed to the outer edge.

3.4.2 Ambient time series plots

For each case study we compare the text in the relevant corpus to the non-relevant corpus with three figure types. The first are ambient sentiment time series plots, shown in Figs. 3.2, 3.3, and 3.4. By sentiment we broadly mean the semantic differential of good-bad (or positive-negative). In these plots we show dynamic changes in language use for tweets containing the selected anchor keyword over time. On the top panel, we show the number of n-gram tokens with LabMT sentiment scores within each time bin [44]. In the center panel, we plot the ambient sentiment, Φ, using a dictionary of LabMT sentiment values ϕ_τ. For each word τ. Wee compute the ambient sentiment as the weighted average,

$$\Phi_{\text{avg}} = \sum_{\tau} \phi_\tau p_\tau, \tag{3.1}$$

where p_τ is the probability or normalized frequency of occurrence. Error bars represent the standard deviation of the mean, with N set conservatively as the number of tweets, rather than number of tokens.

In the lower panel, we plot the standard deviation of ambient sentiment, which could help indicate when the distribution of sentiment is becoming narrower, broader, or even bimodal, indicating polarization. We plot three measurements for three corpora, tweets classified as relevant (R), non-relevant (NR), and the combined dataset (R + NR), with the latter reflecting the measurements we would have obtained without training a classifier.

3.4.3 Lexical calculus: Word shift plots

To examine how the average sentiment differs between the relevant and non-relevant corpora, we present three sentiment shift plots in Fig. 3.5 [55]. Word shifts allow us to visualize how words individually contribute to differences in average sentiment between two texts, a reference and a comparison text. Words that contribute to the comparison text having a higher sentiment than the reference, are shown having a positive contribution, $\delta\Phi_\tau$. Bars corresponding to words with a higher rated sentiment score than the average of the reference text are colored yellow, or blue if lower. Finally, we rank words by the absolute value of their contribution to the difference in average sentiment, $\delta\Phi_{\text{avg}}$, giving a list of the top contributing words.

3.4.4 Allotaxonometry

We further compare language usage using an allotaxonograph in Fig. 3.6, an interpretable instrument that provides a rank-rank histogram of word usage and a ranked list of rank-turbulence divergence (RTD) contributions from individual words. Being able to compare the 1-gram or 2-gram distributions of two corpora with RTD allows us to extract characteristic words at all scales [45]. To compute RTD, we take each distinct word, τ, and compute the ranks with each corpus, $r_{\tau,1}$ and $r_{\tau,2}$. RTD is the sum the difference between inverse ranks, scaled with a parameter, α, and normalized to lie between 0 and 1, having the form:

$$D_\alpha(R_1 \| R_2) \propto \sum \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \tag{3.2}$$

x While α is a continuously tunable parameter with $0 \leq \alpha \leq \infty$, where $\alpha = 0$ represents the limit case where common words contribute the most to rank-turbulence divergence as compared to uncommon words, and $\alpha = \infty$ represents the limit case where uncommon words dominate in the divergence measurement. We set $\alpha = 1/4$ for social media corpus comparisons [45], which we found was an acceptable trade-off between between these extremes to extract meaningful words based on their divergence contributions from all scales within the Zipfian ranked word distributions. While there is not an explicit optimization to set α, we do observe an approximate visual fit between the contours of constant α and top contributing words when comparing social media datasets.

We intend that the following cases studies may serve as an example set of procedures and provide diagnostic tools for computational social scientists to adopt this approach to social media corpus curation.

3.4.5 Solar Energy Case Study

Solar tweets were nearly evenly split with 47% of the corpus being relevant and 53% being non-relevant by volume of words. The solar tweet corpus also achieved the highest classification performance with an F1 score of 0.95, as shown in Tab. 3.2.

Of the three case studies, we find the R '`solar`' tweets corpus evolves most relative to the corresponding NR corpus. Looking at the sentiment time series in Fig. 3.2, we see little difference between the ambient sentiment of the R and NR corpora prior to 2019.

In May of 2019, NR ambient sentiment, shown in red, sharply falls while the R corpus appears to remain on trend. For the standard deviation of ambient sentiment, which measures the width of the distribution of sentiment scores for each LabMT word in the ambient corpus, we also observe a dramatic increase in 2019.

We find that this shift in language use in the NR corpus occurs without a change in query terms, and demonstrates how simple keyword queries can fail. We contend that the process of selecting relevant social media documents to include in a corpus is just as important as the NLP measurement tools used to quantify sentiment. The difference in resulting sentiment measurements, between what would have been measured without a classifier (the R + NR corpus in purple) and the improved measurement after filtering with a classifier (the R corpus in blue) is stark. Looking at only the combined R + NR measurement, researchers could incorrectly conclude that language surrounding '`solar`' has decreased in sentiment dramatically since 2019.

Focusing on only the R '`solar`' sentiment time series, we see clearly that there was in fact no dramatic drop in sentiment around '`solar`', and the relevant language around

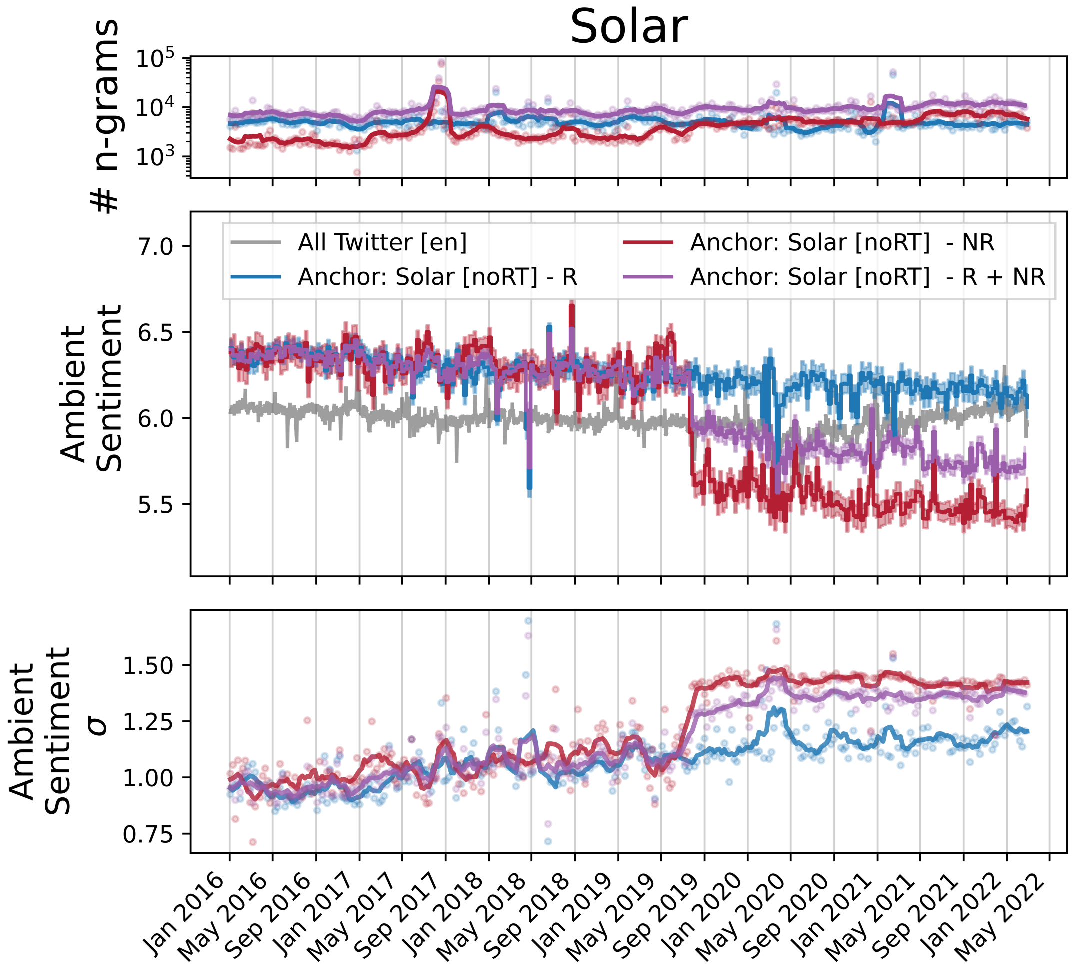


Figure 3.2: ***Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, containing the keyword 'solar'.*** *In the top panel, we show the number of tokens with LabMT [42] sentiment scores in each corpus on each day. 'Relevant' tweets, in blue, have more scored tokens early on, but the number tokens in 'non-relevant' tweets increase in relative proportion over time. The center panel shows the average sentiment for each corpus, including a measurement of English language tweets as a whole in gray for comparison. Before 2019, the measured sentiment for both corpora are comparable, but subsequently the mean sentiment of 'non-relevant' tweets drops. In the bottom panel we plot the standard deviation of the sentiment measurement, which captures a broader distribution of sentiment scores for 'non-relevant' tweets. Without classification filtering, the ambient sentiment measurement would be entirely misleading, appearing as though the sentiment contained in tweets containing the word 'solar' dropped dramatically in 2019, when in fact sentiment has only modestly declined.*

solar remains more positive relative to English language tweets in general. The decrease in observed NR sentiment is related to an influx of weather bots, which provide updates as often as hourly on local weather conditions and contain '`solar`' used in the context of measuring current solar radiation. In Fig. 3.5 we see terms like 'radiation', 'pressure', and 'humidity' are contributing to a lower average sentiment for the NR corpus.

Examining the rank-turbulence divergence shift for '`solar`' from January 2020 to March 2021 in Fig. 3.6, we can see terms like 'energy', 'power', and 'panels' are much more common in the R corpus, all being among the top 15 most frequently used terms. On the other side of the ledger, we find weather related terms like 'mph', 'uv', 'radiation', and 'gust' to be top words in the NR corpus. We also observe that function words—e.g., 'the', 'to', and 'for'—are more common in the R corpus, skewing the rank-rank histogram to the left. The lack of function words is another result of weather bots dominating in the latter period of our study.

3.4.6 Wind Energy Case Study

The unclassified '`wind`' tweets corpus had the lowest proportion of relevant tweets. Only 5% of the human labeled subset was related to clean energy. The n-gram '`wind`' is used in many different contexts besides energy generation, from casual discussion of today's weather to figurative uses like references to athletes getting their 'second wind' and the anticipatory rotational phrase 'wind up' where 'wind' rhymes with 'kind'. In the top panel of Fig. 3.3, we see that the number of n-grams in relevant tweets with corresponding sentiment scores is consistently around 10^3, while the NR corpus contains more than an order of magnitude more text.

We found the ambient sentiment of the R '`wind`' corpus has been slightly more positive than average language use on Twitter. The NR corpus had distinctly lower sentiment, but is more dynamic, rising from a low of 5.5 in 2016, to 5.9 in 2020. Because the proportion

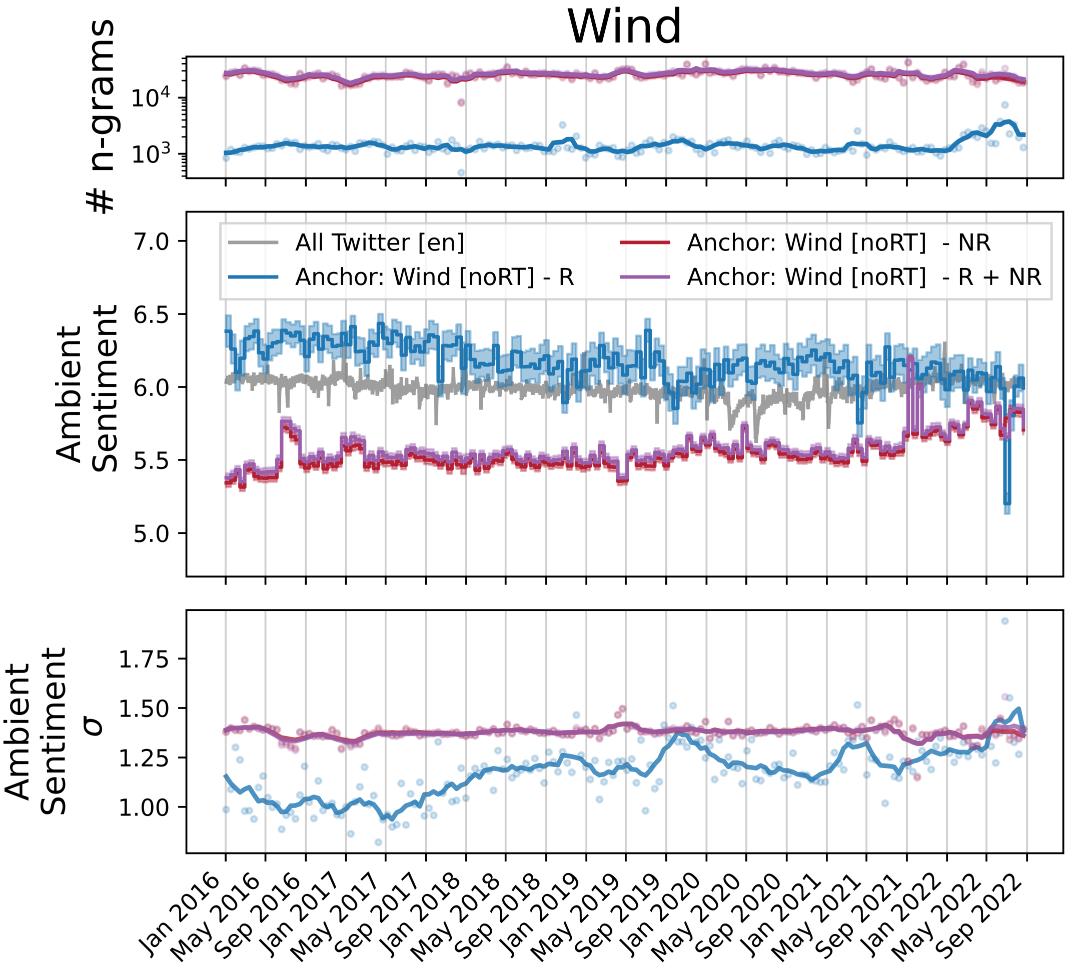


Figure 3.3: ***Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, all containing the keyword ‘`wind`’.*** *In the top panel, we show the number of tokens with LabMT sentiment scores for each corpus during each two week period [42]. R tweets, in blue, have more than an order of magnitude fewer tokens per time window over the entire study period. The center panel shows the average sentiment for each corpus, including measurement of English language tweets as a whole in gray for comparison. R ‘`wind`’ tweets are more positive than Twitter on average early on, but this difference is reduced over time. Because most ‘`wind`’ tweets are non-relevant, sentiment of the combined corpus closely follows the NR sentiment. In the bottom panel we plot the standard deviation of the sentiment measurement, which captures a broader distribution of sentiment scores for ‘non-relevant’ tweets, as was the case for all case-studies we examined. Without classification filtering, the ambient sentiment measurement would have been dominated by NR tweets.*

of tweets relevant to energy is so low, the combined sentiment time series measurement is dominated by the NR corpus. The standard deviation of sentiment, σ, for the R corpus also increases from around 1.0 in 2016, before leveling off around 1.2, slightly under the NR corpus.

The choice of ‘`wind`’ could seem to be a poor choice of keyword, given that the vast majority of matching tweets are non-relevant. Under a paradigm of expert-crafted lists of keywords, we would indeed agree such a generously matching term would not be suitable. However, by choosing a potentially ambiguous term, we are able to capture a wider range of users. Those who do not wish to project their thoughts into a global conversation by attaching a hashtag, but are content with discussing among their local network, are still included with this methodology. Also included are users writing informally or using context of a threaded conversation, who might not use a high precision keyword phrase, like ‘wind power’, ‘wind generation’, or ‘wind energy’. These cases make up a significant proportion of conversation around any given topic; researchers studying more obscure topics could benefit from the increased sample size, and temporal resolution of a higher recall set of keywords.

3.4.7 Nuclear Energy Case Study

The ‘`nuclear`’ case study had the lowest classification performance after fine-tuning, achieving an F1 score of 0.86. The proportion of relevant tweets, 16%, was higher than for the ‘`wind`’ corpus. We believe the performance was impacted negatively by the close proximity and overlap of nuclear energy and nuclear weapons topics in the semantic embedding space.

The ambient sentiment time series, in Fig. 3.4, for the R ‘`nuclear`’ corpus was much lower than average sentiment on Twitter for the entire study period, but higher than the NR corpus. It appears that ambient sentiment around R nuclear energy tweets has been increasing, with a higher stable level since fall 2020. We found that the standard deviation of sentiment is also decreasing slightly, though it starts from a much higher level of around

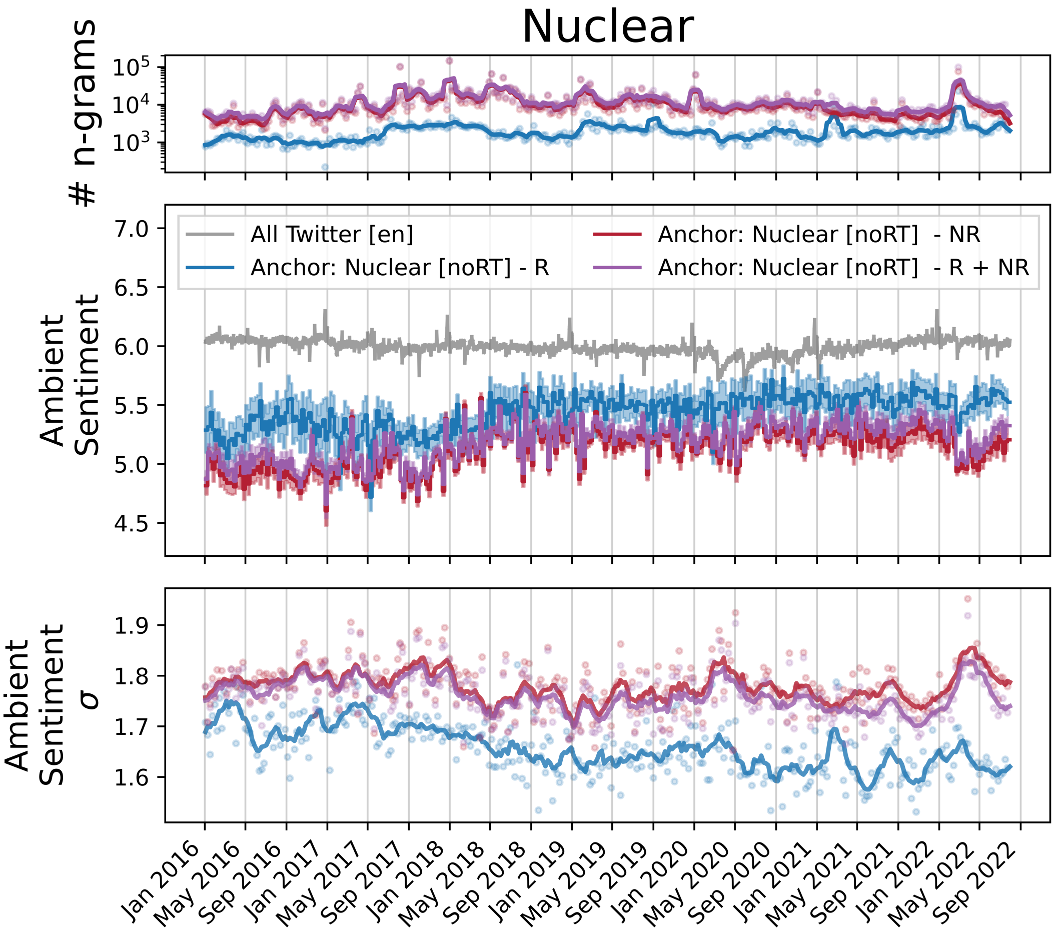


Figure 3.4: ***Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, all containing the keyword 'nuclear'.*** *In the top panel, we show the number of tokens with LabMT [42] sentiment scores for each corpus in each two week period. The number of relevant n-grams, in blue, is consistently lower than non-relevant n-grams. The center panel shows the average sentiment for each corpus, including measurement of English language tweets as a whole in gray. We found that R tweets had higher sentiment than NR tweets containing 'nuclear', but had much lower sentiment than Twitter as a whole. Sentiment appears relatively stable for both corpora with periods of higher sentiment around 2017 and 2020-2022 for the R corpus. In the bottom panel, we plot the standard deviation of the sentiment measurement, which shows a broader distribution of sentiment scores for NR tweets, as well as sentiment for both corpora trending down slightly.*

1.7, when compared with wind and solar.

In Fig. 3.5, we can see that the '`nuclear`' R corpus's higher sentiment relative to that the NR corpus is driven by more positive words like 'power' and 'energy', but also fewer negative words, like 'war' and 'weapons'. Going against the grain is the word 'nuclear' itself as well as term 'waste' which are both negatively scored words that are used much more frequently in the R corpus relative to the NR corpus.

3.5 Concluding remarks

Disambiguating relevant tweets has been a challenge for researchers, especially when a natural keyword choice has a commonly used homograph [62]. We have demonstrated that text classifiers can be trained on top of pre-trained contextual sentence embeddings, which can accurately encode researcher discretion and infer the relevance of millions of messages on a laptop.

Rather than defining the boundaries of a corpus by a set of expert chosen keywords or expert crafted query rules, researchers can look at a sample of data, label messages as relevant as they see see fit, and communicate their reasoning directly. Reviewers and skeptical readers would be empowered to make their own judgments of what qualifies as a relevant tweet, by labeling themselves and comparing the resulting text measurements.

Classification for social media datasets is not a panacea; Twitter's user base remains a non-representative sample of populations, skewing younger, more male, and more educated [116,146]. A small proportion of prolific users generate an outsized proportion of text, while most users rarely tweet [170]. Despite these problems, the platform remains a critical source of data on public conversations at the time of writing with a low barrier to entry compared to traditional media.

Future work could explore better sampling methods for humans labeling tweets to reduce the amount of labeled data needed to train the text classifier. Sampling messages by shuffling

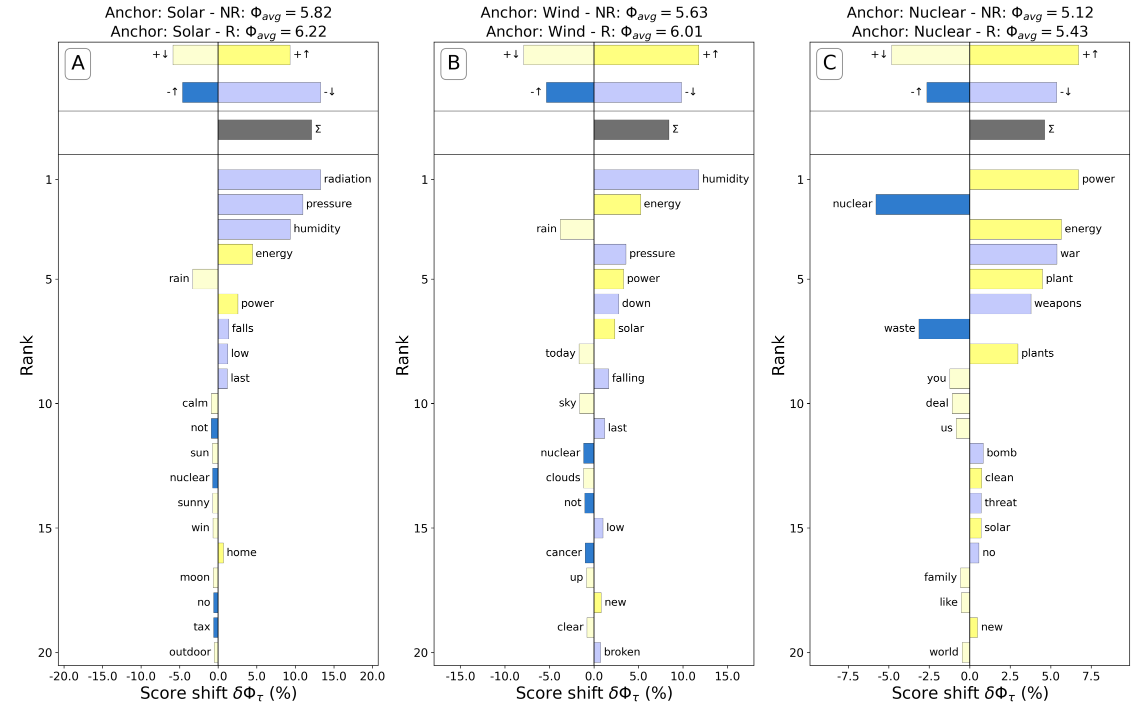


Figure 3.5: ***Sentiment shift plots comparing the classified relevant (R) and non-relevant (NR) tweet corpora for tweets containing the keywords '`solar`', '`wind`', and '`nuclear`'.*** *We show the top 20 words contributing to the difference in LabMT sentiment between the corpora.* ***A.*** *Relevant tweets that are related to clean energy are more positive on average for all keywords when compared to non-relevant tweets. Sad words that are less common in relevant '`solar`' tweets are 'radiation', 'pressure', and 'humidity', which largely refer to the weather. Happy words like 'energy' and 'power' are more common in relevant tweets compared to tweets non-relevant to solar energy.* ***B.*** *For '`wind`', relatively sad terms like 'humidity' and 'pressure' are less common in relevant tweets (these appear in clearly non-related tweets about the weather), while happy terms like 'energy', 'power', and 'solar' are more common in tweets relevant to wind as a renewable energy source.* ***C.*** *For '`nuclear`', relevant tweets are on average more positive due to sad words like 'war', 'weapons', and 'bomb' being less common in relevant tweets, while happy words like 'power' and 'energy' are more common. The two prominent sad words 'nuclear' and 'waste' go against the positive difference in moving from non-relevant to relevant tweets as they both occur more frequently in relevant tweets.*

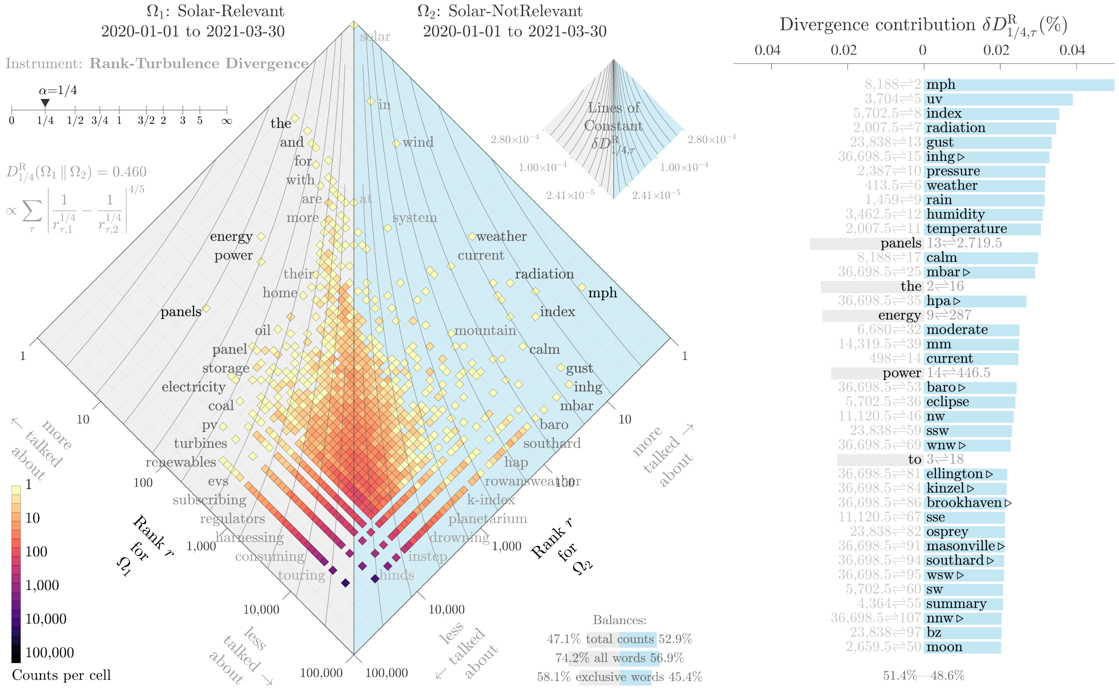


Figure 3.6: ***Allotaxonograph comparing the rank divergence of words classified as relevant to solar energy discourse to those containing the keyword '`solar`' but classified as non-relevant.*** *On the main 2D rank-rank histogram panel, words appearing on the right have a higher rank in the 'relevant' subset than in 'non-relevant', while phrases on the left appeared more frequently in the 'non-relevant' tweets. The panel on the right shows the words which contribute most to the rank divergence between each corpus. We observe that many words associated with weather bots, such as 'mph,' 'uv,' and 'pressure,' are more frequently used in non-relevant posts, while words like 'panels,' 'energy,' and 'power,' used more in tweets relevant to solar energy. Notably, commonly used function words, such as 'the,' 'and,' and 'are,' are off-center in the rank-rank histogram, a further indication that many of the 'non-relevant' tweets are from automated accounts publishing weather data rather than using conversational English. The balance of the words in these two subsets is noted in the bottom right corner of the histogram, showing the percentage of total counts, all words, and exclusive words. For this example the two subsets are nearly balanced, indicating that the filtered corpus contains less than 50% of word tokens from the raw query. See Dodds et al. [45] for a full description of the allotaxonometric instrument.*

risks oversampling from dense regions of the semantic embedding space. The coder sees repetitive messages that provide little marginal information to the model. This would have negative impacts on the generalizability of the classifier, and we would be skeptical of real-time measurements as conversation could drift into under-explored regions of the semantic embedding space. Other work could explore the trade-offs between optimizing for high recall and high precision when curating social media datasets, and the impacts on resulting measurements.

For online applications of relevance classifiers, such work would be useful in identifying when more training data is needed. By measuring changes in language use, both by measuring rank-turbulence or probability-turbulence divergence [45, 48] between the training corpus and incoming data, and by measuring changes in the distribution of messages within a semantic embedding, thresholds for train data updates could be determined.

Finally, researchers could explore viewing social media datasets as having uncertain boundaries, and running measurements over data set ensembles to better capture the uncertainly in researcher discretion inherent in corpus curation.

Overall, we hope our work here highlights a viable alternative corpus curation method for computational social scientists studying social media datasets.

Keyword	Class	Example Tweet
Solar	(R)	The decreasing costs of solar and batteries mean a sustainable future is closer than we think.
	(NR)	Looks like there's a solar eclipse down here. The space nerds bought all the hotel rooms.
Wind	(R)	At this time of year wind makes up only a fraction of the state's energy generation mix.
	(NR)	His mom caught wind of what they were up to and shut down their plans pretty quickly.
Nuclear	(R)	Nuclear activists are questioning #MAYankee's accelerated decommissioning plan.
	(NR)	The global nuclear arsenal stands around 10,000 warheads, down from 70,000 at the peak of the Cold War.

Table 3.1: ***Paraphrased example tweets for relevant (R) and non-relevant (NR) examples in each case study.***

To label the training data, we defined relevant tweets as those which are related to the topic of electricity generation or clean energy. Non-relevant tweets contained the keyword, but were wholly or primarily unrelated.

	‘solar’	‘wind’	‘nuclear’
% Relevant	43.7%	4.7%	16.0%
F1 - MPNet	0.951	**0.903**	0.860
F1 - MiniLM-L12	0.933	0.839	0.879
F1 - MiniLM-L6	0.949	0.828	0.857
F1 - DistilRoberta	**0.956**	**0.903**	0.857
F1 - paraphrase-MiniLM-L6	0.943	0.800	0.826
F1 - paraphrase-MiniLM-L3	0.918	0.714	0.814
F1 - distiluse-multilingual	0.929	0.759	**0.912**
F1 - e5-base	0.949	0.867	0.881
F1 - e5-large	0.949	0.828	0.895

Table 3.2: ***Summary statistics and model performance for each of the three case studies.***

First, we report the proportion of human labeled tweets that are labeled relevant to clean energy from our thousand tweet subsample. The ‘solar’ corpus is most evenly split, while the ‘wind’ corpus is the most imbalanced. Second, we detail F1 evaluation scores for a range of fine-tuned text classifiers trained on our labeled data. The model performance does not necessarily degrade dramatically for corpora with a small proportion of relevant documents, such as for ‘wind’.

Chapter 4

Selected contributions to published work

Studying at the Vermont Complex Systems Center has given me the opportunity to collaborate on nearly 2 dozen studies published by the Computational Story Lab related to the topic of this thesis. In what follows, I briefly describe a few key findings and my contributions to a subset of these papers.

The papers are organized into three categories. First, a set of papers, starting with Storywrangler (subsection 4.1.1), explore case studies using n-gram usage rates from Twitter, which are broadly aggregated by language communities. These studies primarily use n-gram usage rates as proxies of collective attention.

Second, a set of papers based on n-gram usage rate time series subset into location-based communities with user provided location data matched to US states. The additional spatial information allowed us to study more location specific phenomona, from community level stress, sleep pattern changes around the spring change to daylight savings time, and state-level estimates of homelessness rates.

Finally, I present two studies that introduce novel methods for text data, created to solve problems as we explored the potential of social media datasets.

4.1 Collective Attention and n-gram Usage Rate Studies

In the following section we explore a selection of works that use Twitter derived n-gram usage rates as a proxy for collective attention paid towards topics of interest. The Storywrangler paper introduced this dataset to the world, which enabled external researchers and the general public to explore popularity-weighted word usage rates, without needing to access and process hundreds of billions of tweets.

Further studies examined this dataset's immense breadth and depth. We studied how broad patterns of online interaction transitioned from originally authored text to amplification through retweets becoming a dominant behavior across dozens of the most used language communities. We conducted in-depth case studies on a wide range of topics, from pandemics to politics.

As valuable as many of these collected contributions are, certain limitations exist within these studies. Because usage rates have already been aggregated by language group in the Storywrangler dataset, there is no opportunity to subset the n-gram usage rates, into smaller communities of interest. Even more problematic is the inability to easily assess the quality of an n-gram usage rate as a proxy of attention for some topic of interest. Further work (and computational hardware) was needed to address these short-comings when researchers find that a Storywrangler proxy is ill-suited for their needs, whether due to polysemy in a keyword of interest or issues of a non-representative sample.

Despite these limitations, social media derived n-gram time series can sometimes provide an immensely rich window into the dynamics of online attention. Day-level aggregation allowed us to track the emergence of nationally impactful events with high temporal resolution. The relative accessibility of the data allows for rapid, inter-disciplinary exploration of a wide variety of topics, as demonstrated by the collection of co-authors in the following section.

4.1.1 Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter

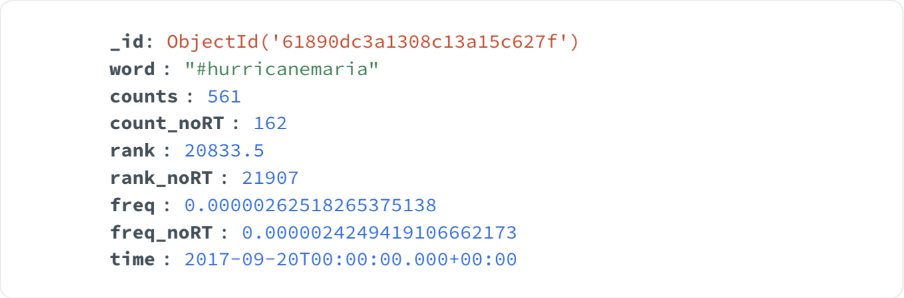
The first paper is *Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter* by Thayer Alshabbi, Jane L. Adams, Michael V. Arnold, Joshua R. Minot, David R. Dewhurst, Andrew J. Reagan, Chrisopher M. Danforth, and Peter Sheridan Dodds, cited as [8].

Abstract

In real time, Twitter strongly imprints world events, popular culture, and the day-to-day, recording an ever-growing compendium of language change. Vitally, and absent from many standard corpora such as books and news archives, Twitter also encodes popularity and spreading through retweets. Here, we describe Storywrangler, an ongoing curation of over 100 billion tweets containing 1 trillion 1-grams from 2008 to 2021. For each day, we break tweets into 1-, 2-, and 3-grams across 100+ languages, generating frequencies for words, hashtags, handles, numerals, symbols, and emojis. We make the dataset available through an interactive time series viewer and as downloadable time series and daily distributions. Although Storywrangler leverages Twitter data, our method of tracking dynamic changes in n-grams can be extended to any temporally evolving corpus. Illustrating the instrument's potential, we present example use cases including social amplification, the sociotechnical dynamics of famous individuals, box office success, and social unrest.

Contribution

For this paper I contributed to the design of the backend for our n-gram database, which is publicly accessible on the storywrangler website. I contributed to designing the storywrangler n-gram parser to capture Twitter-specific tokens of interest, such as hashtags and user handles. We also engineered counters to separately count tokens that were originally



```
_id: ObjectId('61890dc3a1308c13a15c627f')
word : "#hurricanemaria"
counts : 561
count_noRT : 162
rank : 20833.5
rank_noRT : 21907
freq : 0.00000262518265375138
freq_noRT : 0.0000024249419106662173
time : 2017-09-20T00:00:00.000+00:00
```

Figure 4.1: An example document showing a 'storyon' counter objected represented within the database. For each language, we store a collection of n-gram counters. By querying for `word` *we can assemble word usage rate time series. Querying for* `time`*, we can assemble a daily Zipf distribution. More complex queries are also enabled; future studies could query by* `rank` *to study the emergence of slang or other new types entering a language.*

authored from those with social amplification through retweets. This involved engineering a schema that enabled queries for both n-gram time series and daily distribution queries, for 1-grams, 2-grams, and 3-grams at daily resolution with separate collections for over 100 languages. Flexibility is prioritized with this schema, but fast responses are enabled by building indexes. When data streams were still incoming and insert performance was a concern, we avoided building unnecessary indexes. Currently, that cost-benefit equilibrium may have shifted, where it makes sense to pre-compute any likely to be used index.

I researched and directed the purchase of appropriately sized hardware to enable interactive queries, and administered both hardware, databases, and insert software to ensure continuing daily updates with minimal insert times, while we retained access to Twitter's decahose feed. After the team collectively decided to preserve case-sensitivity when counting n-grams, I chose our rank truncation threshold to be one million $(10{*}{*}6)$, set to ensure daily inserts across languages could execute comfortably within 24 hours, given some variance in the number of tweets per day.

I helped to visualize the n-gram time series plots, exploring potential case-studies for further exploration. Additionally, I worked with Thayer Alshaabi to extend this work to a

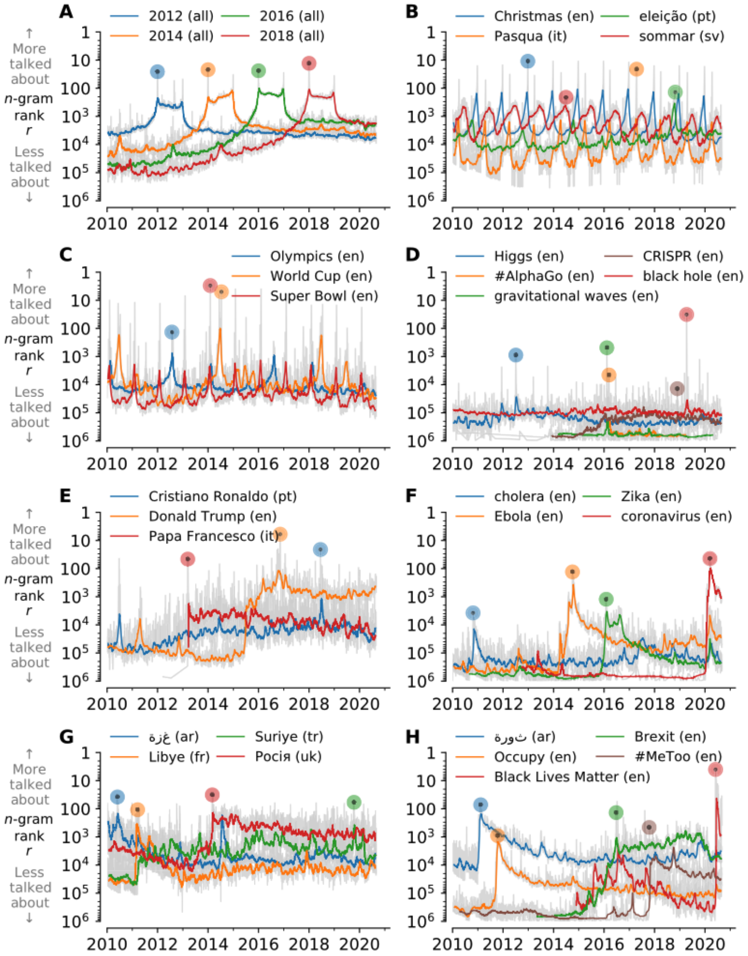
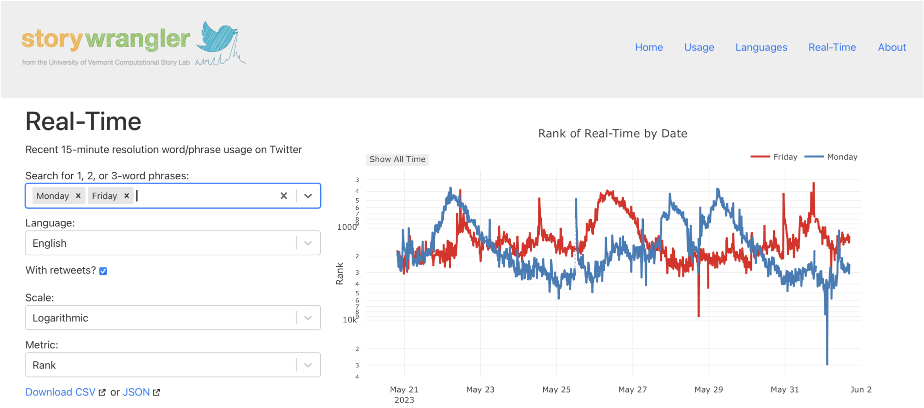


Figure 4.2: Reprint of Figure 1 from [8], with caption as follows: "For each n-gram, we display daily rank in gray overlaid by a centered monthly rolling average (colored lines), and highlight the n-gram's overall highest rank with a solid disk. ***A.*** *Anticipation and memory of calendar years for all of Twitter.* ***B.*** *Annual and periodic events: Christmas in English (blue), Easter in Italian (orange), election in Portuguese (green), and summer in Swedish (red).* ***C.*** *Attention around international sports in English: Olympics (blue), FIFA world cup (orange), and Super Bowl (red).* ***D.*** *Major scientific discoveries and technological innovations in English.* ***E.*** *Three famous individuals in relevant languages: Ronaldo (Portuguese), Trump (English), and Pope Francis (Italian).* ***F.*** *Major infectious disease outbreaks.* ***G.*** *Conflicts: Gaza in Arabic (blue), Libya in French (orange), Syria in Turkish (green), and Russia in Ukrainian (red).* ***H.*** *Protest and movements: Arab Spring (Arabic word for 'revolution', blue), Occupy movement (English, orange), Brexit campaign (English, green), #MeToo movement (English, brown), and Black Lives Matter protests (English, red)."*



Real-Time

Recent 15-minute resolution word/phrase usage on Twitter

Search for 1, 2, or 3-word phrases: Monday × Friday ×

Language: English

With retweets? ☑

Scale: Logarithmic

Metric: Rank

Download CSV or JSON

Figure 4.3: Screenshot of the realtime, 15 minute resolution n-gram viewer available on the storywrangler website. The database is no longer realtime, due to the end of our data sharing agreement with Twitter, but the final two weeks of 15 minute resolution data, from May 20th, 2023 to June 1st, 2023, remains publicly available.

realtime, 15-minute resolution n-gram viewer, as shown in Figure 4.3.

4.1.2 How the worlds collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter

Paper number two is *How the worlds collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter* by Thayer Alshaabi, Michael V. Arnold, Joshua R. Minot, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, Roby Muhamad, Christopher M. Danforth, and Peter Sheridan Dodds, cited as [9].

Abstract

In confronting the global spread of the coronavirus disease COVID-19 pandemic we must have coordinated medical, operational, and political responses. In all efforts, data is crucial. Fundamentally, and in the possible absence of a vaccine for 12 to

18 months, we need universal, well-documented testing for both the presence of the disease as well as confirmed recovery through serological tests for antibodies, and we need to track major socioeconomic indices. But we also need auxiliary data of all kinds, including data related to how populations are talking about the unfolding pandemic through news and stories. To in part help on the social media side, we curate a set of 2000 day-scale time series of 1- and 2-grams across 24 languages on Twitter that are most 'important' for April 2020 with respect to April 2019. We determine importance through our allotaxonometric instrument, rank-turbulence divergence. We make some basic observations about some of the time series, including a comparison to numbers of confirmed deaths due to COVID-19 over time. We broadly observe across all languages a peak for the language-specific word for 'virus' in January 2020 followed by a decline through February and then a surge through March and April. The world's collective attention dropped away while the virus spread out from China. We host the time series on Gitlab, updating them on a daily basis while relevant. Our main intent is for other researchers to use these time series to enhance whatever analyses that may be of use during the pandemic as well as for retrospective investigations.

Contribution

For this paper, I worked closely with Thayer Alshaabi to compile lists of terms potentially relevant to the emerging COVID-19 pandemic by measuring rank-divergences between year-seperated, day-scale 1-, 2-, and 3-gram Zipf distributions of Twitter. We discussed how to create a robust measurement to select emerging words, and decided on averaging rank divergence contributions by n-gram over a month long study period. Rank divergence was computed on the full daily Zipf distributions, rather than the truncated daily distributions stored in the database. The daily zipf distribution for each date in 2020 was compared to the same calendar date in 2019. I repeated these measurements for tweets written in each of the top languages on Twitter as classified by FastText [81].

After the lists of top contributing words were compiled, I built database indexes for each language to enable fast n-gram queries, and wrote and diagnosed queries with Thayer to collect n-gram time series beginning from September 1, 2019. We to run daily updates to share publicly online so external researchers could have up-to-date data as the pandemic continued to evolve.

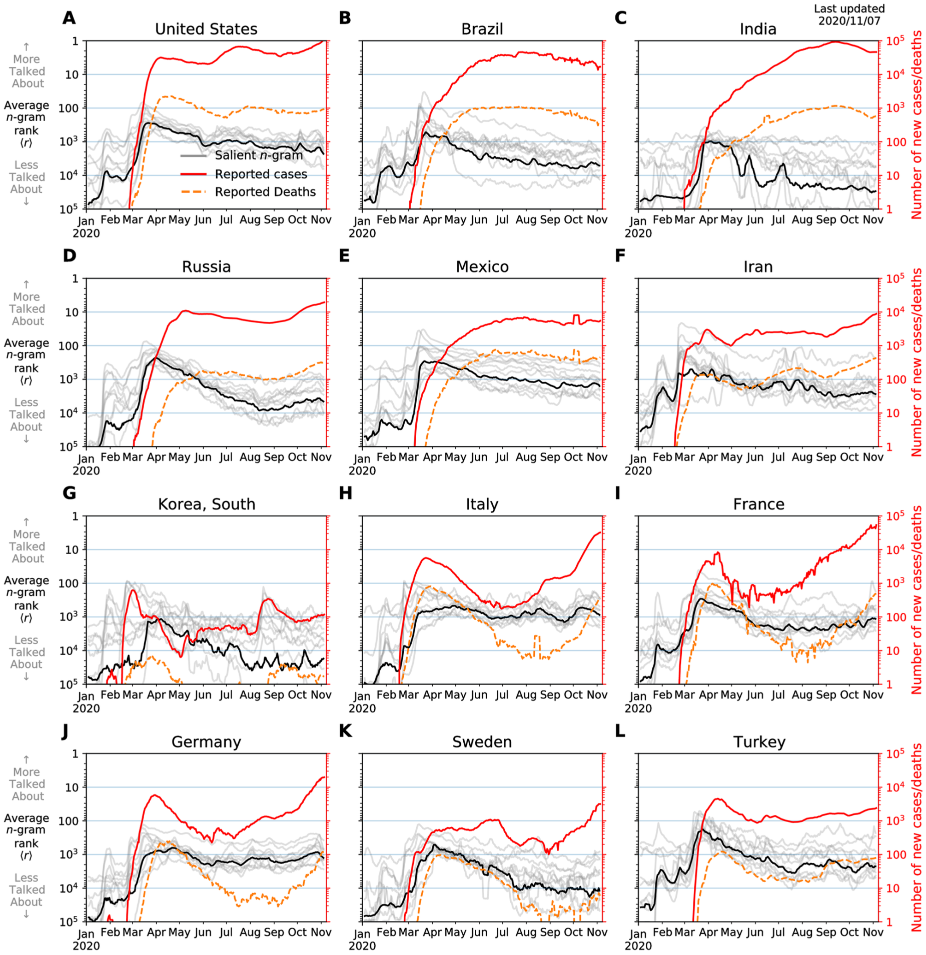
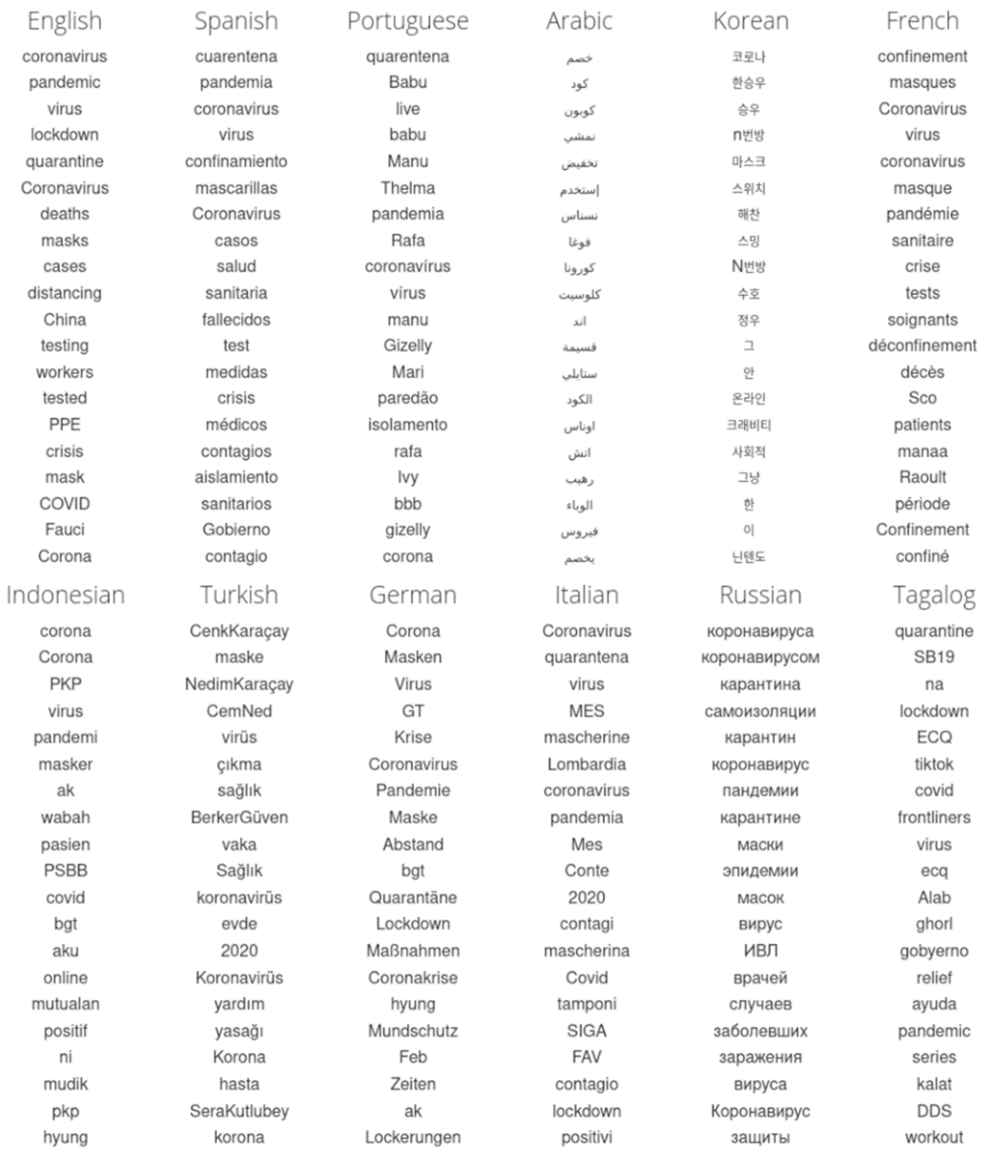


Figure 4.4: Reprint of Figure 6 from [9], with caption as follows: "Time series for daily reported case loads and death compared with a list of 10 salient 1-grams for the top language spoken in each country. For each n-gram, we display a weekly rolling average of usage ranks at the day scale in gray overlaid by an average of all these 1-grams in black marking their corresponding ranks using the left vertical axis. Similarly, we use the right vertical axis to display a weekly rolling average of daily new cases (red solid-line), and reported new deaths (orange dashed-line)."



English	Spanish	Portuguese	Arabic	Korean	French
coronavirus	cuarentena	quarentena	خصم	코로나	confinement
pandemic	pandemia	Babu	كود	한승우	masques
virus	coronavirus	live	كوبون	승우	Coronavirus
lockdown	virus	babu	نمشي	n번방	virus
quarantine	confinamiento	Manu	تخفيض	마스크	coronavirus
Coronavirus	mascarillas	Thelma	إستخدم	스위치	masque
deaths	Coronavirus	pandemia	نسناس	해찬	pandémie
masks	casos	Rafa	فوغا	스밍	sanitaire
cases	salud	coronavírus	كورونا	N번방	crise
distancing	sanitaria	vírus	كلوسيت	수호	tests
China	fallecidos	manu	اند	정우	soignants
testing	test	Gizelly	قسيمة	그	déconfinement
workers	medidas	Mari	ستايلي	안	décès
tested	crisis	paredão	الكود	온라인	Sco
PPE	médicos	isolamento	اوناس	크래비티	patients
crisis	contagios	rafa	انش	사회적	manaa
mask	aislamiento	Ivy	رهيب	그냥	Raoult
COVID	sanitarios	bbb	الوباء	한	période
Fauci	Gobierno	gizelly	فيروس	이	Confinement
Corona	contagio	corona	يخصم	닌텐도	confiné

Indonesian	Turkish	German	Italian	Russian	Tagalog
corona	CenkKaraçay	Corona	Coronavirus	коронавируса	quarantine
Corona	maske	Masken	quarantena	коронавирусом	SB19
PKP	NedimKaraçay	Virus	virus	карантина	na
virus	CemNed	GT	MES	самоизоляции	lockdown
pandemi	virüs	Krise	mascherine	карантин	ECQ
masker	çıkma	Coronavirus	Lombardia	коронавирус	tiktok
ak	sağlık	Pandemie	coronavirus	пандемии	covid
wabah	BerkerGüven	Maske	pandemia	карантине	frontliners
pasien	vaka	Abstand	Mes	маски	virus
PSBB	Sağlık	bgt	Conte	эпидемии	ecq
covid	koronavirüs	Quarantäne	2020	масок	Alab
bgt	evde	Lockdown	contagi	вирус	ghorl
aku	2020	Maßnahmen	mascherina	ИВЛ	gobyerno
online	Koronavirüs	Coronakrise	Covid	врачей	relief
mutualan	yardım	hyung	tamponi	случаев	ayuda
positif	yasağı	Mundschutz	SIGA	заболевших	pandemic
ni	Korona	Feb	FAV	заражения	series
mudik	hasta	Zeiten	contagio	вируса	kalat
pkp	SeraKutlubey	ak	lockdown	Коронавирус	DDS
hyung	korona	Lockerungen	positivi	защиты	workout

Figure 4.5: Reprint of Figure 2 from [9], with caption as follows: "Top 20 (of 1,000) 1-grams for our top 12 languages for the first three weeks of April 2020 relative to a year earlier. Our intent is to capture 1-grams that are topically and culturally important during the COVID-19 pandemic. While overall, we see pandemic-related words dominate the lists across languages, we also find considerable specific variation. Words for virus, quarantine, protective equipment, and testing show different orderings (note that we do not employ stemming). Unrelated 1-grams but important to the time of April 2020 are in evidence; the balance of these are important for our understanding of how much the pandemic is being talked about."

4.1.3 Divergent modes of online collective attention to the COVID-19 pandemic are associated with future caseload variance

Paper number three is *Divergent modes of online collective attention to the COVID-19 pandemic are associated with future caseload variance* by David Rushing Dewhurst, Thayer Alshaabi, Michael V. Arnold, Joshua R. Minot, Christopher M. Danforth, Peter Sheridan Dodds, citied as [36].

Abstract

Using a random 10% sample of tweets authored from 2019-09-01 through 2020-04-30, we analyze the dynamic behavior of words (1-grams) used on Twitter to describe the ongoing COVID-19 pandemic. Across 24 languages, we find two distinct dynamic regimes: One characterizing the rise and subsequent collapse in collective attention to the initial Coronavirus outbreak in late January, and a second that represents March COVID-19-related discourse. Aggregating countries by dominant language use, we find that volatility in the first dynamic regime is associated with future volatility in new cases of COVID-19 roughly three weeks (average 22.49 ± 3.26 days) later. Our results suggest that surveillance of change in usage of epidemiology-related words on social media may be useful in forecasting later change in disease case numbers, but we emphasize that our current findings are not causal or necessarily predictive.

Contribution

My primary contribution for this paper was curating the 1-gram usage time series related to COVID-19 that were analyzed in this study. Additionally, I participated in discussions regarding the interpretations of the resulting time series clusters, shown in Figure 4.6.

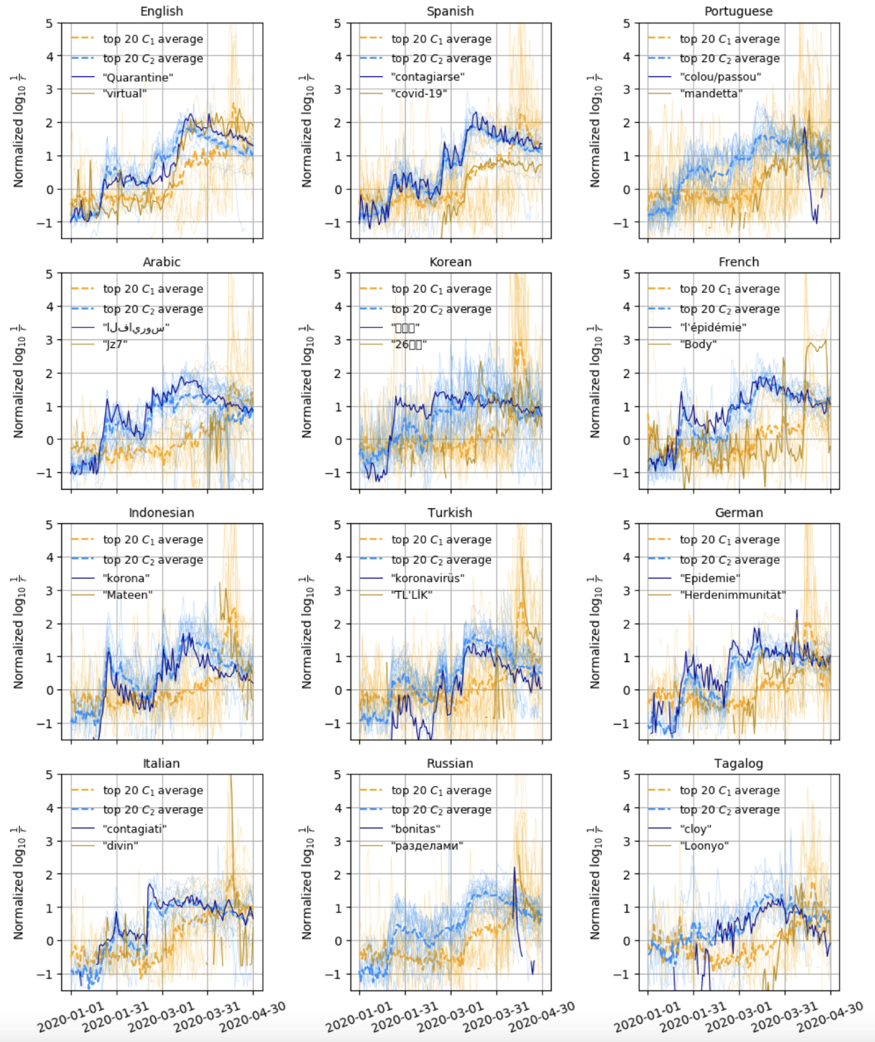


Figure 4.6: Figure 2 from [36], with caption as follows: "We display the mean normalized log rank timeseries of the top 20 words closest to each of $E[C_1]$ and $E[C_2]$ in dashed curves and the single word closest to each of $E[C_1]$ and $E[C_2]$ in thin solid curves for each of the first 12 of 24 languages. The divergent modes of dynamic behavior are consistent across most languages, with some languages (English, French, German, and Indonesian) displaying prominently larger peaks in words closest to $E[C_2]$ during late January through early February 2020. Other languages, such as Korean and Tagalog, do not display this behavior."

4.1.4 Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter

Paper number four is *Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter* by Joshua R Minot, Michael V. Arnold, Thayer Alshaabi, Christopher M. Danforth, and Peter Sheridan Dodds, cited as [115].

Abstract

The past decade has witnessed a marked increase in the use of social media by politicians, most notably exemplified by the 45th President of the United States (POTUS), Donald Trump. On Twitter, POTUS messages consistently attract high levels of engagement as measured by likes, retweets, and replies. Here, we quantify the balance of these activities, also known as “ratios”, and study their dynamics as a proxy for collective political engagement in response to presidential communications. We find that raw activity counts increase during the period leading up to the 2016 election, accompanied by a regime change in the ratio of retweets-to-replies connected to the transition between campaigning and governing. For the Trump account, we find words related to fake news and the Mueller inquiry are more common in tweets with a high number of replies relative to retweets. Finally, we find that Barack Obama consistently received a higher retweet-to-reply ratio than Donald Trump. These results suggest Trump’s Twitter posts are more often controversial and subject to enduring engagement as a given news cycle unfolds.

Contribution

My contribution to this paper began as a class project for Principles of Complex Systems, where Sarah Howerter and I scraped tweets to visualize tweets in ternary plots, where the three axes represented the ratios of likes, retweets, and replies. We measured happiness scores using labMT, and examined this distribution in ternary space. Later, I worked with Josh Minot and Thayer Alshaabi to store relevant political tweets on our databases after we found that searching the decahose for retweets could provide multiple snapshots of engagement. I also assisted in designing the visualizations for the paper.

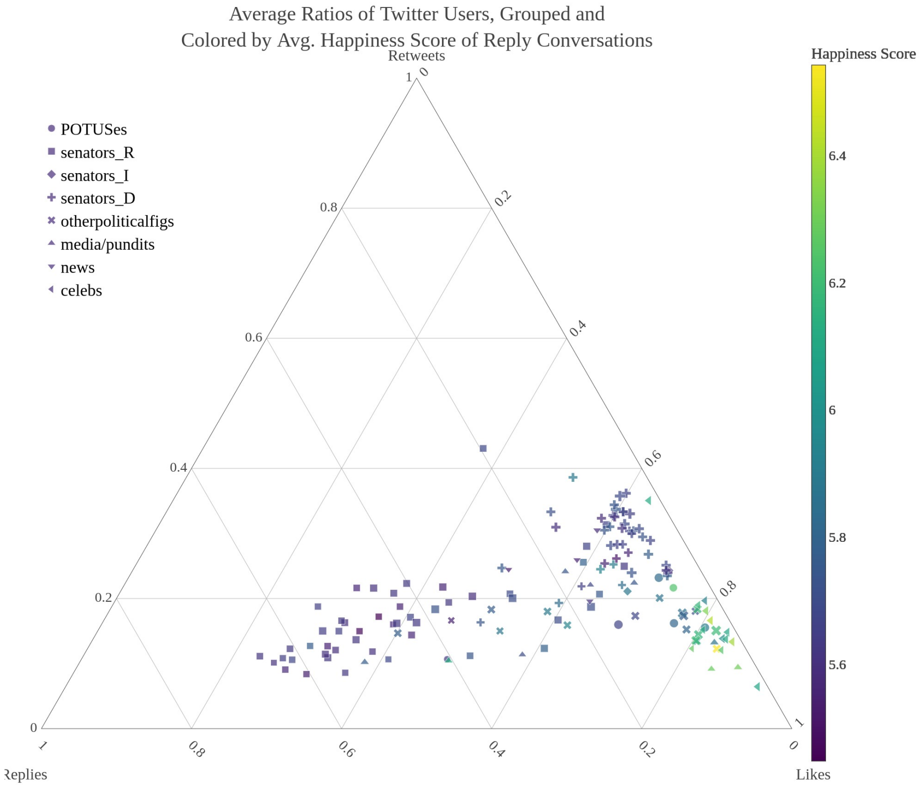


Figure 4.7: Figure from preliminary class project, showing average happiness values of reply threads averaged by user and represented with markers related to account type. We found that tweets with a lower ratio of replies to likes were more likely to have higher happiness scores as measured by the LabMT sentiment lexicon.

4.1.5 Twitter misogyny associated with Hillary Clinton increased throughout the 2016 U.S. election campaign

Paper number eight *Twitter misogyny associated with Hillary Clinton increased throughout the 2016 U.S. election campaign* by Morgan Weaving, Thayer Alshaabi, Michael V. Arnold, Khandis Blake, Christopher M. Danforth, Peter S. Dodds, Nick Haslam and Cordelia Fine, cited as [163].

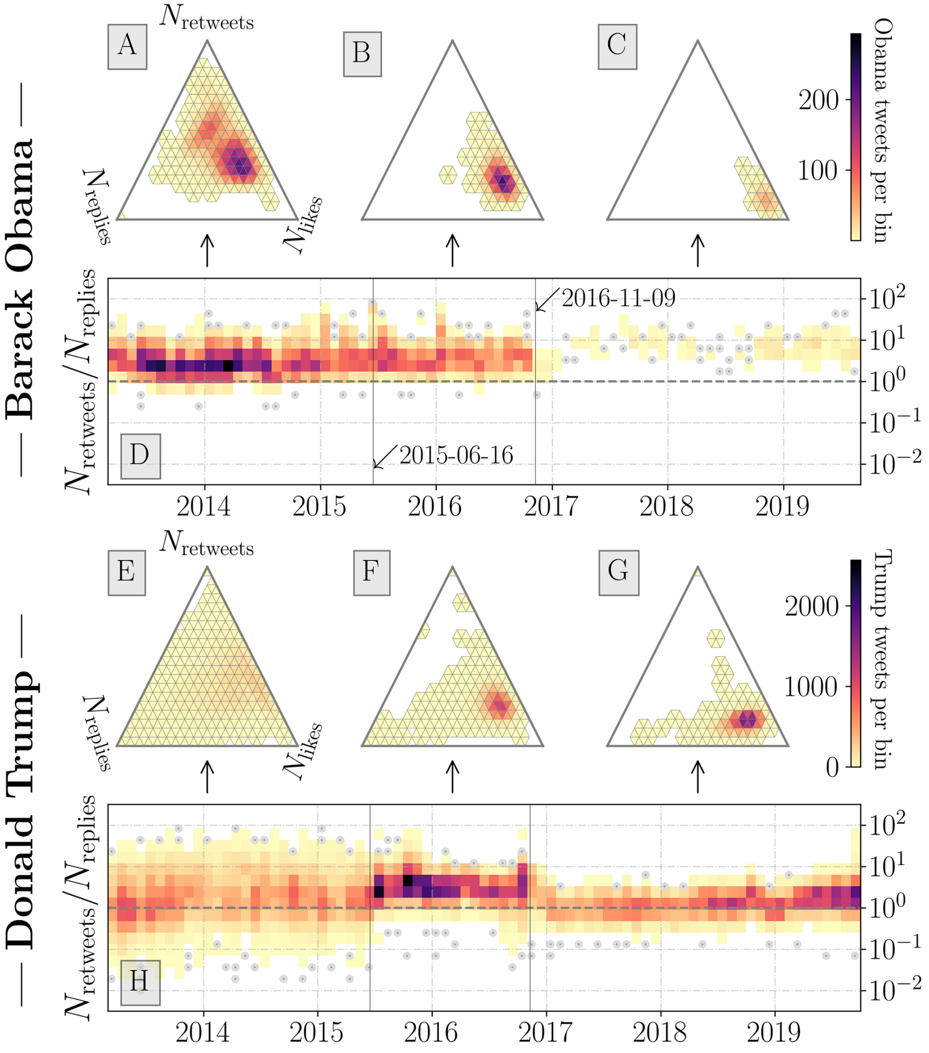


Figure 4.8: Reprint of Figure 2 from [115], with caption as follows: Ternary histograms and $N_{retweets}/N_{replies}$ ratio time series for the ***@BarackObama*** *(A–D) and* ***@realDonaldTrump*** *(E–H) Twitter accounts. The ternary histograms (A–C and E–H) represent the count of retweet, favorite, and reply activities normalized by the sum of all activities. White regions indicate no observations over the given time period. See Fig 4 for examples of full time series for response activity for example tweets. Heatmap time series (D and H) consist of monthly bins representing the density of tweets with a given ratio value. Single observations (bin counts <2) are represented by grey points. The two dates annotated correspond to the date of Trump's declaration of candidacy (2015–05–16) and the 2016 general election (2016–11–09). We show the tendency for Trump tweets to have ternary ratio values with a greater reply component—with pre-candidacy tweets having higher variability and pre-election tweets having a higher $N_{retweets}/N_{replies}$ ratio value. Post-election Obama tweets have ternary ratio values with more likes than other periods for both Obama and Trump.*

Abstract

Online misogyny has become a fixture in female politicians' lives. Backlash theory suggests that it may represent a threat response prompted by female politicians' counterstereotypical, power-seeking behaviors. We investigated this hypothesis by analyzing Twitter references to Hillary Clinton before, during, and after her presidential campaign. We collected a corpus of over 9 million tweets from 2014 to 2018 that referred to Hillary Clinton, and employed an interrupted time series analysis on the relative frequency of misogynistic language within the corpus. Prior to 2015, the level of misogyny associated with Clinton decreased over time, but this trend reversed when she announced her presidential campaign. During the campaign, misogyny steadily increased and only plateaued after the election, when the threat of her electoral success had subsided. These findings are consistent with the notion that online misogyny towards female political nominees is a form of backlash prompted by their ambition for power in the political arena.

Contribution

For this paper I participated in discussions with Morgan Weaving and Thayer Alshaabi to determine what social media data would be useful in measuring misogynistic language surrounding female candidates on Twitter. I created exploratory figures, including ambient sentiment plots of tweets mentioning the keywords Clinton, Trump and Biden. I worked to parse tweets match a list of keywords relevant to Hillary Clinton into n-grams and assisted in creating an indicator of misogyny defined by the frequency of misogynistic terms being used within this corpus.

4.1.6 Fame and Ultrafame: Measuring and comparing daily levels of ‘being talked about’ for United States’ presidents, their rivals, God, countries, and K-pop.

Paper number nine is *Fame and Ultrafame: Measuring and comparing daily levels of ‘being talked about’ for United States’ presidents, their rivals, God, countries, and K-pop* by Peter Sheridan Dodds, Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, Christopher M. Danforth, cited as [41].

Abstract

When building a global brand of any kind – a political actor, clothing style, or belief system – developing widespread awareness is a primary goal. Short of knowing any of the stories or products of a brand, being talked about in whatever fashion – raw fame – is, as Oscar Wilde would have it, better than not being talked about at all. Here, we measure, examine, and contrast the day-to-day raw fame dynamics on Twitter for US Presidents and major US Presidential candidates from 2008 to 2020: Barack Obama, John McCain, Mitt Romney, Hillary Clinton, Donald Trump, and Joe Biden. We assign “lexical fame” to be the number and (Zipfian) rank of the (lowercased) mentions made for each individual across all languages. We show that all five political figures have at some point reached extraordinary volume levels of what we define to be “lexical ultrafame”: An overall rank of approximately 300 or less which is largely the realm of function words and demarcated by the highly stable rank of ‘god’. By this measure, ‘trump’ has become enduringly ultrafamous, from the 2016 election on. We use typical ranks for country names and function words as standards to improve perception of scale. We quantify relative fame rates and find that in the eight weeks leading up the 2008 and 2012 elections, ‘obama’ held a 1000:757 volume ratio over ‘mccain’ and 1000:892 over ‘romney’, well short of the 1000:544 and 1000:504 volumes favoring ‘trump’ over ‘hillary’ and ‘biden’ in the 8 weeks leading up to the 2016 and 2020 elections. Finally, we track how only one other entity has more sustained ultrafame than ‘trump’ on Twitter:

The K-pop (Korean pop) band BTS. We chart the dramatic rise of BTS, finding their Twitter handle '@bts_twt' has been able to compete with 'a' and 'the'. Our findings for BTS more generally point to K-pop's growing economic, social, and political power.

Contribution

For this paper I consulted with the co-authors as part of the larger umbrella of Storywrangler related projects, joining discussions and giving feedback as the figures were developed. I contributed in cleaning and parsing tweets into n-grams by language, and creating storage solutions for the resulting 21TB dataset to enable further analysis. We were recently informed by the Journal of Quantitative Description: Digital Media that this paper received their inaugural award for best paper ever published by the journal: "Best Paper Engaged in Quantitative Description on an Under-studied Phenomenon."

4.1.7 Say their names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd

Paper number thirteen is *Say their names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd* by Henry H. Wu, Ryan J. Gallagher, Thayer Alshaabi, Jane L. Adams, Joshua R. Minot, Michael V. Arnold, Brooke Foucault Welles, Randall Harp, Peter Sheridan Dodds, Christopher M. Danforth, cited as [174].

Abstract

The murder of George Floyd by police in May 2020 sparked international protests and brought unparalleled levels of attention to the Black Lives Matter movement. As we show, his death set record levels of activity and amplification on Twitter, prompted the saddest day in the platform's history, and caused his name to appear among the

ten most frequently used phrases in a day, where he is the only individual to have ever received that level of attention who was not known to the public earlier that same week. Importantly, we find that the Black Lives Matter movement's rhetorical strategy to connect and repeat the names of past Black victims of police violence—foregrounding racial injustice as an ongoing pattern rather than a singular event—was exceptionally effective following George Floyd's death: attention given to him extended to over 185 prior Black victims, more than other past moments in the movement's history. We contextualize this rising tide of attention among 12 years of racial justice activism on Twitter, demonstrating how activists and allies have used attention and amplification as a recurring tactic to lift and memorialize the names of Black victims of police violence. Our results show how the Black Lives Matter movement uses social media to center past instances of police violence at an unprecedented scale and speed, while still advancing the racial justice movement's longstanding goal to "say their names."

Contribution

I contributed to this paper mostly in the exploration phase. I created ambient sentiment plots with our tweet subsamples, but these were low resolutions and not high enough quality to be published. I advised Henry Wu about some of the tools we had created to quantify spikes in attention, but ultimately advised him against pursuing this direction.

4.1.8 The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020

Paper number eleven is *The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020* by Thayer Alshaabi, David Rushing Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane L. Adams, Christopher M. Danforth and Peter Sheridan Dodds, cited as [10].

Abstract

Working from a dataset of 118 billion messages running from the start of 2009 to the end of 2019, we identify and explore the relative daily use of over 150 languages on Twitter. We find that eight languages comprise 80% of all tweets, with English, Japanese, Spanish, Arabic, and Portuguese being the most dominant. To quantify social spreading in each language over time, we compute the 'contagion ratio': The balance of retweets to organic messages. We find that for the most common languages on Twitter there is a growing tendency, though not universal, to retweet rather than share new content. By the end of 2019, the contagion ratios for half of the top 30 languages, including English and Spanish, had reached above 1—the naive contagion threshold. In 2019, the top 5 languages with the highest average daily ratios were, in order, Thai (7.3), Hindi, Tamil, Urdu, and Catalan, while the bottom 5 were Russian, Swedish, Esperanto, Cebuano, and Finnish (0.26). Further, we show that over time, the contagion ratios for most common languages are growing more strongly than those of rare languages.

Contribution

I collaborated with the paper's coauthors to conceptualize the retweet balance measurement. I contributed in creating the paper's primary dataset, Storywrangler, which predicts tweet language and parses and counts n-grams at daily resolution over the study period. Additionally, I designed a database schema and inserted a language database, which stores metadata on language distributions by language for each day for further analysis, including the number of tweets, n-gram tokens, unique speakers, unique n-gram types, for both Twitter's changing language identification algorithm, and a consistently applied language classifier, FastText [81].

4.2 Location-based Social Media Studies

In this section we look at a collection of papers that build on the prior Storywrangler parser but leverage user provided location metadata to infer locations within US cities. This additional location data allowed us to build state-level proxies of happiness, sleep behavior based on aggregated user activity patterns, and homelessness.

While having additional metadata was helpful, continuing to subset the data into smaller and smaller communities led to challenges. For the study of homelessness, a phenomenon which impacts less than 1% of the US population at any given time, we found we had a limited number of tweets after both searching for keywords and grouping by geography. For any relatively uncommon or under-discussed phenomena, these challenges would likely emerge.

This state level twitter dataset is still accessible and I believe it holds great potential. States are often referred to 'Laboratories of Democracy', and enabling being able to compare the spectrum of policies with population level data originating from the populations governed should be valuable [1].

4.2.1 Expecting the Unexpected: Predicting Panic Attacks from Mood and Twitter

Paper number five is *Expecting the Unexpected: Predicting Panic Attacks from Mood and Twitter* by Ellen W. McGinnis, Bryn Loftness, Shania Lunna, Isabel Berman, Skylar Bagdon, Genevieve Lewis, Michael Arnold, Christopher M. Danforth, Peter S. Dodds, Matthew Price, William E. Copeland and Ryan S. McGinnis, cited as [106].

Abstract

Panic attacks are an impairing mental health problem that affects about one in 10 US adults every year. Current DSM criteria describe panic attacks as unexpected, occur-

> ring without warning or triggering events. The unexpected nature of panic attacks not only leads to increased anxiety for the individual but has also made panic attacks particularly challenging to study. However, recent evidence suggests that individuals who experience such attacks could identify attack triggers. We aimed to explore both retrospectively and prospectively, qualitative, and quantitative factors associated with the onset of panic attacks. We remotely recruited a diverse sample of 87 individuals who regularly experienced panic attacks from 30 states in the US. Participants responded to daily questions relating to their panic attacks and wellness behaviors each day for 28 days. We also considered daily community level factors captured by the Hedonometer, a metric which estimates population-level happiness daily using a random 10% of all public tweets. Consistent with our prior work, most participants (95%) were able to retrospectively identify a trigger for their attack. Worse individual mood was associated with greater likelihood of experiencing a same-day panic attack over and above other individual wellness factors. Worse individually reported mood and state-based population level mood as indicated by the Hedonometer were associated with greater likelihood of next-day panic attack. These promising results suggest that individuals who experience panic attacks may be able to expect the unexpected. The importance of individual and state-based population level mood in panic attack risk could be used to ultimately inform future prevention and intervention efforts.

Contribution

For this paper I curated a dataset of daily, state-level Twitter sentiment time series, based on location inferred from user provided text biography fields. Additionally, I created exploratory static and video visualizations to show the spatial change in sentiment over time.

4.2.2 The sleep loss insult of Spring Daylight Savings in the US is observable in Twitter activity

Paper number six, *The sleep loss insult of Spring Daylight Savings in the US is observable in Twitter activity* by Kelsey Linnell, Michael Arnold, Thayer Alshaabi, Thomas McAndrew,

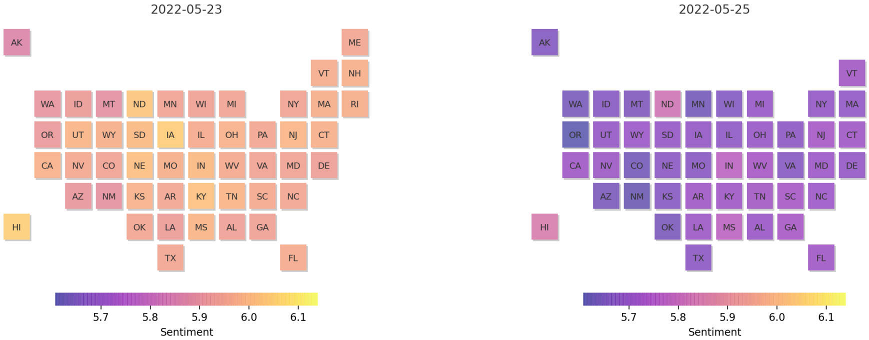


Figure 4.9: State sentiment maps for two selected days in the study period. May 23, 2022 was the day prior to the Uvalde, TX school shooting, and May 25, the day following the shooting.

Jeanie Lim, Peter Sheridan Dodds, and Christopher M. Danforth, cited as [97].

Abstract

Sleep loss has been linked to heart disease, diabetes, cancer, and an increase in accidents, all of which are among the leading causes of death in the United States. Population-scale sleep studies have the potential to advance public health by helping to identify at-risk populations, changes in collective sleep patterns, and to inform policy change. Prior research suggests other kinds of health indicators such as depression and obesity can be estimated using social media activity. However, the inability to effectively measure collective sleep with publicly available data has limited large-scale academic studies. Here, we investigate the passive estimation of sleep loss through a proxy analysis of Twitter activity profiles. We use “Spring Forward” events, which occur at the beginning of Daylight Savings Time in the United States, as a natural experimental condition to estimate spatial differences in sleep loss across the United States. On average, peak Twitter activity occurs 15 to 30 min later on the Sunday following Spring Forward. By Monday morning however, activity curves are realigned with the week before, suggesting that the window of sleep opportunity is compressed in Twitter data, revealing Spring Forward behavioral change.

Contribution

For this paper, I worked with Kelsey Linnell to test a user location classification tool both for accuracy and performance characteristics. With extracted city and state metadata, we were able to store tweets with identifiable user locations into an indexed database for further analysis. Around 5% of all tweets had an identifiable US location with our method, which was 12% of English language tweets.

4.2.3 An assessment of measuring local levels of homelessness through proxy social media signals

Paper number seven *An assessment of measuring local levels of homelessness through proxy social media signals* by Yoshi Meke Bird, Sarah E. Grobe, Michael V. Arnold, Sean P. Rogers, Mikaela I. Fudolig, Julia Witte Zimmerman, Christopher M. Danforth, Peter Sheridan Dodds, cited as [17].

Abstract

Recent studies suggest social media activity can function as a proxy for measures of state-level public health, detectable through natural language processing. We present results of our efforts to apply this approach to estimate homelessness at the state level throughout the US during the period 2010-2019 and 2022 using a dataset of roughly 1 million geotagged tweets containing the substring “homeless.” Correlations between homelessness-related tweet counts and ranked per capita homelessness volume, but not general-population densities, suggest a relationship between the likelihood of Twitter users to personally encounter or observe homelessness in their everyday lives and their likelihood to communicate about it online. An increase to the log-odds of “homeless” appearing in an English-language tweet, as well as an acceleration in the increase in average tweet sentiment, suggest that tweets about homelessness are also affected by trends at the nation-scale. Additionally, changes to the lexical content of tweets over

time suggest that reversals to the polarity of national or state-level trends may be detectable through an increase in political or service-sector language over the semantics of charity or direct appeals. An analysis of user account type also revealed changes to Twitter-use patterns by accounts authored by individuals versus entities that may provide an additional signal to confirm changes to homelessness density in a given jurisdiction. While a computational approach to social media analysis may provide a low-cost, real-time dataset rich with information about nationwide and localized impacts of homelessness and homelessness policy, we find that practical issues abound, limiting the potential of social media as a proxy to complement other measures of homelessness.

Contribution

For this paper I consulted with the primary author, Yoshi Bird, to conceptualize how social media data could be leveraged to estimate Homelessness rates, and generate text measurements like sentiment, and distributional comparisons using rank-turbulence divergence.

To further refine the corpus of homelessness related tweets, I met with Yoshi to discuss labeling training data for a supervised classification task. After Yoshi provided labeled data I trained a classifier to label tweets that were strictly addressing literal human homelessness, as opposed to a wide range of different usages, from abandoned pets to usages as a hyperbolic adjective.

Additionally, I helped to provide Twitter data matching relevant keywords with US state tags based on user-provided text.

4.3 Novel Methods for Social Media Data

In this section, we present two studies presenting methods for social media derived data. Over the course of our exploration of Twitter data, we often found that existing methods were not quite sufficient to answer a question of interest. In the following section are two examples. First is a similarity search method built to extract time-series with interesting

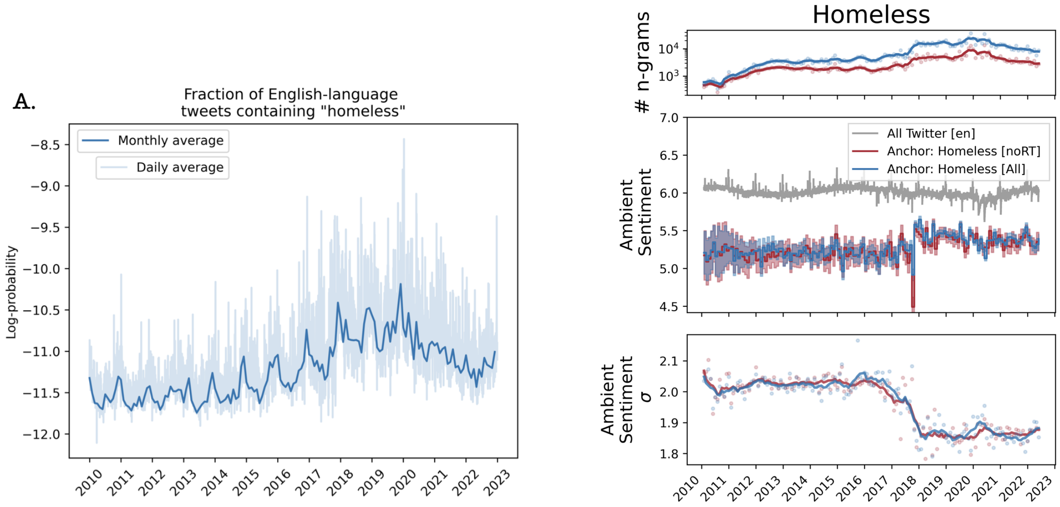


Figure 4.10: Measures of collective attention and sentiment for US tweets containing homelessness from the paper.

dynamics, specifically the kinds of dramatic shifts in attention observed in sociotechnical systems. Second is a method to expand semantic lexicons to unrated words by leveraging pre-trained word embeddings. This was a capability that we desired after seeing extremely negative words like ‘pandemic’ rise in popular usage, and recognizing that our lexicons needed to continue to be updated to reflect current language.

While these studies are quite dissimilar, dealing with time-series methods and natural language processing, respectively, the type of method work they represent is an important component of this dissertation and the broader work of the Computational Story lab.

4.3.1 The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series

Paper number ten *The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series* by David Rushing Dewhurst, Thayer Alshaabi, Dilan Kiley, Michael V. Arnold, Joshua R. Minot, Christopher M. Danforth and Peter Sheridan Dodds, cited as [37].

Abstract

We introduce a qualitative, shape-based, timescale-independent time-domain transform used to extract local dynamics from sociotechnical time series—termed the Discrete Shocklet Transform (DST)—and an associated similarity search routine, the Shocklet Transform And Ranking (STAR) algorithm, that indicates time windows during which panels of time series display qualitatively-similar anomalous behavior. After distinguishing our algorithms from other methods used in anomaly detection and time series similarity search, such as the matrix profile, seasonal-hybrid ESD, and discrete wavelet transform-based procedures, we demonstrate the DST's ability to identify mechanism-driven dynamics at a wide range of timescales and its relative insensitivity to functional parameterization. As an application, we analyze a sociotechnical data source (usage frequencies for a subset of words on Twitter) and highlight our algorithms' utility by using them to extract both a typology of mechanistic local dynamics and a data-driven narrative of socially-important events as perceived by English-language Twitter.

Contribution

For this paper I contributed to the analysis of social media time series, generating Figure 3 in the paper, reproduced here as Figure 4.11. I contributed to parsing, storing, and querying the relevant socio-technical time series in this world. I also created a website with supplementary interactive figures, such as the one shown in Figure 4.12.

4.3.2 Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning

Paper number twelve is *Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning* by Thayer Alshaabi, Colin M. Van Oort, Mikaela Irene Fudolig, Michael V.

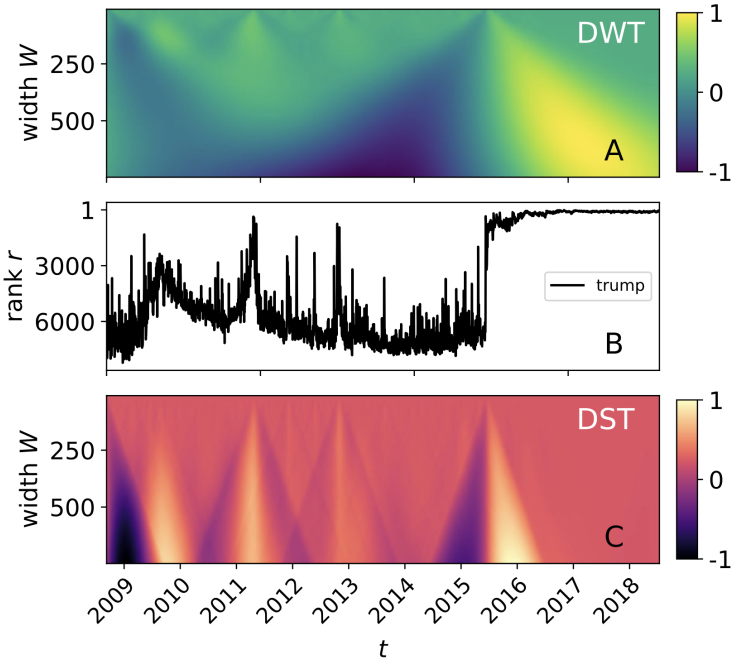


Figure 4.11: Reprint of Figure 3 from [37], with capition as follows: A comparison between the standard discrete wavelet transform (DWT) and our discrete shocklet transform (DST) of a sociotechnical time series. Panel (B) displays the daily time series of the rank of the word "trump" on Twitter. As a comparison with the DST, we computed the DWT of using the Ricker wavelet and display it in panel (A). Panel (C) shows the DST of the time series using a symmetric power shock.

Arnold, Christopher M. Danforth, and Peter Sheridan Dodds, cited as [11].

Abstract

Sentiment-aware intelligent systems are essential to a wide array of applications. These systems are driven by language models which broadly fall into two paradigms: Lexicon-based and contextual. Although recent contextual models are increasingly dominant, we still see demand for lexicon-based models because of their interpretability and ease of use. For example, lexicon-based models allow researchers to readily determine which words and phrases contribute most to a change in measured sentiment. A challenge for any lexicon-based approach is that the lexicon needs to be routinely expanded with new words and expressions. Here, we propose two models for automatic lexicon expansion.

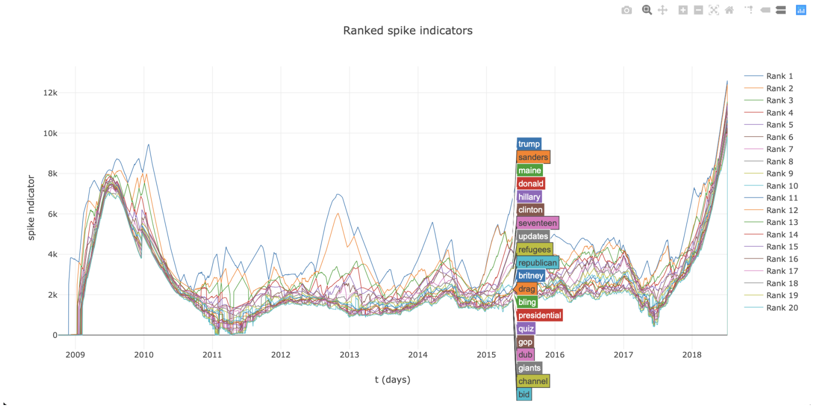


Figure 4.12: Screenshot of one of the paper's associated interactive online supplementary figure showing ranked spike indicators.

Our first model establishes a baseline employing a simple and shallow neural network initialized with pre-trained word embeddings using a non-contextual approach. Our second model improves upon our baseline, featuring a deep Transformer-based network that brings to bear word definitions to estimate their lexical polarity. Our evaluation shows that both models are able to score new words with a similar accuracy to reviewers from Amazon Mechanical Turk, but at a fraction of the cost.

Contribution

My main contribution to this paper was consulting with Thayer Alshaabi on potential models and conceptual conversations about the word-level sentiment prediction task for lexicon augmentation. Thayer found promising results using the character level word embeddings based on FastText. Further discussion yielded the second model, which uses dictionary definitions to supplement the semantic features associated with words within pre-trained embeddings.

We also discussed using ambient tweets paired with transformer based classifiers to predict sentiment scores. Although we didn't pursue this direction, I believe it's a promising option to create more dynamic sentiment scores as usage patterns change over time, but avoids the expense of relying on continued surveys.

Chapter 5

Conclusions

Having described my contributions to many of the 23 manuscripts I've co-authored during my time as a PhD student in Complex Systems & Data Science, I conclude here with thoughts about potential future work.

5.1 Collective Attention Prediction

The Chapter 2 case study of U.S. based hurricanes employed n-gram usage rates to measure associations between collective attention and impacts of the storms. Many other studies are amenable to this style of analysis, provided we can be confident in the quality of our proxy for collective attention, and are able to find data reflecting it. The topic of the Chapter 3 case study, validating tweets as relevant to a topic of interest, also gave promising results. In the future, I'd like to explore a number of themes related to attention using n-gram usage rates as a proxy including, for example:

- occupations,
- mortality,
- natural disasters, and
- sports teams.

For example, an occupation study might try to model mentions of job titles as dependant on work force size, salary, and required education. A mortality study might try to model attention to causes of mortality as a function of mortality rate and demographic risk factors.

5.2 Sub-population Culturomics

Our initial attempts to create measures of n-gram timeseries relied on massive parallelization to pre-compute separate counts for 1-grams, 2-grams, and 3-grams. This approach satisfied the criteria of the project, but it was relatively inflexible. We created separate counts for each language group, but this macro-community may not always be the ideal organizing principle of interest (though sometimes language groups are the community of interest, as in Figure 5.1.)

One can imagine situations where researchers are interested in computing proxies of attention for a community limited to a location, such as a country, state, or city. Or comparing proxies of attention in groups defined by demographic features such as political leanings, education level, gender, race, etc. Given an acceptably accurate classifier, it is computationally feasible to interactively compute a proxy for arbitrary communities using database aggregations, rather than the computationally expensive process of parsing n-grams. We haven't used the fraction of tweets containing a given keyword as a proxy, but it correlates strongly with n-gram usage rates.

To enable these aggregations, we'd need to add a new field to each document to store the classification. It is very likely faster to drop and re-insert entire collections than it is to attempt to update documents, due to the latency and inefficiency of document level writes operations.

We should create fields for bot classifications as well, using the same concept and adapting an existing tool like BotometerLite, which classifies accounts based only on metadata, so doesn't require the access to Twitter's API [34, 178, 179].

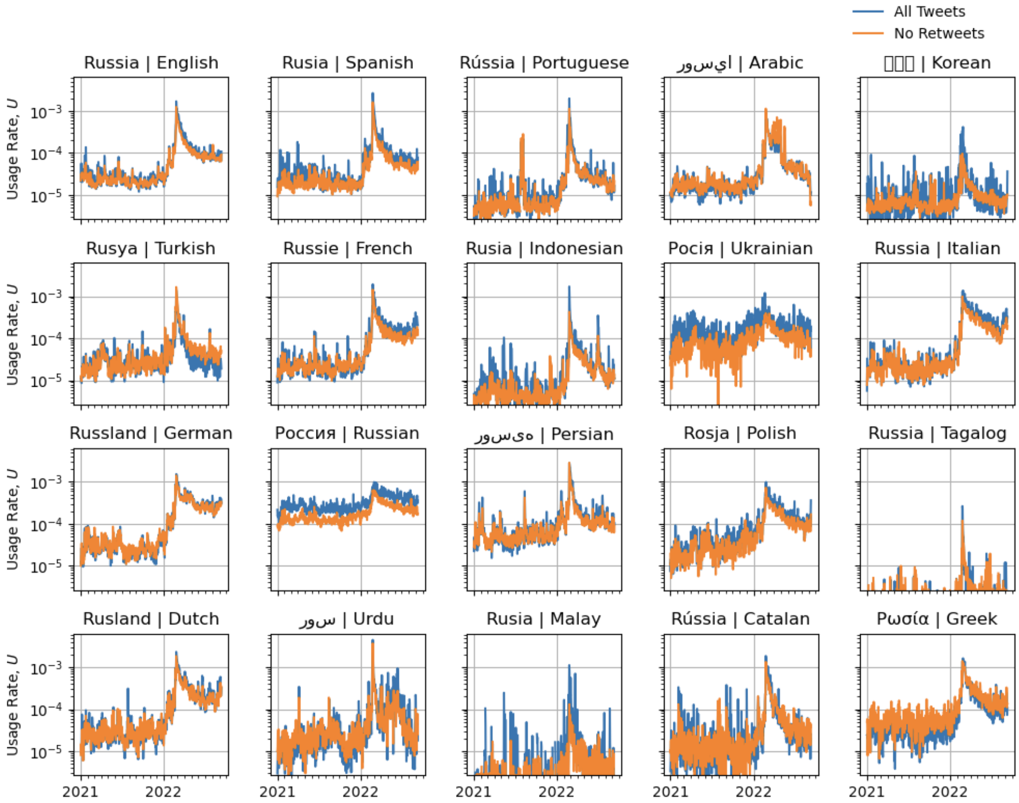


Figure 5.1: Multi-language comparison of collective attention for the 1*-gram "Russia", translated into 20 languages. Notably, all languages see a spike in mentions corresponding to the invasion of Ukraine in Feburary 2022, but Ukranian and Russian see the smallest increase.*

5.3 Opinion Polling

Having accurate classifications of user demographics would also be a first step to being able to post-stratify estimates for populations of interest beyond Twitter users, such as likely voters, or to measure text for target demographics. An alternative would be to create a representative panel of Twitter users, if demographic classification is found to be insufficiently accurate. To make quantitative estimates of opinion, we'd need estimates of target population demographics. Perhaps through partnerships with polling organizations such as Gallup, this functionality could be built out. Regrettably, the loss of real-time

Twitter data due to ongoing litigation associated with the training of LLMs significantly reduces the potential for this concept, since much of the value is derived from the fine temporal resolution of passive human expression. However, ethical considerations associated with building these capabilities should be further explored [103].

5.4 GPU Accelerated NLP

Our earlier contributions were based on bag-of-words methods or small models that could be improved with higher performing contextual NLP tools. GPU acceleration will enable our group to study social phenomenon faster using large social media corpora. We see custom fine-tuned classifiers as playing a critical role in allowing researchers to curate topic focused corpora. We hope to train language model based instruments to measure semantic differentials like valence, arousal, and dominance to help capture trends in public opinion. Multi-GPU inference will be needed to scale these instruments to process our 100 billion document scale corpora. In addition to ML inference tasks, GPU acceleration is increasingly available at all stages of the data analysis pipeline, including direct to GPU data loading, dimensionality reduction, and clustering [131]. GPU resources would enable us to pre-compute semantic embeddings to be stored in a vector database to allow for semantic similarity search and lower latency interactive visualizations for exploratory research [12].

5.5 Past, Present, and Future Hedonometer

In the future, I'd like to build three versions of the Hedonometer, adding distinct estimates of sentiment when people talk about the future and separately when they talk about the past. Indeed, the labMT dataset on which the Hedonometer is based did demonstrate a slight sentiment improvement associated with moving tense from past to future, a trajectory recently shown to appear in conversations as well [133]. Twitter data would be an obvious choice, but we could make these distinct sentiment timeseries measurements for any text.

On a technical level, there are two steps. We would run the Storywrangler parser on text data, creating separate counters for n-grams based on the verb tense within each clause. Then we would perform measurements like sentiment analysis for each corpus. While it will be more memory intensive to compute, it may be worth breaking down the tense categories from past, present, and future into smaller groups. We could then compare difference in language usage using sentiment shifts and rank-turbulence divergence.

This could also be structured as an ambient sentiment study, where the sentiment of ambient text around particular verbs in difference tenses is contrasted. Unfortunately, it would be computationally difficult to have noun-centered ambient text broken out by verb tense, since each tweet could potentially contain verbs of multiple tenses, so documents could not be pre-indexed.

5.6 Tweets with Locations

Our user location inferred tweet datasets are operational, representing around 5% of all tweets, but they only match cities and states within the United States if users' declared this information to their biographical profile. We have done some limited validation comparing tweets from users who both added this profile location and opted into adding precise GPS-based metadata from their mobile device.

There are many potential future projects utilizing U.S. state-level subsets. We're exploring looking at state-level language usage related to masking, social distancing, and vaccinations during the COVID-19 pandemic to state-level public health outcomes. Measuring associations between language measures and state-level policy would also be a natural future direction [6]. Finally, expanding our location matching to include more countries would be valuable, though inter-language text comparisons remain challenging.

Tweets with geospatial metadata also offer possibilities, especially during the period from 2012 to 2015 when it was more common for individuals to GPS tag their messages.

This data is stored on DataMountain and is indexed for fast queries, but has not yet been adequately explored. Combining this data with other spatial datasets would enable measuring associations in language use. One example we've explored is using the National Transportation Noise Map to examine the association between sentiment and exposure to traffic noise.

Additionally, I would like to generate relative usage rate and sentiment maps for keyword queries. For example, searching for the keyword '`Traffic`' one might expect to see it used relatively more frequently in urban areas relative to the overall usage rate. In Figure 5.2, we show spatially aggregated tweet counts for two anchor *n*-grams, '`Farm`' and '`Traffic`' as an example. Admittedly, Gelman warns that **all** maps of parameter estimates are misleading, since higher variation due to small sample sizes tends to make low density areas extreme values, while correcting for sample sizes makes low density areas too uniform [60].

5.7 Ousiometric Lexicon Generation

Our lab has put a lot of work into developing lexicons for sentiment (or other semantic differential) analysis. We've also built multiple tools to understand text using these lexicons. However, often the lexicons don't exactly correspond to a measurement we are interested in making. We also know that much of variance in scores for semantic differentials can be explained by just a few orthogonal dimensions, related to 'power', 'danger' and 'structure' [39]. It would be worth exploring using linear combinations of scored words along these dimensions to create specialized lexicons, without the expense of paying human raters.

5.8 BERTopic Weighting Scheme

BERTopic is a transformer powered, topic modeling technique. It has a few modular steps, beginning with embedding documents, using dimensionality reduction on those embeddings, and clustering. I believe we could contribute to the algorithm, which currently uses cluster

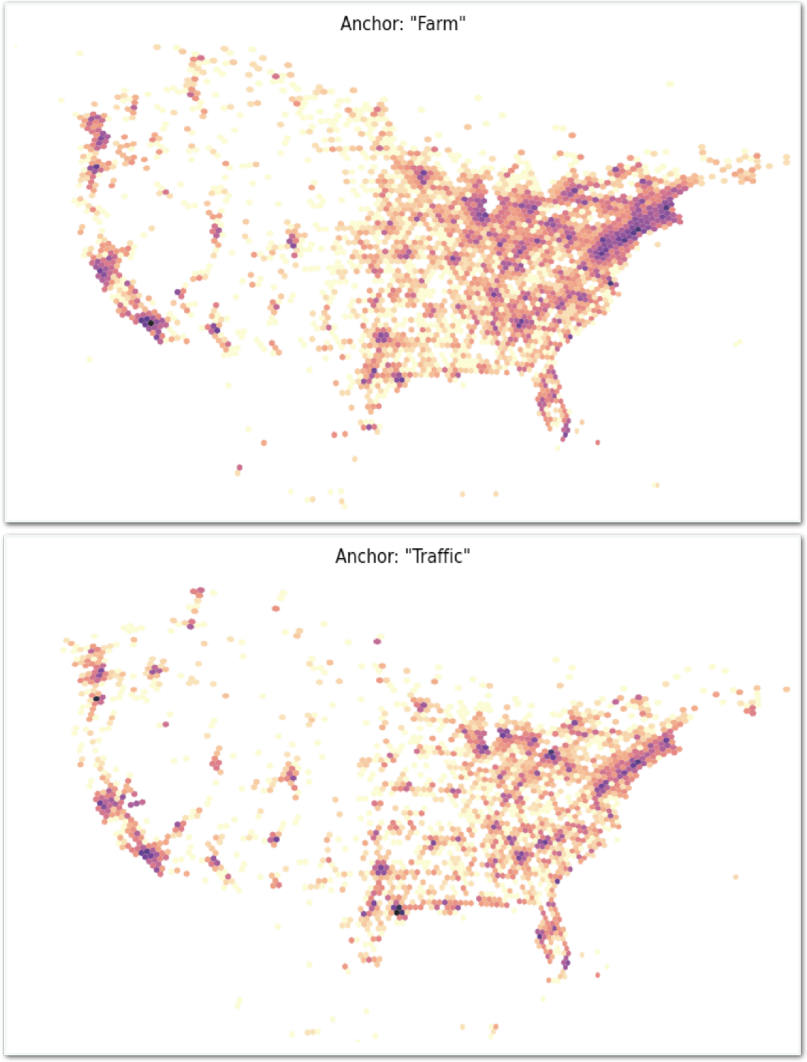


Figure 5.2: Spatially aggregated counts of tweets containing anchors, ‘`Farm`’ and ‘`Traffic`’. While both counts are highly correlated with population density, ‘`Traffic`’ counts seem to be more intensely peaked in urban areas, while ‘`Farm`’ counts appear to be more broadly distributed. There seems to be potential for further studies of spatial variation of language usage.

Term Frequency - Inverse Document Frequency (c-TF-IDF) weighting to summarize the topics [68]. This leads to many function words being included as representative terms for the cluster. We could replace c-TF-IDF with a tunable turbulence divergence weighting scheme that better describes document clusters at the scale most useful to researchers.

5.9 Additional Corpora

There would be great benefit in adding new text corpora within the same DataMountain software environment that we currently use to access tweets. Our text comparison tools, such as allotaxonographs and word shifts utilizing specialized lexicons, will be more accessible to interdisciplinary researchers with varying levels of programming expertise.

Reddit data will likely be the first extension. As another large social media platform, it will be natural to contextualize insights gleaned from Twitter data using comments posted to Reddit as well. With different sociotechnical algorithms moderating interactions, one could imagine the language usage distributions could be quite distinct [65]. The sub-reddit community structure is substantively different from Twitter's single public square model.

Another corpus I would like to acquire is a music lyrics dataset. The first data product would be an n-gram viewer from the perspective of songwriters, where n-gram frequency is computed for each publishing date, likely at the year scale [43]. This would a cultural record, similar to how Google books encodes the words of authors or how Storywrangler encodes the words of Twitter users.

The second music lyrics product would require a partnership with a streaming platform like Spotify. Daily play counts of songs could be used to create a popularity weighted n-gram count, from the perspective of listeners. From here, we can measure sentiment or other lexicons of interest. Do people listen to happier music on holidays? Sad music during national tragedies? With play counts aggregated by locations we could even map the sentiment of experienced music [124].

Of course there are many more potential corpora to add, either as raw text or as parsed n-grams. A short, non-comprehensive list:

- newspapers,
- legal texts,

- Front Porch Forum / Nextdoor,
- TV news,
- Google Trends,
- Google Books, and
- Wikipedia.

Chapter 6

Supporting information for Hurricanes and Hashtags

6.1 Summary Tables for Regressions

Provided for the reader here are tables of summary statistics of the estimated parameters in the regression models in Section 2.4.3 and Section 2.4.4.

Mean Regression Parameters – Deaths

	Tropical Storms	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	All Hurricanes
a_{deaths}	0.25	0.61	0.31	0.72	1.39	1.35	1.16
a_0	-7.65	-6.63	-6.58	-6.25	-6.01	-6.91	-6.56

Mean Regression Parameters – Damages

	Tropical Storms	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	All Hurricanes
a_{damage}	0.06	0.17	0.17	0.24	0.37	0.46	0.31
a_0	-7.91	-7.41	-7.27	-7.21	-7.60	-8.22	-7.92

Table 6.1: Mean Regression Parameters fit for storms of each category. See Fig. 2.4 for full parameter distributions.

a_0	a_{death}	a_{damage}
$\mathbf{normal}(-8,3)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$

Table 6.2: Priors for Regression 1: Linear in deaths and damages

	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
a_0	-7.57	0.52	0.01	-8.60	-6.56	4182	1.0
Deaths	0.49	0.16	0.00	0.16	0.80	4660	1.0
Damage	0.24	0.08	0.00	0.08	0.40	4108	1.0
sd	0.89	0.08	0.00	0.75	1.05	8449	1.0

Table 6.3: Results for Regression 1: Linear in deaths and damages

a_0	a_{death}	a_{damage}	$a_{\text{d,D}}$
$\mathbf{normal}(-8,3)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$

Table 6.4: Priors for Regression 2: Additional interaction term

	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
a_0	-7.58	0.51	0.01	-8.58	-6.58	8085	1.0
Deaths	0.05	0.34	0.00	-0.65	0.70	8326	1.0
Damage	0.22	0.08	0.00	0.06	0.38	8151	1.0
Interaction	0.06	0.04	0.00	-0.02	0.14	8676	1.0
sd	0.88	0.08	0.00	0.74	1.04	10843	1.0

Table 6.5: Results for Regression 2: Additional interaction term

6.2 2-gram Attention Proportion of “hurricane” Usage Rate

Examining the top 2-grams matching the pattern “hurricane *” in Fig. 6.1, we can get a sense of what are the top storms during the season, and how much attention is allocated to each at a given time. For English tweets, the first major spike of the 2017 hurricane season is surrounding Hurricane Harvey, though attention also spikes for Hurricane Katrina, in reference to the 2005 storm that affected a nearby region of the gulf coast. As attention begins to decay for Hurricane Harvey, a spike in usage for the 2-gram “hurricane relief” is observed, though it reaches only $f = 3 * 10^{-5}$. Next, attention turns to Hurricane Irma, which reaches the highest 2-gram usage rate of any hurricane in our dataset. Finally, one

a_0	a_{death}	a_{damage}	$a_{\text{d}\times\text{D}}$	a_{C_i}
$\mathbf{normal}(-8,3)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$	$\mathbf{normal}(0,1)$

Table 6.6: Priors for Regression 3: Additional categorical term for hurricane category

	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
a_0	-7.64	0.51	0.01	-8.60	-6.60	9916	1.0
Deaths	0.09	0.36	0.00	-0.60	0.81	9892	1.0
Damage	0.20	0.08	0.00	0.05	0.35	10580	1.0
Interaction	0.05	0.04	0.00	-0.04	0.13	10424	1.0
Cat2	0.07	0.31	0.00	-0.55	0.66	15415	1.0
Cat3	0.21	0.26	0.00	-0.32	0.72	14877	1.0
Cat4	0.76	0.28	0.00	0.20	1.29	15063	1.0
Cat5	0.66	0.44	0.00	-0.17	1.57	13237	1.0
sd	0.84	0.08	0.00	0.70	1.00	14240	1.0

Table 6.7: Results for Regression 3: Additional categorical term for hurricane category

week after attention for Irma begins to decay, attention spikes for Hurricane maria, though at a level noticeably lower than for Harvey or Irma.

We notice that during storm events the 2-gram usage rates for storms “hurricane *” is often between only half or a fifth the usage rate of the 1-gram “hurricane”, meaning that about one in every 5 times the name of the storm follows the word hurricane in English tweets during active storms.

In Spanish tweets the usage rates of “Huracàn Harvey” only reach a maximum of around $f \sim 10^{-4}$, while “Huracàn Irma” receives much more relative attention. “Huracàn Marìa” receives about as much attention as Harvey, and also occupies a space similar to “Hurricane Maria” in English, around $f \sim 10^{-4}$.

6.3 Bi-exponential Decays

To quantify the characteristic time scales of attention given to storms, we examined usage rates by fitting the bi-exponential model introduced by Candia et al. [23]. Not all storms receive enough attention, but 50 of 75 in the Atlantic basin recorded at least 6 days of

	Integrated Frequency	Max Frequency	Deaths	Damage	Quantile 0.99	Quantile 0.9
2017 Harvey	2.3×10^{-3}	3.5×10^{-4}	107	1.2×10^{11}	126	14
2017 Maria	4.9×10^{-4}	5.0×10^{-5}	3057	9.1×10^{10}	363	166
2017 Irma	1.6×10^{-3}	4.6×10^{-4}	134	7.7×10^{10}	75	15
2012 Sandy	3.7×10^{-4}	1.5×10^{-4}	286	6.8×10^{10}	157	13
2018 Michael	3.7×10^{-4}	1.1×10^{-4}	72	2.5×10^{10}	201	13
2018 Florence	9.3×10^{-4}	1.8×10^{-4}	57	2.4×10^{10}	44	15
2016 Matthew	9.9×10^{-4}	2.6×10^{-4}	603	1.6×10^{10}	136	15
2011 Irene	2.0×10^{-4}	8.0×10^{-5}	58	1.4×10^{10}	14	8
2019 Dorian	5.7×10^{-4}	1.2×10^{-4}	70	4.6×10^{9}	36	12
2012 Isaac	2.6×10^{-5}	6.1×10^{-6}	41	3.1×10^{9}	192	97
2010 Alex	5.8×10^{-6}	2.5×10^{-6}	52	1.5×10^{9}	15	7
2017 Nate	6.3×10^{-5}	3.1×10^{-5}	48	7.8×10^{8}	8	5
2019 Barry	1.1×10^{-5}	3.8×10^{-6}	1	6.0×10^{8}	8	4
2016 Hermine	4.1×10^{-5}	1.9×10^{-5}	5	5.5×10^{8}	7	3
2019 Lorenzo	4.1×10^{-6}	1.0×10^{-6}	16	3.6×10^{8}	11	9
2014 Gonzalo	1.5×10^{-5}	6.4×10^{-6}	6	3.1×10^{8}	14	11
2015 Joaquin	3.7×10^{-5}	1.1×10^{-5}	34	2.0×10^{8}	11	5
2017 Ophelia	2.7×10^{-5}	1.2×10^{-5}	5	8.7×10^{7}	15	7
2009 Bill	1.6×10^{-5}	9.4×10^{-6}	2	4.6×10^{7}	11	7
2010 Earl	1.9×10^{-5}	4.9×10^{-6}	8	4.5×10^{7}	8	6
2014 Arthur	2.5×10^{-5}	1.3×10^{-5}	1	1.6×10^{7}	9	5
2016 Nicole	1.1×10^{-5}	5.3×10^{-6}	1	1.5×10^{7}	13	9
2017 Katia	4.0×10^{-6}	1.1×10^{-6}	3	3.2×10^{6}	7	4
2017 Jose	2.9×10^{-5}	4.7×10^{-6}	1	2.8×10^{6}	22	13
2014 Bertha	2.7×10^{-6}	1.1×10^{-6}	4	0.0	11	8
2015 Danny	4.0×10^{-6}	1.8×10^{-6}	0	NaN	6	3

Table 6.8: The unnormalized values associated with radar plots in Section 2.4, sorted by estimated damage.

consecutive 2-gram usage within the year of the hurricane, and these storms were had both their hashtag and 2-gram usage rate fit with the bi-exponential model of Candia et al.The model here assumes two populations, u and v, which become interested in a given event. Population u, comparable to the general population starts with a peak interest, and losses attention as $\frac{du}{dt} = -(p+r)u$. During every unit time pu attention is lost from the system and ru is transferred to population v. The dynamics of population v are as follows: $\frac{dv}{dt} = ru - qv$, so attention decays from v with rate q, but increases proportionally to the total attention of population u. The final bi-exponential model is

$$S(t) = \frac{N}{p+r-q}[(p-q)e^{-(p+r)t} + re^{-qt}],$$

and we present the half-lives associated with this model as $\tau_1 = \frac{\ln(2)}{(p+r)}$ and $\tau_2 = \frac{\ln(2)}{q}$, which are the rates of decay from the two populations u and v. The distributions of τ_1 and τ_2 for both hashtag usage rates and 2-gram usage rates are shown in Fig. 6.3. The mean half-life for population u, the population with faster attention decay, is $\bar{\tau_1} = 1.3$ days for hashtags, and $\bar{\tau_1} = 1.1$ days for 2-grams. The decays for population v were not uni-modal, due to some storms regaining attention long after their initial impact, deviating from the model and receiveing poor fits, and resulting in very large values of τ_2, but median values were approximately 24 days. All summary statistics are reported in Table 6.10. We speculate that for this model the population u is largely people effected by the storm, while population v is largely people writing about the storms or sharing information about the storm response, eg, reporters and non-profit professionals. Further work could look to confirm who is behind the tweets.

The fitting procedure was to first find the maximum value of the usage rate for each storm, before fitting the above model to the decay of log usage rate after this maximum. The resulting fits are shown in Fig. 6.6 and Fig. 6.7. The fits generally appear sensible, but there are sometimes issues for noisy time series, where the rate parameter r becomes very small, corresponding to a very long half-life, and misfitting the early decay. This occurs in the time series for Hurricane Florence. The distributions of Mean Squared Error (MSE) are shown in Fig. 6.5.

Looking at the decay half-lives in Table 6.10 we notice can see that most hurricane hashtags lose half their volume on the order of 1 or 2 days. The storms with relatively more attention on Twitter, Harvey, Irma, Matthew, and Sandy, all initially decay quickly, with a half-life on the order of a few days, but then have much longer decays associated with τ_2, on the order of a few weeks. There are some aberrations where the bi-exponential model does a poor job of explaining the data, such as for hurricane Joaquin, where a fight between Governor Bobby Jindal and the Obama administration over the size of a recovery package spurred news stories and attention long after the initial activity associated with the storm itself. This leads to increases in hashtag usage rate, and thus negative half-lives. The longest half-life is associated with hurricane Maria, τ_2 was approximately twice as long as the next largest hurricane. The extended crisis in Puerto Rico caused by Maria may be a reason this exceedingly long lifetime, even though the initial attention received by the hashtag was less than storms of comparable strength.

We also fit a simple exponential model $S(t) = Ne^{-pt}$. For high attention storms for which we have more than a week of data, this model is unable to capture decays occurring on different time scales, and thus has poor fits. For smaller storms for which attention is lower than the resolution of our data set, the exponential model is perhaps more appropriate. A distribution of half-lives for hashtags and 2-grams is shown in Fig. 6.4. While for larger storms, the fits did not capture the changing rates of attention decay, it was adequate for smaller storms that decay quickly below our instrument's resolution. However, for storms for which we have data for an extended decay, the bi-exponential model is more appropriate.

	Max Usage Rate	τ_1 [Days]	τ_2 [Days]	Season
#hurricanealex	2.5×10^{-6}	0.7	8.6	2010
#hurricanearthur	1.3×10^{-5}	0.9	190.3	2014
#hurricanebarry	3.8×10^{-6}	0.7	16.0	2019
#hurricanebertha	1.1×10^{-6}	0.6	6.9	2014
#hurricanebill	9.4×10^{-6}	0.2	693.1	2009
#hurricanechris	8.9×10^{-7}	0.6	693.1	2018
#hurricanecristobal	2.0×10^{-7}	2.0	6.9	2014
#hurricanedanielle	1.9×10^{-7}	0.7	693.1	2010
#hurricanedanny	1.8×10^{-6}	0.7	6.9	2015
#hurricanedorian	1.2×10^{-4}	1.6	8.8	2019
#hurricaneearl	5.0×10^{-6}	0.4	6.9	2010
#hurricaneflorence	1.8×10^{-4}	2.8	323.3	2018
#hurricanegert	3.6×10^{-7}	0.4	6.9	2017
#hurricanegonzalo	6.4×10^{-6}	0.9	693.1	2014
#hurricaneharvey	3.5×10^{-4}	2.5	30.6	2017
#hurricanehermine	1.9×10^{-5}	0.8	15.9	2016
#hurricaneida	8.3×10^{-7}	0.8	9.7	2009
#hurricaneigor	2.2×10^{-7}	1.1	693.1	2010
#hurricaneirene	8.0×10^{-5}	0.7	26.5	2011
#hurricaneirma	4.6×10^{-4}	1.0	20.0	2017
#hurricaneisaac	6.1×10^{-6}	0.7	693.1	2012
#hurricanejoaquin	1.1×10^{-5}	1.2	57.7	2015
#hurricanejose	4.7×10^{-6}	2.0	23.1	2017
#hurricanekarl	7.4×10^{-8}	0.6	68.9	2010
#hurricanekatia	8.7×10^{-7}	0.2	6.9	2011
#hurricanelorenzo	1.0×10^{-6}	1.3	64.2	2019
#hurricanemaria	5.0×10^{-5}	4.1	43.4	2017
#hurricanematthew	2.6×10^{-4}	1.4	27.4	2016
#hurricanemichael	1.1×10^{-4}	1.8	20.2	2018
#hurricanenate	3.1×10^{-5}	0.5	10.6	2017
#hurricanenicole	5.3×10^{-6}	0.6	6.9	2016
#hurricaneophelia	1.2×10^{-5}	0.3	6.9	2017
#hurricanesandy	1.5×10^{-4}	1.1	23.0	2012
#hurricanetomas	3.0×10^{-7}	0.9	6.9	2010

Table 6.9: Fitted half-lives τ_1 and τ_2 for all storms with at least 10 days of hashtag usage.

	Max Usage Rate	τ_1 [Days]	τ_2 [Days]	Season
Hurricane Alex	4.1×10^{-5}	0.8	9.3	2010
Hurricane Arthur	2.8×10^{-5}	1.0	693.1	2014
Hurricane Barry	8.9×10^{-6}	0.6	6.9	2019
Hurricane Bertha	8.2×10^{-6}	0.4	693.1	2014
Hurricane Bill	8.2×10^{-5}	0.8	9.7	2009
Hurricane Chris	3.0×10^{-5}	0.6	693.1	2018
Hurricane Cristobal	1.9×10^{-6}	1.5	693.1	2014
Hurricane Danielle	1.0×10^{-5}	0.9	7.1	2010
Hurricane Danny	7.6×10^{-6}	0.6	693.1	2015
Hurricane Dorian	1.1×10^{-4}	2.6	18.2	2019
Hurricane Earl	1.7×10^{-4}	1.2	9.5	2010
Hurricane Florence	1.3×10^{-4}	3.5	37.1	2018
Hurricane Gert	1.0×10^{-6}	2.1	321.9	2017
Hurricane Gonzalo	1.4×10^{-5}	1.7	693.1	2014
Hurricane Harvey	4.0×10^{-4}	2.9	29.3	2017
Hurricane Hermine	2.0×10^{-5}	0.4	6.9	2016
Hurricane Ida	4.5×10^{-5}	0.7	17.1	2009
Hurricane Igor	1.1×10^{-5}	1.0	25.2	2010
Hurricane Irene	3.3×10^{-4}	1.2	21.8	2011
Hurricane Irma	5.0×10^{-4}	2.3	24.1	2017
Hurricane Isaac	3.8×10^{-5}	1.6	21.1	2012
Hurricane Joaquin	4.4×10^{-5}	1.2	144.5	2015
Hurricane Jose	2.4×10^{-5}	1.3	7.1	2017
Hurricane Karl	1.6×10^{-5}	0.3	6.9	2010
Hurricane Katia	9.3×10^{-6}	2.1	7.4	2011
Hurricane Lorenzo	2.7×10^{-6}	1.7	8.1	2019
Hurricane Maria	1.1×10^{-4}	0.7	6.9	2017
Hurricane Matthew	2.9×10^{-4}	1.7	22.4	2016
Hurricane Michael	9.3×10^{-5}	2.5	27.2	2018
Hurricane Nate	3.5×10^{-5}	0.5	693.1	2017
Hurricane Nicole	1.2×10^{-5}	0.3	6.9	2016
Hurricane Ophelia	1.9×10^{-5}	0.5	6.9	2017
Hurricane Sandy	5.3×10^{-4}	2.1	28.5	2012
Hurricane Tomas	1.4×10^{-5}	0.9	6.9	2010

Table 6.10: Fitted half-lives τ_1 and τ_2 for all storms with at least 10 days of 2*-gram usage.*

6.4 Hurricane Attention Maps

The remaining Hurricane Attention Map and time series from 2009 to 2018 are presented for the reader's perusal. Only storms reaching at least Category 2 are shown, and Seasons 2013 and 2014 are omitted. Earlier storms in our dataset mostly did not make landfall, and thus appear to recieve relatively little attention. The scale of attention on the maps is held constant between years.

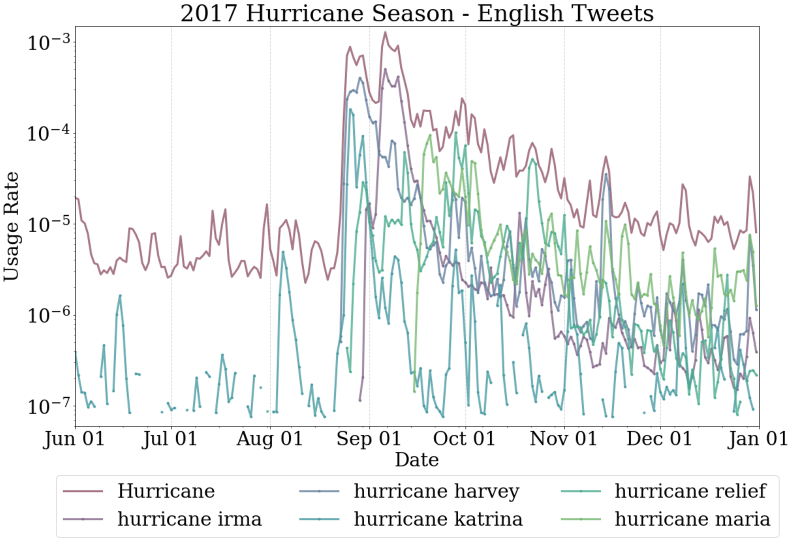


Figure 6.1: Word usage rate proportions of “`hurricane *`” in English tweets

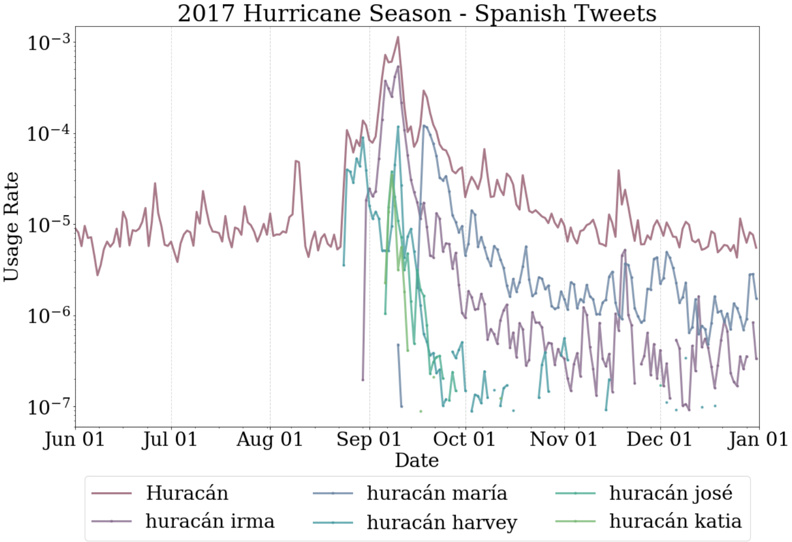


Figure 6.2: Attention proportions of “`Huracàn *`” in Spanish. We can see that the word usage rate surrounding “`Hurricane Maria`” captures a similar amount of the total attention for the 1*-gram hurricane as “`Huracàn Marìa`” captures. Additionally, hurricane Harvey’s* 2*-gram usage rate is lower in Spanish than in English, while Hurricane Katrina is talked about considerably in English but does not rise about the 50000th most used* 2*-gram in Spanish. As always, usage rates are case-insensitive.*

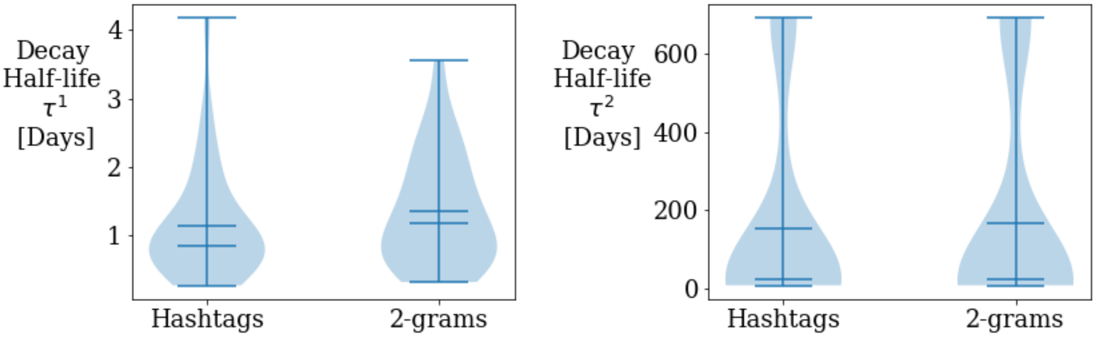


Figure 6.3: ***Bi-exponential Hurricane decay half-lives:*** *Distributions of fitted half-lifes for the populations u and v. The mean half-lives for $\tau_1 = 1.3$ days and $\tau_2 = 156$ days for hashtags and $\tau_1 = 1.1$ days and $\tau_2 = 241$ days for 2-grams. For τ_2 the median half-lives are also interesting since we suspect the longest half-lives are due to poor fits. For hashtags $\tau_2 = 23$ days, and for 2-grams $\tau_2 = 24$ days.*

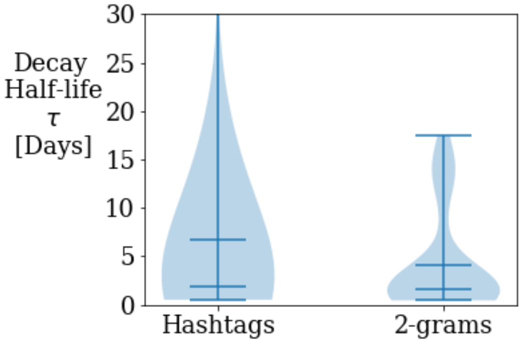


Figure 6.4: ***Simple Exponential Hurricane decay half-lives:*** *Distributions of fitted half-lives for a single population. The median half-lives for $\tau = 5.3$ days a for hashtags and $\tau = 5.2$ days for 2-grams. The simple exponential model fails to explain the break in attention decay for larger storms, receiving more attention. The bi-modal distribution of half-lives for 2-grams suggests that there are two categories of storms, ones with larger half-lives have more data, and thus the longer decay increases the fitted half-life. Meanwhile, smaller storms receive so little attention, that we don't measure any after a week or so, leading to a much smaller half-live, which corresponds to τ_1 in our bi-exponential fit.*

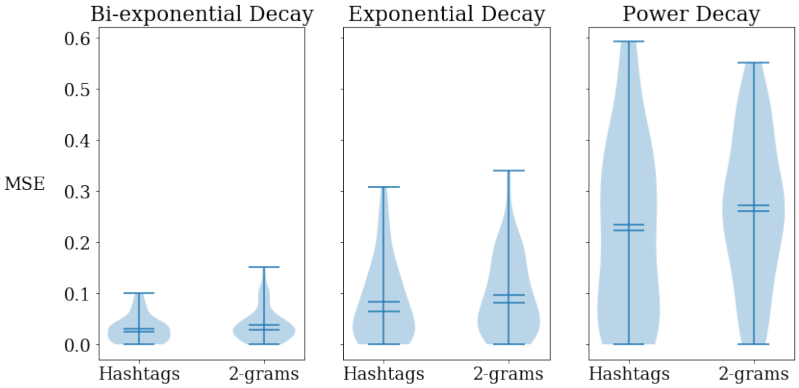


Figure 6.5: ***Decay Model Comparision:*** *Distributions of Mean Squared Error (MSE). The bi-exponential model has the lowest average MSE, followed by the simple exponential decay. The power law decay fails to capture the dynamics of attention decay, when the fit is compared to the data visually, and is reflected in the higher average MSE.*

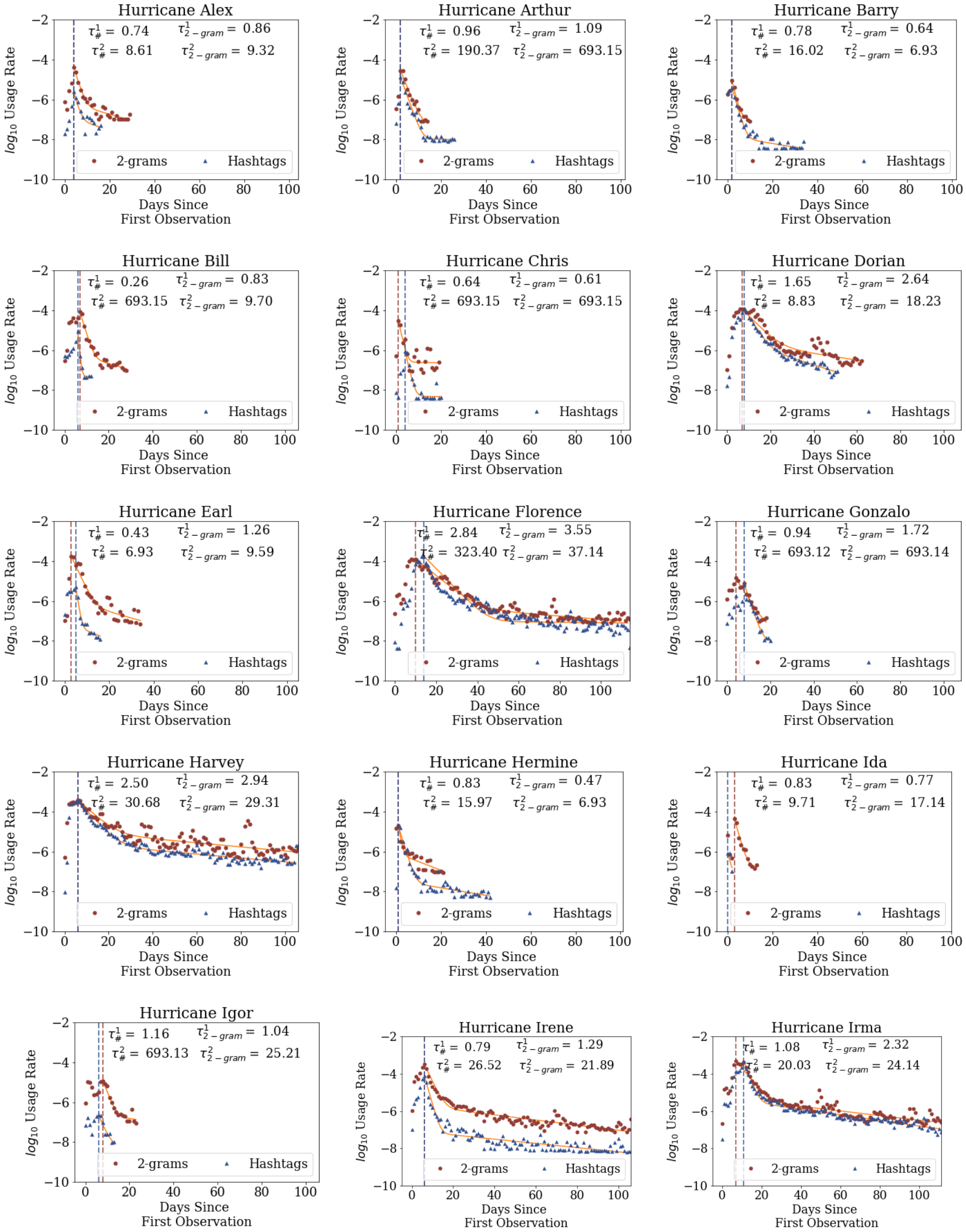


Figure 6.6: Hurricane bi-exponential decay fits for hashtag usage rates and 2-gram usage rates for “`hurricane *`”

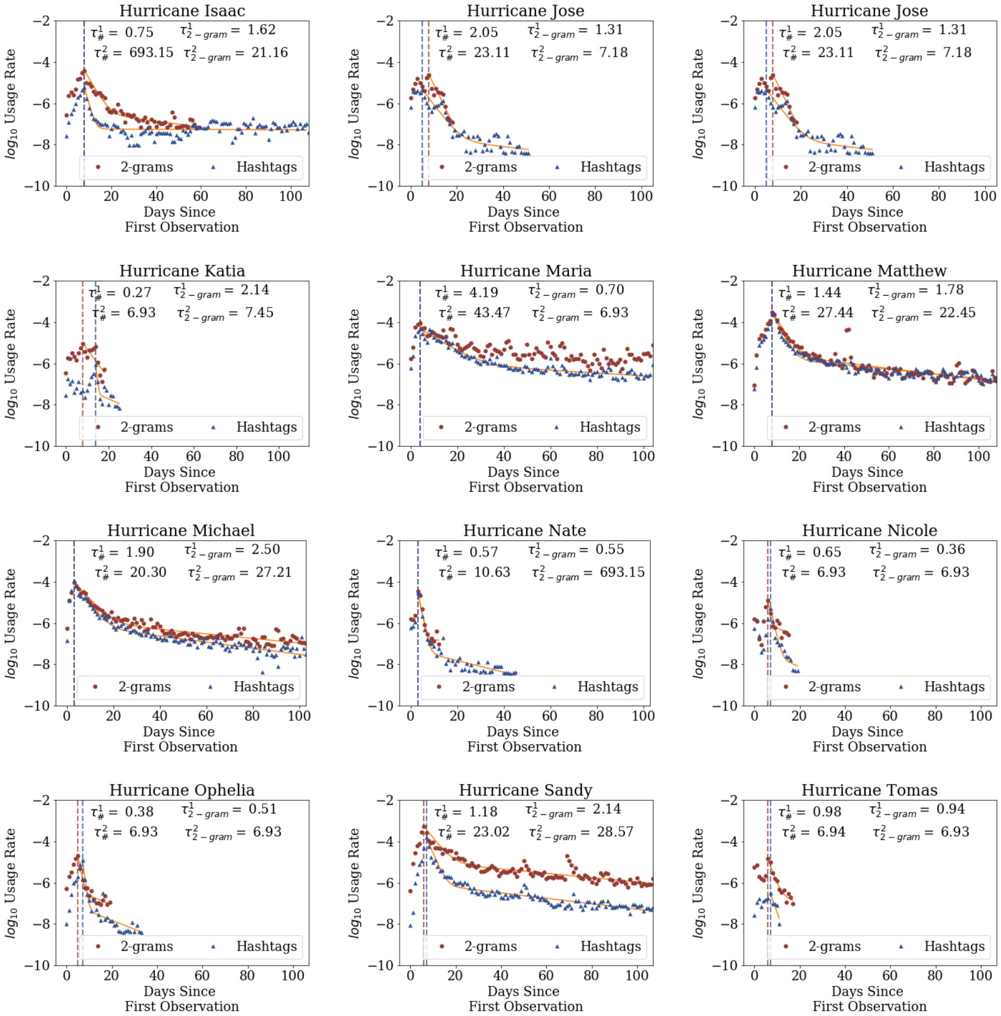


Figure 6.7: Hurricane decays fits for all hurricanes for which we have at least 10 days of 2-gram usage rate data. Fits are performed for the function $y = \frac{N}{p+r-q}[(p-q)e^{-(p+r)t} + re^{-qt}]$, a simple two population decay model as proposed by Candia et al. [23]. Here p would be interpreted as rate of decay from population 1, r would be the transfer rate from population 1 to population 2, and r would be the rate of decay from population 2. Population 1 might be thought of as bystandards with a shorter attention span, while population two are those living with the ramifications, or working on the recovery who lose attention more slowly. Reported on the graph are the half lives associated with fitting this model for both the hashtag usage rate and 2-gram usage rate, $\tau_1 = \frac{\ln 2}{p+r}$ and $\tau_2 = \frac{\ln 2}{q}$

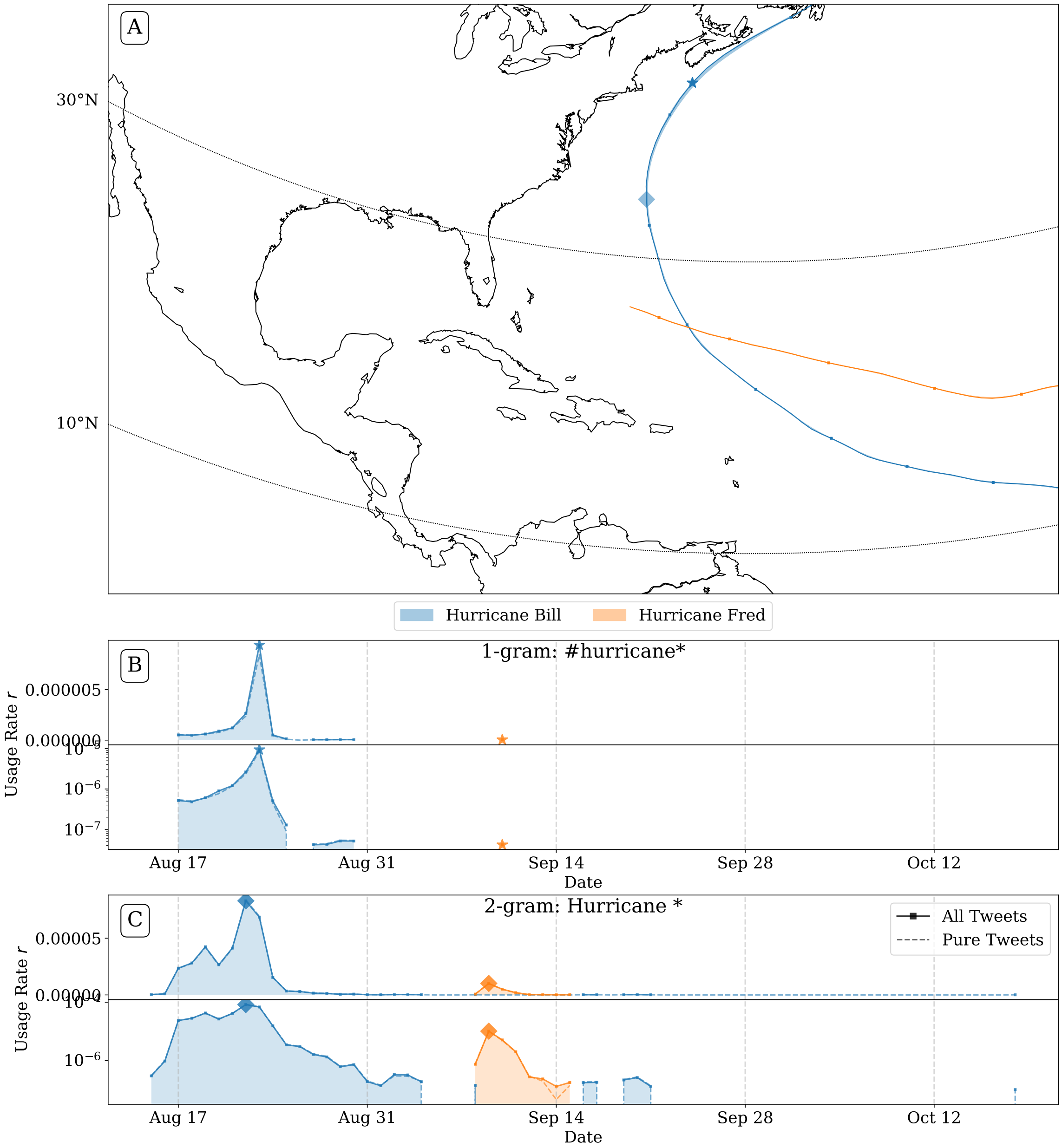


Figure 6.8: Hurricane Attention Map and time series for 2009

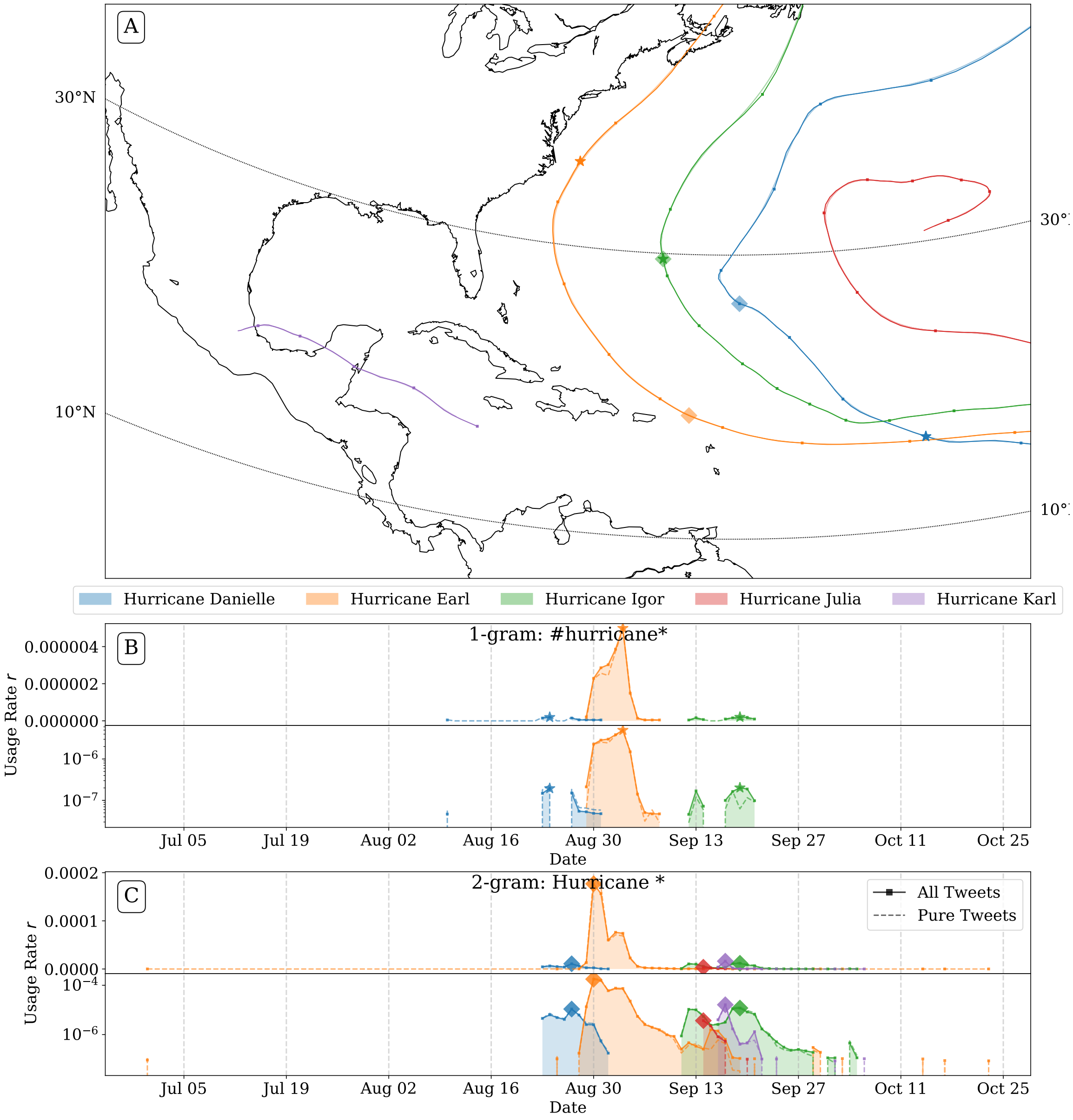


Figure 6.9: Hurricane Attention Map and time series for 2010

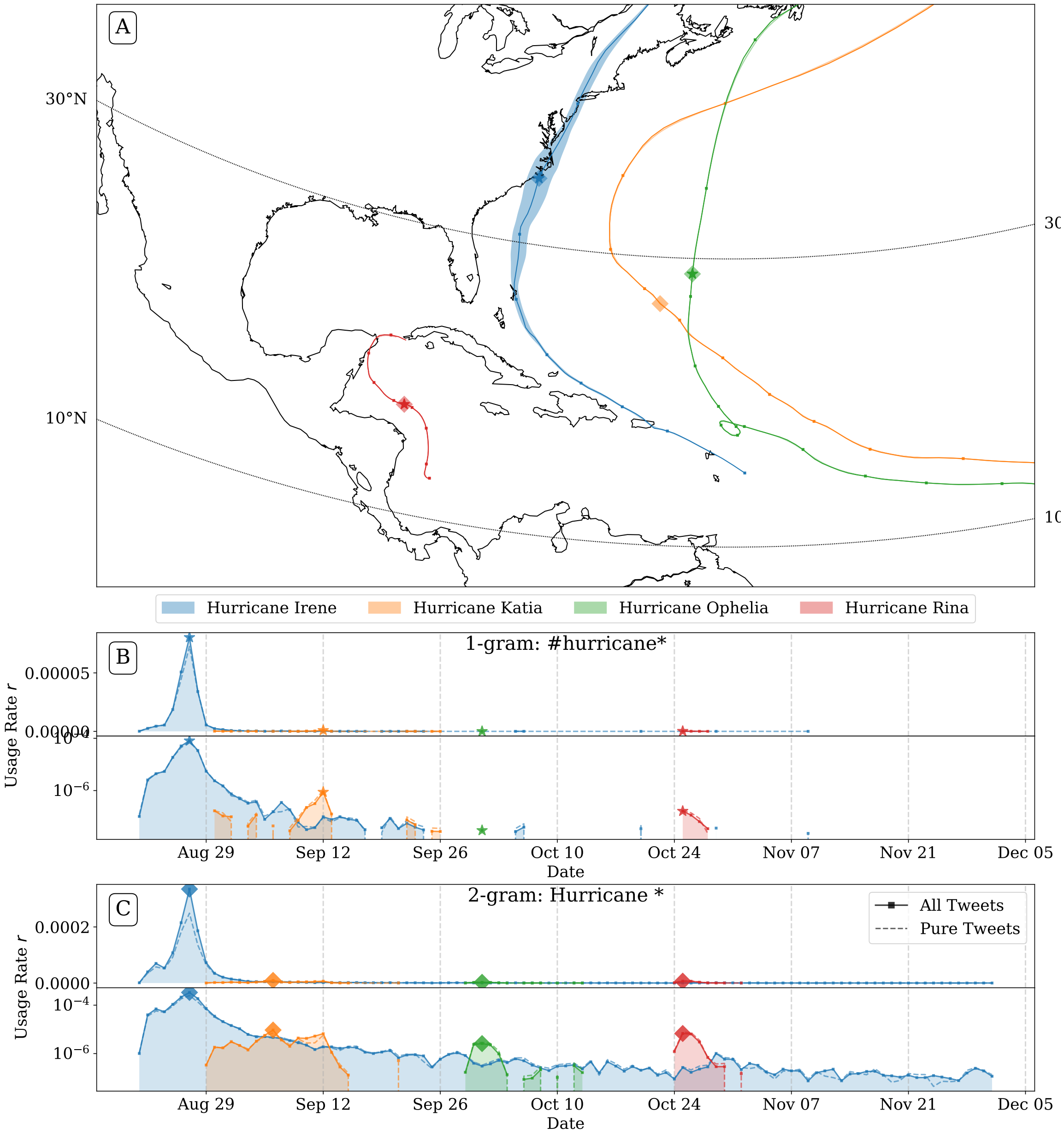


Figure 6.10: Hurricane Attention Map and time series for 2011

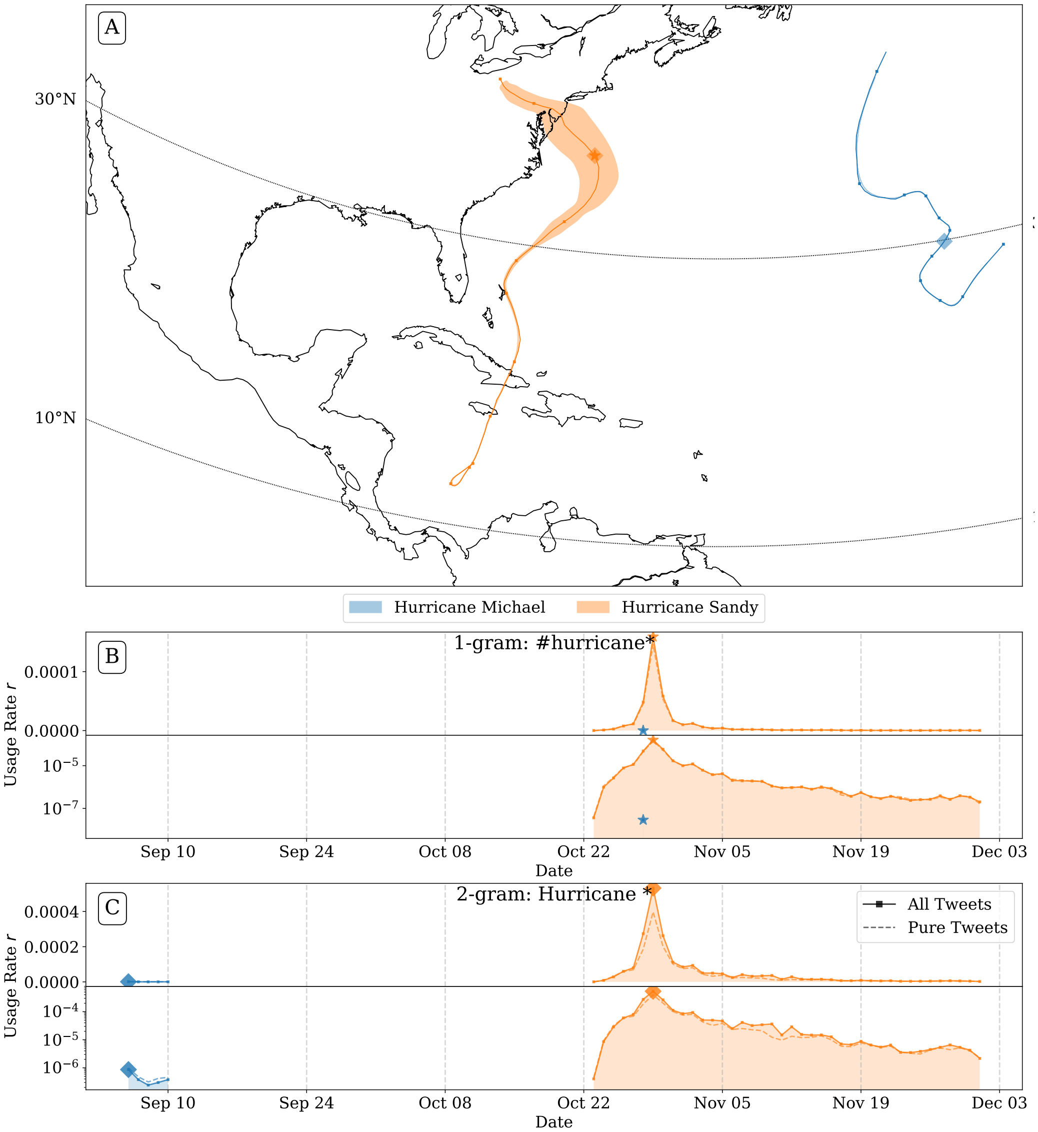


Figure 6.11: Hurricane Attention Map and time series for 2012

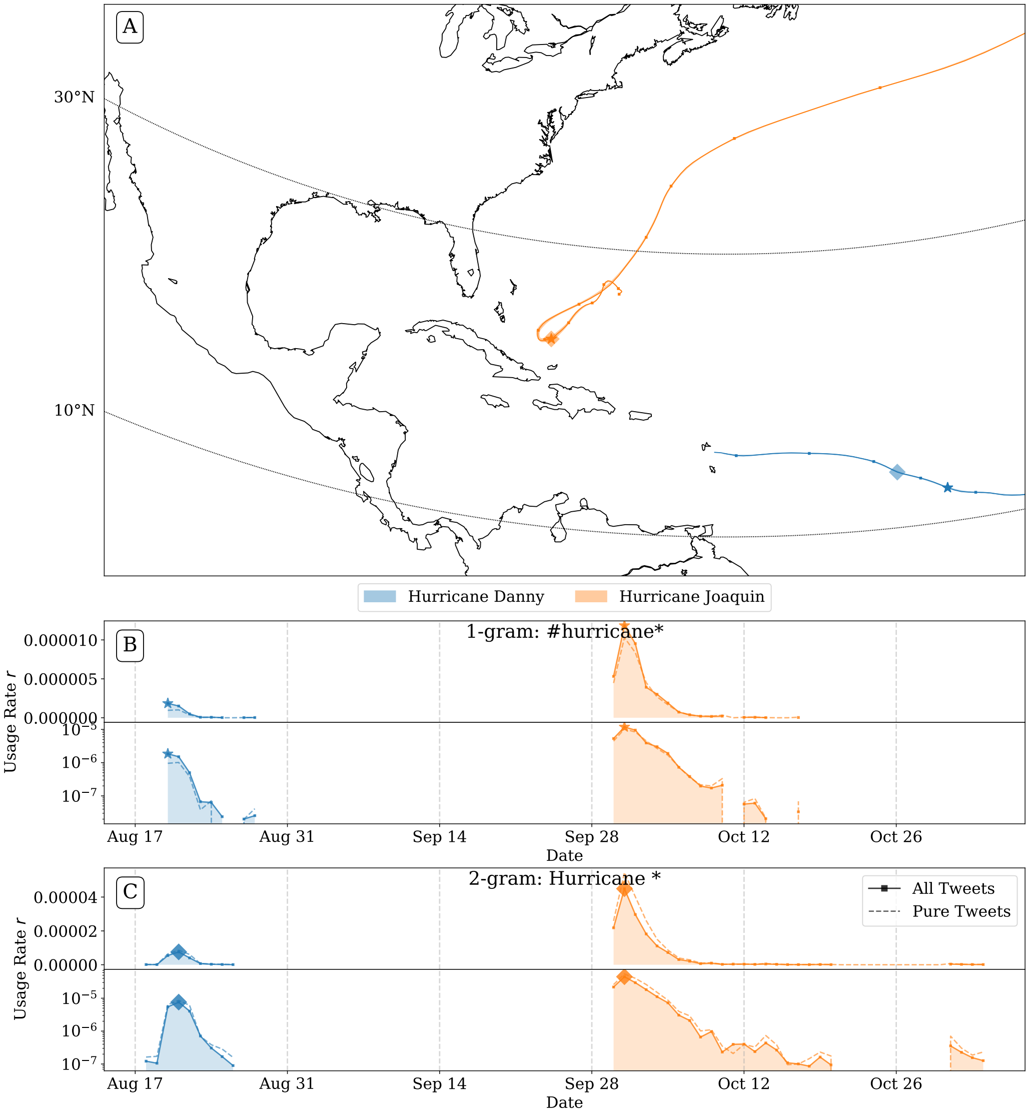


Figure 6.12: Hurricane Attention Map and time series for 2015

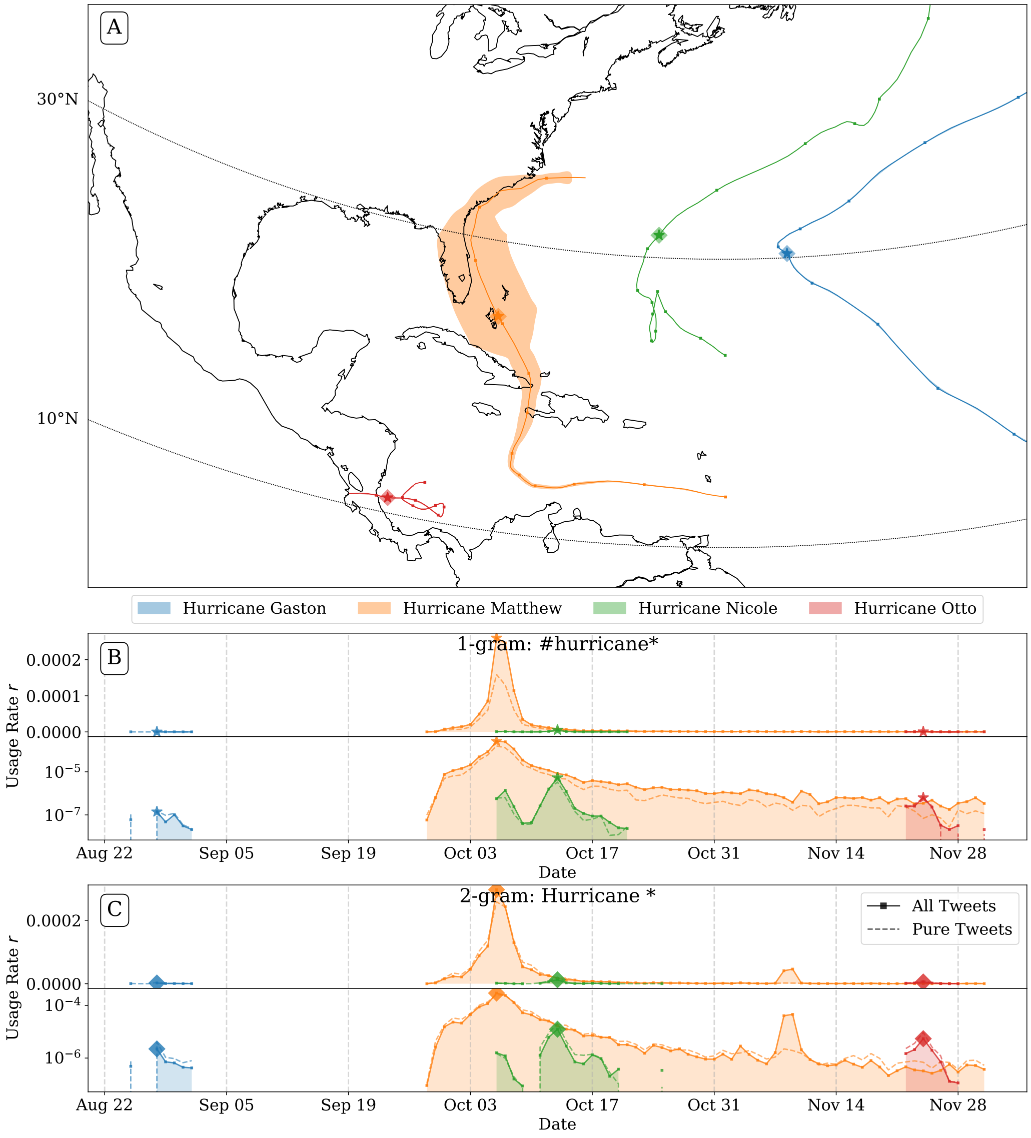


Figure 6.13: Hurricane Attention Map and time series for 2016

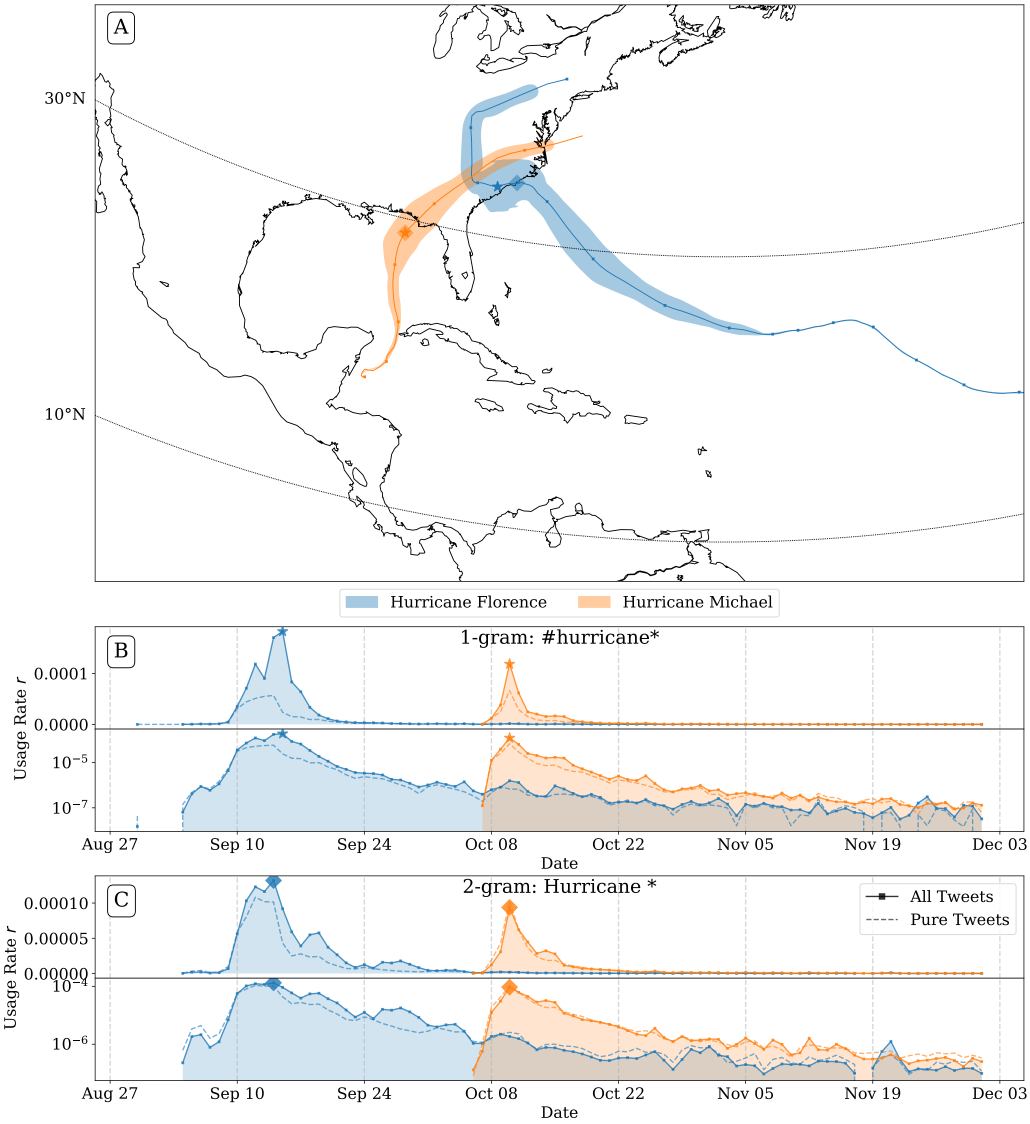


Figure 6.14: Hurricane Attention Map and time series Map and time series for 2018

Bibliography

[1] New state ice co. v. liebmann.

[2] Md Ashraf Ahmed, Arif Mohaimin Sadri, and Piyush Pradhananga. Social media communication patterns of construction industry in major disasters. *Pre-print*, 02 2020.

[3] Laurenz Aisenpreis, Gustav Gyrst, and Vedran Sekara. How do us congress members advertise climate change: An analysis of ads run on meta's platforms. *arXiv preprint arXiv:2304.03278*, 2023.

[4] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[5] Sharon Alajajian, Jake Williams, Andrew Reagan, Stephen Alajajian, Morgan Frank, Lewis Mitchell, Jacob Lahne, Christopher Danforth, and Peter Dodds. The lexicocalorimeter: Gauging public health through caloric input and output on social media. *PLOS ONE*, 12, 07 2015.

[6] Sharon E. Alajajian, Jake Ryland Williams, Andrew J. Reagan, Stephen C. Alajajian, Morgan R. Frank, Lewis Mitchell, Jacob Lahne, Christopher M. Danforth, and Peter Sheridan Dodds. The lexicocalorimeter: Gauging public health through caloric input and output on social media. *PLOS ONE*, 12(2):1–25, 02 2017.

[7] David E Allen and Michael McAleer. President trump tweets supreme leader kim jong-un on nuclear weapons: A comparison with climate change. *Sustainability*, 10(7):2310, 2018.

[8] Thayer Alshaabi, Jane L Adams, Michael V Arnold, Joshua R Minot, David R Dewhurst, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. *Science advances*, 7(29):eabe6534, 2021.

[9] Thayer Alshaabi, Michael V. Arnold, Joshua R. Minot, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, Roby Muhamad, Christopher M. Danforth, and Peter Sheridan Dodds. How the world's collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter. *PLoS ONE*, 16(1):e0244476, 2021.

[10] Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ data science*, 10(1):15, 2021.

[11] Thayer Alshaabi, Colin M. Van Oort, Mikaela Irene Fudolig, Michael V. Arnold, Christopher M. Danforth, and Peter Sheridan Dodds. Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning. *Frontiers in Artificial Intelligence*, 4:783778, 2022.

[12] Zaira Hassan Amur, Yew Kwang Hooi, Hina Bhanbhro, Kamran Dahri, and Gul Muhammad Soomro. Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6):3911, 2023.

[13] Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. Twitter topic classification. *arXiv preprint arXiv:2209.09824*, 2022.

[14] Michael V Arnold, David Rushing Dewhurst, Thayer Alshaabi, Joshua R Minot, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. Hurricanes and hashtags: Characterizing online collective attention for natural disasters. *PLoS one*, 16(5):e0251762, 2021.

[15] Michael V Arnold, Peter Sheridan Dodds, and Christopher M Danforth. Curating corpora with classifiers: A case study of clean energy sentiment online. *arXiv preprint arXiv:2305.03092*, 2023.

[16] Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33(6):712–729, 2015.

[17] Yoshi Meke Bird, Sarah E Grobe, Michael V Arnold, Sean P Rogers, Mikaela I Fudolig, Julia Witte Zimmerman, Christopher M Danforth, and Peter Sheridan Dodds. An assessment of measuring local levels of homelessness through proxy social media signals. *arXiv*, 2023.

[18] Matthew Blaszka, Lauren M Burch, Evan L Frederick, Galen Clavio, and Patrick Walsh. # worldseries: An empirical examination of a Twitter hashtag during a major sporting event. *International Journal of Sport Communication*, 5(4):435–453, 2012.

[19] Leticia Bode and Kajsa E Dalrymple. Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing*, 15(4):311–332, 2016.

[20] David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384, 2018. PMID: 30138075.

[21] US Census Bureau. Hurricanes, Dec 2019. [Online; accessed 4. Dec. 2019].

[22] Randolph Burnside, DeMond Shondell Miller, and Jason D Rivera. The impact of information and risk perception on the hurricane evacuation decision-making of greater new orleans residents. *Sociological Spectrum*, 27(6):727–740, 2007.

[23] Cristian Candia, C Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A Hidalgo. The universal decay of collective memory and attention. *Nature Human Behaviour*, 3(1):82–91, January 2019.

[24] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37, 2021.

[25] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171, 2020.

[26] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[27] Seong Cheol Choi, Xanat Vargas Meza, and Han Woo Park. South korean culture goes latin america: Social network analysis of kpop tweets in mexico. *International Journal of Contents*, 10(1):36–42, 2014.

[28] Yves Citton. *The ecology of attention.* John Wiley & Sons, 2017.

[29] Emily M Cody, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. Public opinion polling with twitter. *arXiv preprint arXiv:1608.02024*, 2016.

[30] 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):e0136092, 2015.

[31] Emily M. Cody, Jennie C. Stephens, James P. Bagrow, Peter Sheridan Dodds, and Christopher M. Danforth. Transitions in climate and energy discourse between hurricanes Katrina and Sandy. *Journal of Environmental Studies and Sciences*, 7(1):87–101, Mar 2017.

[32] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2):317–332, 03 2014.

[33] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008.

[34] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[36] David Rushing Dewhurst, Thayer Alshaabi, Michael V Arnold, Joshua R Minot, Christopher M Danforth, and Peter Sheridan Dodds. Divergent modes of online collective attention to the COVID-19 pandemic are associated with future caseload variance. *arXiv*, 2020.

[37] David Rushing Dewhurst, Thayer Alshaabi, Dilan Kiley, Michael V Arnold, Joshua R Minot, Christopher M Danforth, and Peter Sheridan Dodds. The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series. *EPJ Data Science*, 9(1):3, 2020.

[38] Salomé Do, Étienne Ollion, and Rubing Shen. The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, page 00491241221134526, 2022.

[39] P S Dodds, T Alshaabi, M I Fudolig, J W Zimmerman, J Lovato, S Beaulieu, J R Minot, M V Arnold, A J Reagan, and C M Danforth. Ousiometrics and Telegnomics: The essence of meaning conforms to a two-dimensional powerful-weak and dangerous-safe framework with diverse corpora presenting a safety bias. *arXiv*, 2021.

[40] P S Dodds, J R Minot, M V Arnold, T Alshaabi, J L Adams, D R Dewhurst, A J Reagan, and C M Danforth. Long-term word frequency dynamics derived from Twitter are corrupted: A bespoke approach to detecting and removing pathologies in ensembles of time series. *arXiv*, 2020.

[41] Peter Dodds, Joshua Minot, Michael Arnold, Thayer Alshaabi, Jane Adams, David Dewhurst, Andrew Reagan, and Christopher Danforth. Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop. *Journal of Quantitative Description: Digital Media*, 2, 2022.

[42] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394, 2015.

[43] Peter Sheridan Dodds and Christopher M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[44] 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.

[45] Peter Sheridan Dodds, Joshua R Minot, Michael V Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Tyler J Gray, Morgan R Frank, Andrew J Reagan, and Christopher M Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. *EPJ Data Science*, 12(1):37, 2023.

[46] Peter Sheridan Dodds, Joshua R Minot, Michael V Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, and Christopher M Danforth. Fame and ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and k-pop. *arXiv.org*, September 2019.

[47] Peter Sheridan Dodds, Joshua R Minot, Michael V Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, and Christopher M Danforth. Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions. *arXiv preprint arXiv:2008.13078*, 2020.

[48] Peter Sheridan Dodds, Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, and Christopher M. Danforth. Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions, 2020. Available online at https://arxiv.org/abs/2008.13078.

[49] S N Dorogovtsev and J F F Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62(2):1842–1845, August 2000.

[50] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler. Waiting times in quality of experience for web based services. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 86–96, 2012.

[51] T Eisensee and D Strömberg . News droughts, news floods, and US disaster relief. *The Quarterly Journal of Economics*, 2007.

[52] Georg Franck. Scientific communication–a vanity fair? *Science*, 286(5437):53–55, 1999.

[53] Georg Franck. The economy of attention. *Journal of sociology*, 55(1):8–19, 2019.

[54] Deen Freelon, Charlton D McIlwain, and Meredith Clark. Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[55] Ryan J Gallagher, Morgan R Frank, Lewis Mitchell, Aaron J Schwartz, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4, 2021.

[56] Ryan J Gallagher, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *PLOS ONE*, 13(4):e0195644, 2018.

[57] Ryan J Gallagher, Elizabeth Stowell, Andrea G Parker, and Brooke Foucault Welles. Reclaiming stigmatized narratives: The networked disclosure landscape of# metoo. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[58] Glen R Gallaway. Response times to user activities in interactive man/machine computer systems. In *Proceedings of the Human Factors Society Annual Meeting*, volume 25, pages 754–758. SAGE Publications Sage CA: Los Angeles, CA, 1981.

[59] Shang Gao, Mohammed Alawad, M Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, Antoinette Stroup, et al. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3596–3607, 2021.

[60] Andrew Gelman and Phillip N Price. All maps of parameter estimates are misleading. *Statistics in medicine*, 18(23):3221–3234, 1999.

[61] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[62] Antonio A Ginart, Sanmay Das, Jenine K Harris, Roger Wong, Hao Yan, Melissa Krauss, and Patricia A Cavazos-Rehg. Drugs or dancing? using real-time machine learning to classify streamed "dabbing" homograph tweets. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 10–13. IEEE, 2016.

[63] Michael Golosovsky and Sorin Solomon. Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters*, 109(9):098701, August 2012.

[64] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. *European Economic Review*, 136:103772, 2021.

[65] Kelly Gothard, David Rushing Dewhurst, Joshua R. Minot, Jane Lydia Adams, Christopher M. Danforth, and Peter Sheridan Dodds. The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community, 2021.

[66] Tyler J Gray, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. English verb regularization in books and tweets. *PLOS ONE*, 13(12):e0209651, 2018.

[67] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. *Science advances*, 6(28):eabc2717, 2020.

[68] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[69] Anatoliy Gruzd and Jeffrey Roy. Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet*, 6(1):28–45, 2014.

[70] Marianne Halloran. Analysis finds disaster relief support swift but short, recurring donors crucial | classy, Sep 2018. [Online; accessed 3. Dec. 2019].

[71] Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc_ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52, 2013.

[72] K W Higham, M Governale, A B Jaffe, and U Zülicke. Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. *Physical Review E*, 95(4):042309, April 2017.

[73] K W Higham, M Governale, A B Jaffe, and U Zülicke. Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields. *Journal of Informetrics*, 11(4):1190–1200, November 2017.

[74] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[75] William Housley, Rob Procter, Adam Edwards, Peter Burnap, Matthew Williams, Luke Sloan, Omer Rana, Jeffrey Morgan, Alex Voss, and Anita Greenhill. Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*, 1(2):205395171454513, August 2014.

[76] Ashlee Humphreys and Robert V Kozinets. The construction of value in attention economies. *ACR North American Advances*, 2009.

[77] IAB internet advertising revenue report, May 2018.

[78] Sarah J Jackson, Moya Bailey, and Brooke Foucault Welles. *# HashtagActivism: Networks of race and gender justice*. Mit Press, 2020.

[79] Achin Jain and Vanita Jain. Sentiment classification of Twitter data belonging to renewable energy using machine learning. *Journal of Information and Optimization Sciences*, 40(2):521–533, 2019.

[80] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter. In *the 9th WebKDD and 1st SNA-KDD 2007 workshop*, pages 56–65, New York, New York, USA, 2007. ACM Press.

[81] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[82] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1):72–91, 2016.

[83] Dilan Kiley. *Characterizing the Shapes of Collective Attention using Social Media.* PhD thesis, The University of Vermont, May 2014.

[84] Jisu Kim, Dahye Jeong, Daejin Choi, and Eunil Park. Exploring public perceptions of renewable energy: Evidence from a word network model in social network services. *Energy Strategy Reviews*, 32:100552, 2020.

[85] Serena Y Kim, Koushik Ganesan, Princess Dickens, and Soumya Panda. Public sentiment toward solar energy—opinion mining of Twitter using a transformer-based language model. *Sustainability*, 13(5):2673, 2021.

[86] 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):1–7, 01 2012.

[87] Richard J Ladle, Ricardo A Correia, Yuno Do, Gea Jae Joo, Ana CM Malhado, Raphaël Proulx, Jean Michel Roberge, and Paul Jepson. Conservation culturomics. *Frontiers in Ecology and the Environment*, 14(5):269–275, June 2016.

[88] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866):189–196, 2021.

[89] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[90] Dokyun Lee, Kartik Hosanagar, and Harikesh S Nair. Advertising content and consumer engagement on social media: Evidence from facebook. *Management Science*, 64(11):5105–5131, 2018.

[91] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. *Dynamical classes of collective attention in Twitter.* ACM, New York, New York, USA, April 2012.

[92] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009.

[93] Quanzhi Li, Sameena Shah, Merine Thomas, Kajsa Anderson, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. How much data do you need? twitter decahose data analysis. 07 2016.

[94] Ruopu Li, Jessica Crowe, David Leifer, Lei Zou, and Justin Schoof. Beyond big data: Social media challenges and opportunities for understanding social perception of energy. *Energy Research & Social Science*, 56:101217, 2019.

[95] Brianna A Lienemann, Jennifer B Unger, Tess Boley Cruz, and Kar-Hai Chu. Methods for coding tobacco-related Twitter data: A systematic review. *Journal of medical Internet research*, 19(3):e91, 2017.

[96] Kelsey Linnell, Michael Arnold, Thayer Alshaabi, Thomas McAndrew, Jeanie Lim, Peter Sheridan Dodds, and Christopher M Danforth. The sleep loss insult of spring daylight savings in the us is observable in Twitter activity. *Journal of Big Data*, 8:1–17, 2021.

[97] Kelsey Linnell, Michael Arnold, Thayer Alshaabi, Thomas McAndrew, Jeanie Lim, Peter Sheridan Dodds, and Christopher M. Danforth. The sleep loss insult of Spring Daylight Savings in the US is observable in Twitter activity. *Journal of Big Data*, 8(1):121, 2021.

[98] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

[99] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[100] Clare Llewellyn, Claire Grover, Beatrice Alex, Jon Oberlander, and Richard Tobin. Extracting a topic specific dataset from a Twitter archive. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19*, pages 364–367. Springer, 2015.

[101] Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1–9, April 2019.

[102] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:31, 2011.

[103] Juniper L. Lovato, Antoine Allard, Randall Harp, Jeremiah Onaolapo, and Laurent Hébert-Dufresne. Limits of individual consent and models of distributed consent in online social networks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2251–2262, New York, NY, USA, 2022. Association for Computing Machinery.

[104] Yago Martín, Zhenlong Li, and Susan L. Cutter. Leveraging twitter to gauge evacuation compliance: Spatiotemporal analysis of hurricane matthew. *PLOS ONE*, 12(7):1–22, 07 2017.

[105] Luis Marujo, Wang Ling, Isabel Trancoso, Chris Dyer, Alan W Black, Anatole Gershman, David Martins de Matos, Joao P Neto, and Jaime G Carbonell. Automatic keyword extraction on Twitter. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 637–643, 2015.

[106] Ellen W McGinnis, Shania Lunna, Isabel Berman, Skylar Bagdon, Genevieve Lewis, Michael Arnold, Christopher M Danforth, Peter Sheridan Dodds, Matthew Price, William E Copeland, et al. Expecting the unexpected: Predicting panic attacks from mood and twitter. *medRxiv*, pages 2023–01, 2023.

[107] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[108] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

[110] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[111] Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva Presser, Veres, Adrian, Gray, Matthew K, Team, The Google Books, Pickett, Joseph P, Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A, and Aiden, Erez Lieberman. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 2011.

[112] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[113] Lee M Miller. Collective disaster responses to katrina and rita: Exploring therapeutic community, social capital and social control. *Southern Rural Sociology*, 22(2):45–63, 2007.

[114] Joshua Minot, Milo Trujillo, Samuel Rosenblatt, Guillermo De Anda-Jáuregui, Emily Moog, Allison M Roth, Briane Paul Samson, and Laurent Hébert-Dufresne. Distinguishing in-groups and onlookers by language use. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 157–171, 2022.

[115] Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Christopher M. Danforth, and Peter Sheridan Dodds. Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. *PLOS ONE*, 16(4):e0248880, 2021.

[116] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 554–557, 2011.

[117] Mislove, A, Lehmann, S, Ahn, Y Y, and ICWSM, JP Onnela. Understanding the demographics of Twitter users. *aaai.org*, 2011.

[118] Nic Newman. The rise of social media and its impact on mainstream journalism. *Working Paper*, 2009.

[119] Meredith T. Niles, Benjamin F. Emery, Andrew J. Reagan, Peter Sheridan Dodds, and Christopher M. Danforth. Social media usage patterns during natural hazards. *PLOS ONE*, 14(2):1–16, 02 2019.

[120] Andrzej Nowak, Marta Kacprzyk-Murawska, and Ewa Serwotka. Social psychology and the narrative economy. In *Non-Equilibrium Social Science and Policy*, pages 45–58. Springer, Cham, 2017.

[121] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. volume 11, 01 2010.

[122] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 122–129, 2010.

[123] NStratcom NATO StratCom Centre of Excellence and 2015. Internet trolling as a hybrid warfare tool: The case of Latvia.

[124] Minsu Park, Jennifer Thom, Sarah Mennicken, Henriette Cramer, and Michael Macy. Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour*, 3(3):230–236, 2019.

[125] Eitan A. Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10:e0137041, 2015.

[126] Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015.

[127] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[128] Andrew Perrin. Social media usage. *Pew research center*, pages 52–68, 2015.

[129] Andrew J Reagan, Christopher M Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6(1):1, October 2017.

[130] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[131] Bartley Richardson, Bradley Rees, Tom Drabas, Even Oldridge, David A Bader, and Rachel Allen. Accelerating and expanding end-to-end data science workflows with dl/ml interoperability using rapids. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3503–3504, 2020.

[132] Miguel O Román, Eleanor C Stokes, Ranjay Shrestha, Zhuosen Wang, Lori Schultz, Edil A Sepúlveda Carlo, Qingsong Sun, Jordan Bell, Andrew Molthan, Virginia Kalb, et al. Satellite-based assessment of electricity restoration efforts in puerto rico after hurricane maria. *PloS one*, 14(6), 2019.

[133] Lindsay Ross, Christopher M Danforth, Margaret J Eppstein, Laurence A Clarfeld, Brigitte N Durieux, Cailin J Gramling, Laura Hirsch, Donna M Rizzo, and Robert Gramling. Story arcs in serious illness: Natural language processing features of palliative care conversations. *Patient Educ Couns*, 103(4):826–832, Apr 2020.

[134] 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*, WWW '10, page 851–860, New York, NY, USA, 2010. Association for Computing Machinery.

[135] Sara Salinas. Twitter stock plunges as company blames ad targeting problems for earnings miss. *CNBC*, Oct 2019.

[136] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[137] Kazutoshi Sasahara, Yoshito Hirata, Masashi Toyoda, Masaru Kitsuregawa, and Kazuyuki Aihara. Quantifying collective attention from tweet stream. *PLOS ONE*, 8(4):e61823, April 2013.

[138] Barry Schwartz. Self-determination: The tyranny of freedom. *American psychologist*, 55(1):79, 2000.

[139] Michon Scott. Hurricane Maria's devastation of Puerto Rico, Jan 2020. [Online; accessed 6. Jan. 2020].

[140] Steven C Seow. *Designing and engineering time: The psychology of time perception in software*. Addison-Wesley Professional, 2008.

[141] B Shahbazi. *StoryMiner: An Automated and Scalable Framework for Story Analysis and Detection from Social Media*. PhD thesis, UCLA, 2019.

[142] R.J. Shiller. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, 2019.

[143] Robert J Shiller. Narrative economics. *American Economic Review*, 107(4):967–1004, April 2017.

[144] Sarah Shugars, Adina Gitomer, Stefan McCabe, Ryan J Gallagher, Kenneth Joseph, Nir Grinberg, Larissa Doroshenko, Brooke Foucault Welles, and David Lazer. Pandemics, protests, and publics: Demographic activity and engagement on Twitter in 2020. *Journal of Quantitative Description: Digital Media*, 1, 2021.

[145] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLOS ONE*, 10(3):e0115545, March 2015.

[146] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):74–84, 2013.

[147] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[148] Zachary C Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):1–9, 2015.

[149] Anne Marie Stupinski, Thayer Alshaabi, Michael V Arnold, Jane Lydia Adams, Joshua R Minot, Matthew Price, Peter Sheridan Dodds, and Christopher M Danforth. Quantifying changes in the language used around mental health on Twitter over 10 years: Observational study. *JMIR mental health*, 9(3):e33685, 2022.

[150] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. The DARPA Twitter bot challenge. *Computer*, 49(6):38–46, June 2016.

[151] Harvey Thurm Taylor, Bill Ward, Mark Willis, and Walt Zaleski. The Saffir-Simpson hurricane wind scale. *Atmospheric Administration: Washington, DC, USA*, 2010.

[152] Zeynep Tufekci. “Not this one” social movements, the attention economy, and microcelebrity networked activism. *American Behavioral Scientist*, 57(7):848–870, 2013.

[153] Twitter decahose.

[154] Oskar Vågerö, Anders Bråte, Alexandra Wittemann, Jessica Yarin Robinson, Natalia Sirotko-Sibirskaya, and Marianne Zeyringer. Machine learning of public sentiments toward wind energy in norway. *arXiv preprint arXiv:2304.02388*, 2023.

[155] Sergi Valverde, Ricard V Solé, Mark A Bedau, and Norman Packard. Topology and evolution of technology innovation networks. *Physical Review E*, 76(5):056118, November 2007.

[156] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[157] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, October 2013.

[158] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[159] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[160] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[161] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[162] Zheye Wang, Nina S.N. Lam, Nick Obradovich, and Xinyue Ye. Are vulnerable communities digitally left behind in social responses to natural disasters? an evidence from Hurricane Sandy with Twitter data. *Applied Geography*, 108:1 – 8, 2019.

[163] Morgan Weaving, Thayer Alshaabi, Michael V Arnold, Khandis Blake, Christopher M Danforth, Peter S Dodds, Nick Haslam, and Cordelia Fine. Twitter misogyny associated with Hillary Clinton increased throughout the 2016 U.S. election campaign. *Scientific Reports*, 13(1):5266, 2023.

[164] Jessica Weinkle, Chris Landsea, Douglas Collins, Rade Musulin, Ryan P Crompton, Philip J Klotzbach, and Roger Pielke. Normalized hurricane damage in the continental united states 1900–2017. *Nature Sustainability*, 1(12):808–813, 2018.

[165] Wen-Yuan Liu, Bao-Wen Wang, Jia-Xin Yu, Fang Li, Shui-Xing Wang, and Wen-Xue Hong. Visualization classification method of multi-dimensional data based on radar chart mapping. In *2008 International Conference on Machine Learning and Cybernetics*, volume 2, pages 857–862, July 2008.

[166] Wikipedia contributors. 2010 atlantic hurricane season — Wikipedia, the free encyclopedia, 2019. [Online; accessed 31-October-2019].

[167] Jake Ryland Williams, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. Text mixing shapes the anatomy of rank-frequency distributions. *Phys. Rev. E*, 91:052811, May 2015.

[168] Charley E Willison, Phillip M Singer, Melissa S Creary, and Scott L Greer. Quantifying inequities in US federal response to hurricane disaster in Texas and Florida compared with Puerto Rico. *BMJ Global Health*, 4(1):e001191, January 2019.

[169] Stefan Wojcik and Adam Hughes. How Twitter users compare to the general public, Apr 2019. [Online; accessed 7. Jan. 2020].

[170] Stefan Wojcik and Adam Hughes. Sizing up Twitter users. *PEW research center*, 24:1–23, 2019.

[171] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[172] Fang Wu and Bernardo A Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, November 2007.

[173] Henry H. Wu, Ryan J. Gallagher, Thayer Alshaabi, Jane L. Adams, Joshua R. Minot, Michael V. Arnold, Brooke Foucault Welles, Randall Harp, Peter Sheridan Dodds, and Christopher M. Danforth. Say their names: Resurgence in the collective attention

toward Black victims of fatal police violence following the death of George Floyd. *PLOS ONE*, 18(1):1–26, 01 2023.

[174] Henry H. Wu, Ryan J. Gallagher, Thayer Alshaabi, Jane L. Adams, Joshua R. Minot, Michael V. Arnold, Brooke Foucault Welles, Randall Harp, Peter Sheridan Dodds, and Christopher M. Danforth. Say their names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd. *PLOS ONE*, 18(1):e0279225, 2023.

[175] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[176] Clayton Wukich and Alan Steinberg. Nonprofit and public sector participation in self-organizing information networks: Twitter hashtag and trending topic use during disasters. *Risk, Hazards & Crisis in Public Policy*, 4(2):83–109, June 2013.

[177] Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810, 2018.

[178] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103, 2020.

[179] Kai-Cheng Yang, Onur Varol, Alexander C Nwala, Mohsen Sayyadiharikandeh, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Social bots: Detection and challenges. *arXiv preprint arXiv:2312.17423*, 2023.

[180] Yaming Zhang, Majed Abbas, and Wasim Iqbal. Perceptions of ghg emissions and renewable energy sources in europe, australia and the usa. *Environmental Science and Pollution Research*, 29(4):5971–5987, 2022.

[181] Carmen D. Zorrilla. The view from Puerto Rico — hurricane Maria and its aftermath. *New England Journal of Medicine*, 377(19):1801–1803, 2017. PMID: 29019710.