Disentangling Discourse: Networks, Entropy, and Social Movements

A Thesis Presented

by

Ryan J. Gallagher

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Applied Mathematics

May, 2017

Defense Date: March 21, 2017
Dissertation Examination Committee:

Christopher Danforth, Ph.D., Advisor
Peter Sheridan Dodds, Ph.D., Advisor
Josh Bongard, Ph.D., Chairperson
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

Our daily online conversations with friends, family, colleagues, and strangers weave an intricate network of interactions. From these networked discussions emerge themes and topics that transcend the scope of any individual conversation. In turn, these themes direct the discourse of the network and continue to ebb and flow as the interactions between individuals shape the topics themselves. This rich loop between interpersonal conversations and overarching topics is a wonderful example of a complex system: the themes of a discussion are more than just the sum of its parts.

Some of the most socially relevant topics emerging from these online conversations are those pertaining to racial justice issues. Since the shooting of Black teenager Michael Brown by White police officer Darren Wilson in Ferguson, Missouri, the protest hashtag #BlackLivesMatter has amplified critiques of extrajudicial shootings of Black Americans. In response to #BlackLivesMatter, other online users have adopted #AllLivesMatter, a counter-protest hashtag whose content argues that equal attention should be given to all lives regardless of race. Together these contentious hashtags each shape clashing narratives that echo previous civil rights battles and illustrate ongoing racial tension between police officers and Black Americans.

These narratives have taken place on a massive scale with millions of online posts and articles debating the sentiments of "black lives matter" and "all lives matter." Since no one person could possibly read everything written in this debate, comprehensively understanding these conversations and their underlying networks requires us to leverage tools from data science, machine learning, and natural language processing. In Chapter 2, we utilize methodology from network science to measure to what extent #BlackLivesMatter and #AllLivesMatter are "slacktivist" movements, and the effect this has on the diversity of topics discussed within these hashtags. In Chapter 3, we precisely quantify the ways in which the discourse of #BlackLivesMatter and #AllLivesMatter diverge through the application of information-theoretic techniques, validating our results at the topic level from Chapter 2. These entropy-based approaches provide the foundation for powerful automated analysis of textual data, and we explore more generally how they can be used to construct a human-in-the-loop topic model in Chapter 5. Our work demonstrates that there is rich potential for weaving together social science domain knowledge with computational tools in the study of language, networks, and social movements.

Citations

Material from this thesis has been submitted for publication to *EPJ Data Science* on June 28th, 2016 in the following form:

Gallagher, R. J. and Reagan, A. R. and Danforth, C. M. and Dodds, P. S.. (2016). Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter. *EPJ Data Science.*

AND

Material from this thesis has been submitted for publication to *Transactions of the Association of Computational Linguistics (TACL)* on November 30th, 2016 in the following form:

Gallagher, R. J. and Reing, K. and Kale, D. and Ver Steeg, G.. (2016). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Transactions of the Association of Computational Linguistics.*

Dedication

To my mother and father for all the sacrifices they have made for me

Acknowledgements

This thesis and my transformation into a budding complex systems researcher would not have been possible without the support, guidance, and imagination of Chris Danforth, Peter Dodds, and Greg Ver Steeg. I am deeply grateful for their mentorship. The ensemble of Chris Fusting, Abby Ross, Justin Foster, Dilan Kiley, Kelsey Linnell, Jack Dalton, Garvin Gaston, Kewang Chen, and the Computational Story Lab helped me navigate graduate school as a part of a wonderfully supportive team. In addition, Ada Morse stood in as an excellent math mentor who always helped me through some of my hardest problems, and Andy Reagan helped me aspire to be at his level of data wizardry. Thank you to Nicole Beauregard, Devyn Neveux, Graeson McMahon, Marissa Lily, and Spencer Janes for being excellent supportive friends. Finally, much love to Rachel Stewart for providing constant reassurance and encouragement through the hardest parts of this journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Language and Networks as Complex Systems

Our social spheres are one of the most classic examples of a network. The interactions we partake in with our friends, colleagues, teachers, and family form a network that we visualize as a collection of dots and lines: dots represent people, and lines connecting these dots represent an interaction between two people. Typically we instead refer to dots as nodes and lines as edges, and we can imagine embedding a variety of information into these networks. For instance, we may wish to classify student-teacher interactions differently than parent-child interactions or perhaps we would like prioritize the importance of interactions by counting the number of them that occur between any two people. We may label nodes of the network with information about what role that person plays in a network, and we could even allow the network in its entirety to vary over time. It is quite natural to want to ask questions about all of thse aspects and how they affect the structure of a social network.

Indeed, sociologists have considered questions about social networks for quite some time now **??**, but it is only with the uptake of social media that reseachers have been able to study social networks at massive scales. Online network data has allowed reseachers to study a pleothora of fascinating questions, including how emotions spread among friends **??**, how peer influence affects voter turnout **??**, and how groups respond to crisis events **??**. While, of course, these online interactions are only a proxy for face-to-face social networks, they provide an avenue for studying social behavior in a variety of settings.

Language provides one of the primary avenues for which we can carry out our social interactions. As with networks, the widespread adpotion of the internet has produced a wealth of data to study the nuances and impact of language. This has launched studies into how to measure the evolution of language over time **??**, how to quantify the sentiment and emotions of conversations **??**, and how to extract topics and themse from books, articles, and social media posts **??**. Network and textual data coupled together allow us to understand not only *what* is being said, but *who* is saying it and *how* it is spreading.

These intricate interacations and feedbacks between people and their language are hallmarks of complex systems. While the definition of a complex system is notoriously hard to pin down, they often consist of indvidual parts that interact, sometimes in an adaptive manner, in such a way that produces *emergent* phenomenon that would not arise from any individual alone. Who we choose to interact and what we say evolves adaptively over time according to the conversations and interactions that we have had in the past. From these many discussions arise themes and topics that transcends any individual person or conversation, and these topics themselves adapt as the underlying conversations continue to change.

1.2 Black Lives Matter

Perhaps one of the most socially relevant examples of networks and language interacting together is that of social and protest movements. Protest movements have a long history of forcing difficult conversations in order to enact social change, and the increasing prominence of social media has allowed these conversations to be shaped in new and complex ways. Indeed, significant attention has been given to how to quantify the dynamics of such social movements. Recent work has studied movements such as Occupy Wall Street [2–4], the Arab Spring [5], and large-scale protests in Egypt and Spain [6, 7], and how they evolved with respect to their causes. The network structures of movements have also been leveraged to answer questions about how protest networks facilitate information diffusion [8], align with geospatial networks [9], and impact offline activism [10–12]. Both offline and online activists have been shown to be crucial to the formation of protest networks [13, 14] and play a critical role in the eventual tipping point of social movements [15, 16].

The protest hashtag #BlackLivesMatter has come to represent a major social movement. The hashtag was started by three women, Alicia Garza, Patrisse Cullors, and Opal Tometi, following the death of Trayvon Martin, a Black teenager who was shot and killed by neighborhood watchman George Zimmerman in February 2012 [17]. The hashtag was a "call to action" to address anti-Black racism, but it was not until November 2014 when White Ferguson police officer Darren Wilson was not indicted for the shooting of Michael Brown that #BlackLivesMatter saw widespread use. Since then, the hashtag has been used in combination with other hashtags, such as #EricGarner, #FreddieGray, and #SandraBland, to highlight the extrajudicial deaths of other Black Americans. #BlackLivesMatter has organized the conversation surrounding the broader Black Lives Matter movement and activist organization of the same name. Some have likened Black Lives Matter to the New Civil Rights movement [18, 19], though the founders reject the comparison and self-describe Black Lives Matter as a "human rights" movement [20].

Researchers have only just begun to study the emergence and structure of #BlackLivesMatter and its associated movement. To date and to the best of our knowledge, Freelon et al. have provided the most comprehensive data-driven study of Black Lives Matter [21]. Their research characterizes the movement through multiple frames and analyzes how Black Lives Matter has evolved as a movement both online and offline. Other researchers have given particular attention to the beginnings of the movement and its relation to the events of Ferguson, Missouri. Jackson and Welles have shown that the initial uptake of #Ferguson, a hashtag that precluded widespread use of #BlackLivesMatter, was due to the early efforts of "citizen journalists" [22], and Bonilla and Rosa argue that these citizens framed the story of Michael Brown in such a way that facilitated its eventual spreading [23]. Other related work has attempted to characterize the demographics of #BlackLivesMatter users [24], how #BlackLivesMatter activists effect systematic political change [25], and how the movement self-documents itself through Wikipedia [26]. Our work spans a larger time scale than any of the previous works, covering not only events such as the Ferguson and Baltimore protests, but also the death of Sandra Bland, the 2015 Grammy Awards, and the shootings in Chapel Hill, North Carolina and Charleston, South Carolina. Furthermore, our work provides a comprehensive view of the major topics and events occurring within

#BlackLivesMatter during all of these events.

#BlackLivesMatter has found itself contended by a relatively vocal counter-protest hashtag: #AllLivesMatter. Advocates of #AllLivesMatter affirm that equal attention should be given to all lives, while #BlackLivesMatter supporters contend that such a sentiment derails the Black Lives Matter movement. The counter-hashtag #AllLivesMatter has received less attention in terms of research, largely being studied from a theoretical angle. The phrase “All Lives Matter” reflects a “race-neutral” or “color-blind” approach to racial issues [27]. While this sentiment may be “laudable,” it is argued that race-neutral attitudes can mask power inequalities that result from racial biases [28]. From this perspective, those who adopt #AllLivesMatter evade the importance of race in the discussion of Black deaths in police-involved shootings [29, 30]. To our knowledge, our work is the first to engage in a data-driven approach to understanding #AllLivesMatter. This approach not only allows us to substantiate several broad claims about the use of #AllLivesMatter, but to also highlight trends in #AllLivesMatter that are absent from the theoretical discussion of the hashtag.

Together, #BlackLivesMatter and #AllLivesMatter are examples of politically polarized groups. Such polarization has mostly been studied within the context of the more traditional political sphere [31–33], although some research has examined protest polarization of Occupy Wall Street views [34] and secularist versus Islamic views in Egypt during the Arab Spring [35, 36]. Like the aforementioned studies of social movements, much of the research on political polarization has focused on the network structure of these polarized groups. In the cases where textual content analysis has been utilized, researchers have examined trends of various hashtags and important terms specified by domain experts [32,33,35,36]. In studying just hashtag trends and predetermined lists of terms, these analyses have discarded a significant portion of the textual data. Our work systematically analyzes the divergence of #BlackLivesMatter and #AllLivesMatter across all words and hashtags, allowing us to discern important themes that do not otherwise emerge. In particular, this approach gives us a new perspective on the use "hijacking" [37] and "content injection" [31].

Chapter 2

Networked Slacktivism

Chapter abstract goes here.

2.1 Data Collection

We collected tweets containing #BlackLivesMatter and #AllLivesMatter (case-insensitive) from the period August 8th, 2014 to August 31st, 2015 from the Twitter Gardenhose feed. The Gardenhose represents a 10% random sample of all public tweets. Our subsample resulted in 767,139 #BlackLivesMatter tweets from 375,620 unique users and 101,498 #AllLivesMatter tweets from 79,753 unique users. Of these tweets, 23,633 of them contained both hashtags. When performing our analyses, these tweets appear in each corpus.

Previous work has emphasized the importance of viewing protest movements through small time scales [21, 22]. In addition, we do not attempt to characterize all of the narratives that exist within #BlackLivesMatter and #AllLivesMatter. Therefore, we choose to restrict our analysis to eight one-week periods where there were simultaneous spikes in #BlackLivesMatter and #AllLivesMatter. These one-week periods are labeled on Figure 2.1 and are as follows:

1. *November 24th, 2014*: the non-indictment of Darren Wilson in the death of Michael Brown [38].

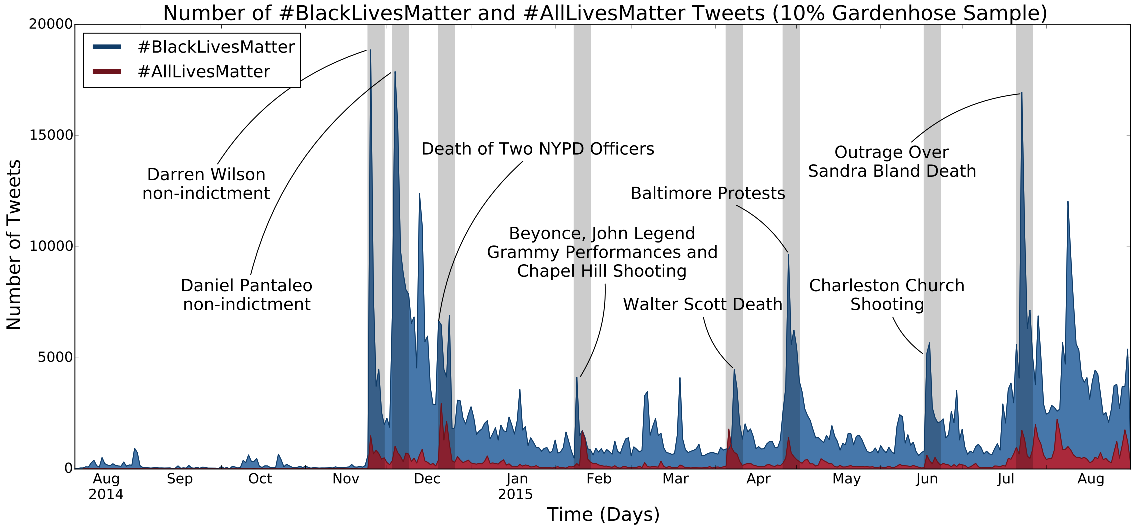


Figure 2.1: Time series showing the number of #BlackLivesMatter and #AllLivesMatter tweets from Twitter's 10% Gardenhose sample. The plot is annotated with several major events pertaining to the hashtags. Shaded regions indicate one-week periods where use of both #BlackLivesMatter and #AllLivesMatter peaked in frequency. These are the periods we focus on in the present study.

2. *December 3rd, 2014*: the non-indictment of Daniel Pantaleo in the death of Eric Garner [39].

3. *December 20th, 2014*: the deaths of New York City police officers Wenjian Liu and Rafael Ramos [40].

4. *February 8th, 2015*: the Chapel Hill shooting and the 2015 Grammy performances by Beyonce and John Legend [41, 42].

5. *April 4th, 2015*: the death of Walter Scott [43].

6. *April 26th, 2015*: the social media peak of protests in Baltimore over the death of Freddie Gray [44].

7. *June 17th, 2015*: the Charleston Church shooting [45].

8. *July 21st, 2015*: outrage over the death of Sandra Bland [46].

2.2 Methods

2.2.1 k-Core Decomposition

2.2.2 Multiscale Backbone

Reminiscent of finding the core of a network, we may also be interested in finding the "backbone" of a weighted network. Unlike an uncovered core though, the nodes and edges contained within the backbone may not necessarily have high centrality, but instead are represenative of the structure of the full network. To extract such a mesoscale structure, we apply the disparity filter, a method introduced by Serrano et al. that yields the multiscale backbone of a weighted network [1]. On a node-by-node basis, the disparity filter compares the distribution of the weighted edges to a uniform null model. More specifically, given node i of degree k, we first normalize the node's weight distribution. We are then interested in edges whose weights deviate significantly from a null assumption that the weights are uniformly distributed. So, for each neighbor j of node i with normalized edge weight p_{ij}, we calculate the quantity

$$\alpha_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}\, dx. \tag{2.1}$$

For a specified significance level α, we say the statistically significant edges with respect to the uniform null model are those satisfying $\alpha_{ij} < \alpha$. The disparity filter provides a method for capturing the most significant and representative aspect of a network while discarding more spurious connections.

2.3 Topic Networks

We wish to understand the broad topics discussed within #BlackLivesMatter and #AllLivesMatter and how these topics evolve with respect to the underlying user network. Previous work on political polarization has used hashtags as a proxy for topics [31, 33, 35, 47, 48] and here we use the same interpretation. However, not all hashtags assist in understanding the broad topics. For example, #retweet and #lol are two such hashtags that frequently appear in tweets, but they provide no evidently relevant information about the events that are being discussed. Thus, we require a way of uncovering the most important topics and how they connect to one another.

To find these topics, we first construct hashtag networks for each of #BlackLivesMatter and #AllLivesMatter, where nodes are hashtags and weighted edges denote co-occurrence of these hashtags within a tweet. We take the topic network to be the largest connected component of the backbone. For significance level $\alpha < 0.03$, the disparity filter begins to force drastic drops in the number of nodes removed from the original hashtag network, as shown in Figure **??**. For this reason, we analyze the backbones only for $0.03 \leq \alpha \leq 0.05$.

Example visualizations of the topic networks following the week of the death of the two NYPD officers are presented in Figures 2.3–2.4. Node sizes are proportional to the square root of the number of times each hashtag was used, and node colors

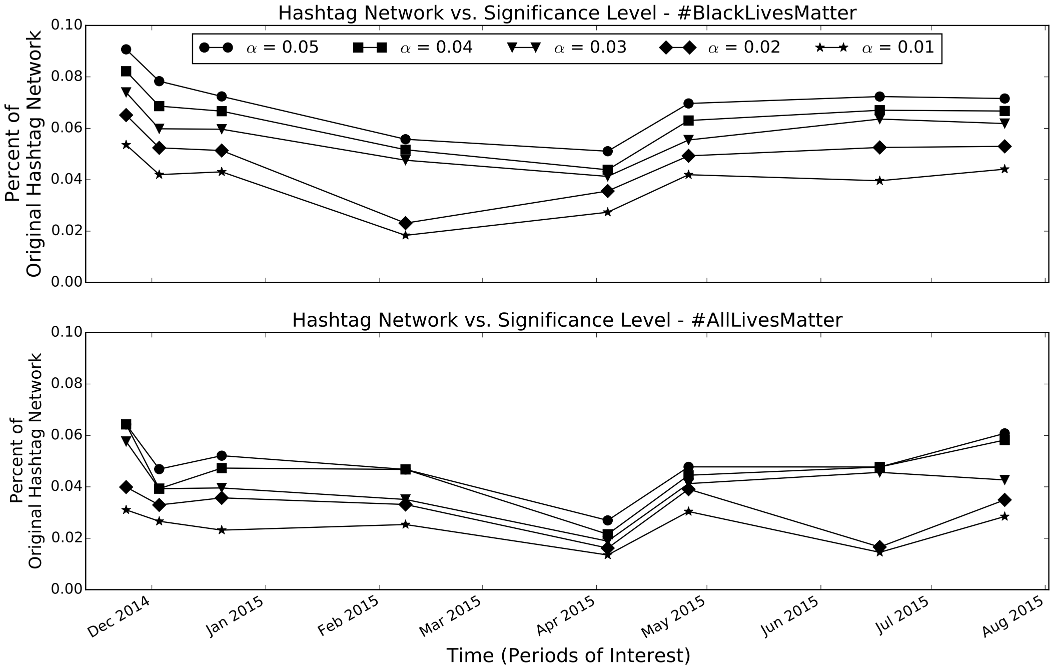


Figure 2.2: Percent of original hashtag network maintained for #BlackLivesMatter (top) and #AllLivesMatter (bottom) at each of the periods of interest for varying levels of the disparity filter significance level. We wish to filter as much of the network as possible, while avoiding sudden reductions in the number of nodes in the network. Note, when going from $\alpha = 0.03$ *to* $\alpha = 0.02$*, the February 8th #BlackLivesMatter and July 21st #AllLivesMatter networks fall in size by a factor of approximately one half. Therefore, we restrict to* $\alpha \geq 0.03$.

are determined by the Louvain structure detection method [49]. The exact assignments of topics to communities is not critical, but rather they provide a visual guide through the networks.

Table 2.1 reports the number of nodes, edges, clustering coefficients, and percentages of nodes maintained from the full hashtag networks for significance level $\alpha = 0.03$. We see that across all time periods of interest, the number of topics in #BlackLivesMatter is higher than that of #AllLivesMatter. We also note that the clustering of the #BlackLivesMatter topics is less that of #AllLivesMatter almost always. Thus, not only are there more topics presented in #BlackLivesMatter, but

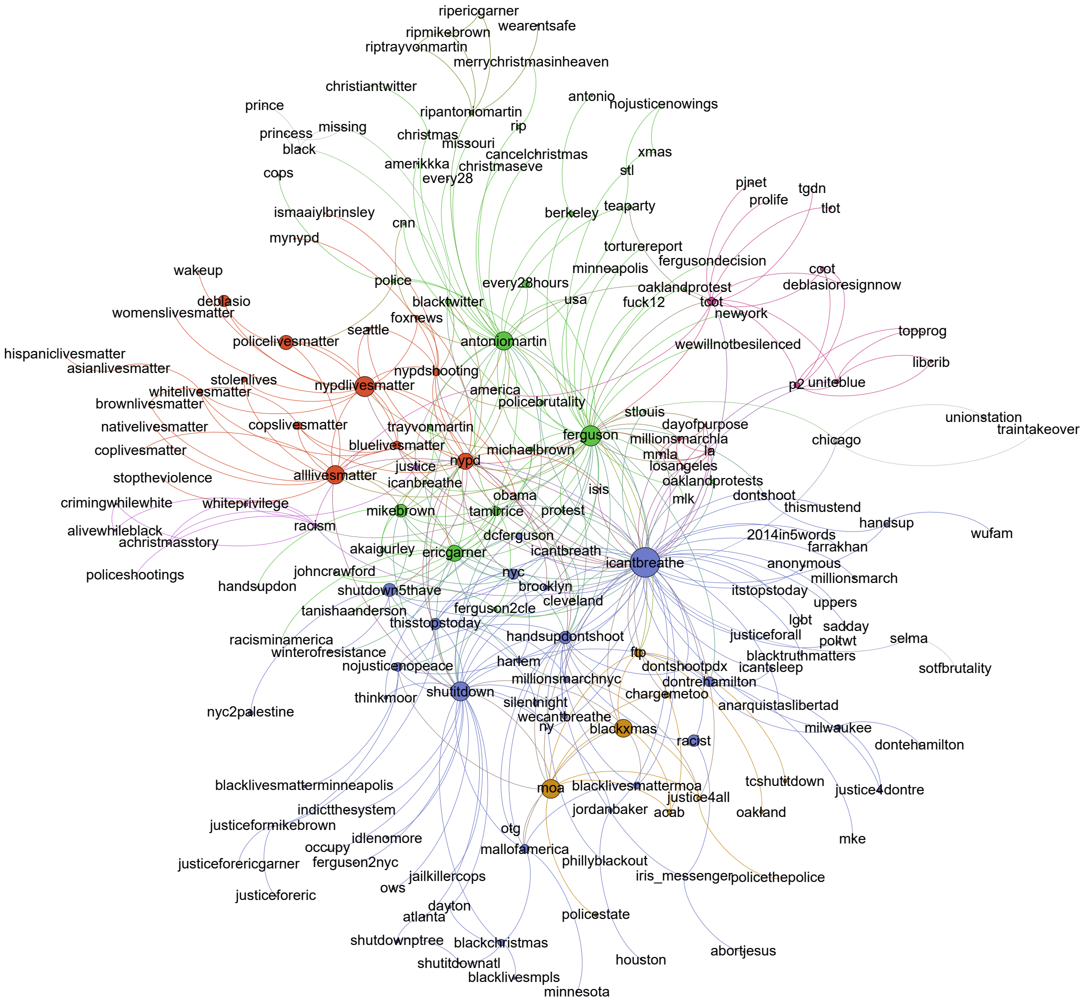


Figure 2.3: #BlackLivesMatter topic network for the week following the death of two NYPD officers. The network is constructed by first constructing a network of co-occurrences of hashtags and then applying the disparity filter to find the multiscale backbone of the hashtag network. This particular network is for significance level α = 0.03*.*

they are more diverse in their connections. In contrast, the stronger #AllLivesMatter ties, as measured by their clustering, suggest that the #AllLivesMatter topics are more tightly connected and revolve around similar themes. We see the clustering is less within #AllLivesMatter during the week of Walter Scott's death, where the topic network has a star-like shape with no triadic

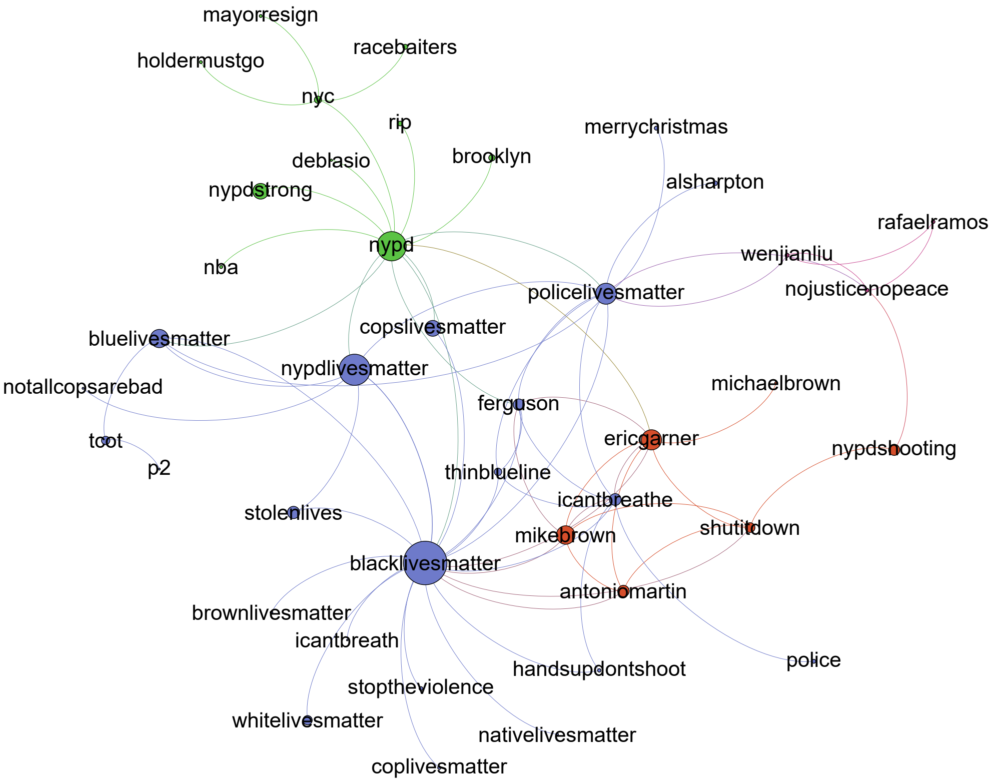


Figure 2.4: #AllLivesMatter topic network for the week following the death of two NYPD officers. The network is constructed by first constructing a network of co-occurrences of hashtags and then applying the disparity filter to find the multiscale backbone of the hashtag network. This particular network is for significance level $\alpha = 0.03$.

closure across all significance levels. This low clustering is not indicative of diverse conversation, as the central node #BlackLivesMatter connects several disparate topics. As we will show, the discussion within #AllLivesMatter was dominated by a retweet not pertaining to the death of Walter Scott, the event of that time period. Note, these conclusions also hold for the topic networks at significance levels $\alpha = 0.04$ and $\alpha = 0.05$.

In order to extract the most central topics of #BlackLivesMatter and #AllLivesMatter during each time period, we compare the results of three centrality

#BlackLivesMatter	Nodes	% Original Nodes	Edges	Clustering
Nov. 24–Nov. 30, 2014	243	7.39%	467	0.0605
Dec. 3–Dec. 9, 2014	339	5.98%	794	0.0691
Dec. 20–Dec. 26, 2014	187	5.96%	391	0.1635
Feb. 8–Feb. 14, 2015	70	4.75%	94	0.1740
Apr. 4–Apr. 10, 2015	80	4.12%	114	0.1019
Apr. 26–May 2, 2015	234	5.54%	471	0.1068
Jun. 17–Jun. 23, 2015	167	6.35%	246	0.0746
Jul. 21–Jul. 27, 2015	216	6.18%	393	0.0914
#AllLivesMatter				
Nov. 24–Nov. 30, 2014	26	5.76%	35	0.1209
Dec. 3–Dec. 9, 2014	31	3.92%	49	0.2667
Dec. 20–Dec. 26, 2014	41	3.95%	70	0.2910
Feb. 8–Feb. 14, 2015	18	3.50%	23	0.1894
Apr. 4–Apr. 10, 2015	7	1.88%	6	0.0000
Apr. 26–May 2, 2015	38	4.12%	62	0.1868
Jun. 17–Jun. 23, 2015	22	4.56%	28	0.3571
Jul. 21–Jul. 27, 2015	33	4.26%	44	0.1209

Table 2.1: Summary statistics for topic networks created from the full hashtag networks using the disparity filter at the significance level α = 0.03.

measures, betweenness centrality, random walk closeness centrality, and PageRank on the topic networks at significance level $\alpha = 0.03$. Through inspection of the rankings of each list, we find the relative rankings of the most central topics in #BlackLivesMatter and #AllLivesMatter are robust to the centrality measure used. Table 2.2 shows the rankings according to random walk centrality.

We see that for both #BlackLivesMatter and #AllLivesMatter, the top identified topics are indicative of the relative events occurring in each time period. In #BlackLivesMatter for instance, #ferguson and #mikebrown are top topics after the non-indictment of Darren Wilson, #walterscott is a top topic after the death of Walter Scott, and #sandrabland and #sayhername are a top topics during the time period following the death of Sandra Bland. However, these major topics rank

	#BlackLivesMatter		#AllLivesMatter	
	Top Hashtags	Betweenness	Top Hashtags	Betweenness
Nov. 24–Nov. 30, 2014	1. ferguson	0.8969	1. blacklivesmatter	0.7229
	2. mikebrown	0.2303	2. ferguson	0.6079
	3. fergusondecision	0.1159	3. fergusondecision	0.1791
	4. shutitdown	0.0854	4. mikebrown	0.1730
	5. justiceformikebrown	0.0652	5. nycprotest	0.0303
	6. tamirrice	0.0583	6. williamsburg	0.0192
	7. blackoutblackfriday	0.0577	7. brownlivesmatter	0.0158
	8. alllivesmatter	0.0512	8. sf	0.0150
	9. boycottblackfriday	0.0442	9. blackfridayblackout	0.0137
	10. blackfriday	0.0439	10. nyc	0.0123
Dec. 3–Dec. 9, 2014	1. ericgarner	0.6221	1. blacklivesmatter	0.6589
	2. icantbreathe	0.5035	2. ericgarner	0.4396
	3. ferguson	0.2823	3. ferguson	0.3684
	4. shutitdown	0.1117	4. icantbreathe	0.2743
	5. mikebrown	0.0745	5. tcot	0.1916
	6. thisstopstoday	0.0505	6. shutitdown	0.1529
	7. handsupdontshoot	0.0469	7. handsupdontshoot	0.0797
	8. nojusticenopeace	0.0449	8. crimingwhilewhite	0.0666
	9. nypd	0.0399	9. miami	0.0666
	10. berkeley	0.0352	10. mikebrown	0.0645
Dec. 20–Dec. 26, 2014	1. icantbreathe	0.5742	1. blacklivesmatter	0.6791
	2. ferguson	0.3234	2. nypd	0.4486
	3. antoniomartin	0.2826	3. policelivesmatter	0.2926
	4. shutitdown	0.2473	4. nypdlivesmatter	0.1990
	5. nypd	0.1496	5. ericgarner	0.1565
	6. alllivesmatter	0.1324	6. nyc	0.1461
	7. ericgarner	0.1265	7. bluelivesmatter	0.1409
	8. nypdlivesmatter	0.1147	8. icantbreathe	0.1254
	9. tcot	0.1082	9. shutitdown	0.0992
	10. handsupdontshoot	0.0986	10. mikebrown	0.0860
Feb. 8–Feb. 14, 2015	1. blackhistorymonth	0.6506	1. muslimlivesmatter	0.8012
	2. grammys	0.5614	2. blacklivesmatter	0.4296
	3. muslimlivesmatter	0.5515	3. chapelhillshooting	0.4192
	4. bhm	0.4987	4. jewishlivesmatter	0.2279
	5. alllivesmatter	0.4932	5. butinacosmicsensenothingreallymatters	0.0431
	6. handsupdontshoot	0.4150	6. whitelivesmatter	0.0112
	7. mikebrown	0.2588	7. rip	0.0073
	8. ferguson	0.1754	8. hatecrime	0.0071
	9. icantbreathe	0.1614	9. ourthreewinners	0.0058
	10. beyhive	0.1325		
Apr. 4–Apr. 10, 2015	1. walterscott	0.9118	1. blacklivesmatter	1.0000
	2. blacktwitter	0.2283		
	3. icantbreathe	0.1779		
	4. ferguson	0.1555		
	5. p2	0.1022		
	6. ericgarner	0.1003		
	7. alllivesmatter	0.0906		
	8. mlk	0.0717		
	9. kendrickjohnson	0.0685		
	10. tcot	0.0617		
Apr. 26–May 2, 2015	1. freddiegray	0.5806	1. blacklivesmatter	0.7227
	2. baltimore	0.4732	2. baltimoreriots	0.4339
	3. baltimoreuprising	0.2625	3. baltimore	0.3463
	4. baltimoreriots	0.2415	4. freddiegray	0.1869
	5. alllivesmatter	0.1053	5. policelivesmatter	0.1106
	6. mayday	0.0970	6. baltimoreuprising	0.1014
	7. blackspring	0.0737	7. tcot	0.0663
	8. tcot	0.0581	8. peace	0.0451
	9. baltimoreuprising	0.0500	9. whitelivesmatter	0.0444
	10. handsupdontshoot	0.0458	10. wakeupamerica	0.0281
Jun. 17–Jun. 23, 2015	1. charlestonshooting	0.8849	1. charlestonshooting	0.6666
	2. charleston	0.2551	2. blacklivesmatter	0.6238
	3. blacktwitter	0.1489	3. bluelivesmatter	0.4900
	4. tcot	0.1379	4. gunsense	0.2571
	5. unitedblue	0.1340	5. pjnet	0.2292
	6. racism	0.1323	6. 2a	0.1857
	7. ferguson	0.1085	7. wakeupamerica	0.1494
	8. usa	0.1011	8. tcot	0.1289
	9. takedowntheflag	0.0790	9. gohomederay	0.0952
	10. baltimore	0.0788	10. ferguson	0.0952
Jul. 21–Jul. 27, 2015	1. sandrabland	0.7802	1. blacklivesmatter	0.8404
	2. sayhername	0.3175	2. pjnet	0.2689
	3. justiceforsandrabland	0.1994	3. tcot	0.2683
	4. unitedblue	0.1870	4. uniteblue	0.2440
	5. blacktwitter	0.1788	5. defundplannedparenthood	0.2437
	6. alllivesmatter	0.1648	6. defundpp	0.1692
	7. tcot	0.1081	7. sandrabland	0.1386
	8. defundpp	0.0827	8. justiceforsandrabland	0.1386
	9. p2	0.0756	9. prolife	0.0881
	10. m4bl	0.0734	10. nn15	0.0625

Table 2.2: The top 10 hashtags in the topic networks as determined by random walk centrality for each time period. Some #AllLivesMatter topic networks have less than 10 top nodes due to the relatively small size of the networks.

differently in both #BlackLivesMatter and #AllLivesMatter. For instance, while #mikebrown, #ericgarner, #icantbreathe, #freddiegray, #baltimore, and #sandrabland all consistently rank higher in #BlackLivesMatter than in #AllLivesMatter.

The most prominent discussion of non-Black lives in the topic networks of #AllLivesMatter is discussion of police lives. We see that in #AllLivesMatter, #nypd, #policelivesmatter, and #bluelivesmatter are ranked higher as topics in #AllLivesMatter than in #BlackLivesMatter during December 20th and April 26th periods, similar to what we found in the JSD word shifts. On the other hand, hashtags depicting strong anti-police sentiment such as #killercops, #policestate, and #fuckthepolice appear almost exclusively in #BlackLivesMatter and are absent from #AllLivesMatter. The alignment of #AllLivesMatter with police lives coincides with a broader alignment with the conservative sphere of Twitter that is apparent through the topic networks. In several periods for #AllLivesMatter, #tcot is a central topic, as well as #pjnet (Patriots Journal Network), #wakeupamerica, and #defundplannedparenthood. The hashtag #tcot also appears in several of the #BlackLivesMatter periods as well. This is to be expected, as Freelon et al. found that a portion of #BlackLivesMatter tweets were hijacked by the conservative sphere of Twitter [21].

However, the hijacking of #BlackLivesMatter and content injection of conservative views is a much smaller component of the #BlackLivesMatter topics as compared to the respective hijacking of the #AllLivesMatter topics. As evidenced both by the network statistics and the network visualizations themselves, the #BlackLivesMatter topic networks show that the conversations are diverse and multifaceted while the #AllLivesMatter networks show conversations that are more limited in scope. Furthermore, #BlackLivesMatter is consistently a more central topic within the #AllLivesMatter networks than #AllLivesMatter is within the #BlackLivesMatter networks. Thus, hijacking is more prevalent within #AllLivesMatter, while #BlackLivesMatter users are able to maintain diverse conversations and delegate hijacking to only a portion of the discourse.

2.4 Slacktivist Reach

We turn now to understanding the relationship between the network of topics and the underlying network of users. Given that we have qualitatively seen that the #BlackLivesMatter topic networks are more diverse than the #AllLivesMatter topic networks (a fact that we further substantiate in Chapter 3), we are especially interested in how the topology of Twitter users may affect the topology of topics. In particular, the extent to which each Lives Matter hashtag consists of "slacktivists" may affect the diversity of topics.

"Slacktivism," short for "slacker activism," is a term describing people who interact with social movements through limited online support, usually in the form of sharing or liking political posts or petitions. Generally, the term has taken on a negative connotation **??**, and there are scholars that have suggested that on-the-ground activism has been replaced by slacktivism, making it more difficult for activists to implement political change **??**. However, while it is unclear what portion of slacktivist efforts reach their goals, there is little evidence to suggest that slacktivism has negatively impacted offline activism **??**. Furthermore, Barbera et al.

demonstrated that the sheer volume of slacktivists help dissemenitate the message of core movement members.

We follow the methodology of Barbera et al. in understanding how periperhy slacktivist users affect the reach and topics of a social movement **??**. We start by constructing a retweet network for each of #BlackLivesMatter and #AllLivesMatter during the month of December 2014, the time period in which both Daniel Pantaleo was not indicted for the death of Eric Garner, and two New York City Police officers were shot. We also construct retweet networks for #PoliceLivesMatter and #BlueLivesMatter which spiked upon the death of the NYPD officers. In these retweet networks, we form a directed edge from one user to another if that user retweeted the second user. This network gives us a proxy for the portrait of conversations and interactions that occured over this time period.

In this retweet network, we measure the *reach* of each user, where a user's reach is the number of retweets they received (their in-degree) over the total number of users in the network. Note, in calculating the reach for each user in the network, we double count some users. While this may seem problematic in terms of measuring each user's individual reach, we take the view that a user often needs to be exposed to a message multiple times before engaging with the message **??**.

With this metric in hand, we perform a k-core decomposition of the retweet network and measure the percentage of remaining reach as a function of the number of cores removed from the network. For this decomposition, we take a user's degree to be the sum of their in-degree and out-degree. Thus, we measure how the reach of each hashtag's message changes as we strip away the network from the periphery to the core. This reach decomposition is depicted in Figure 2.5.

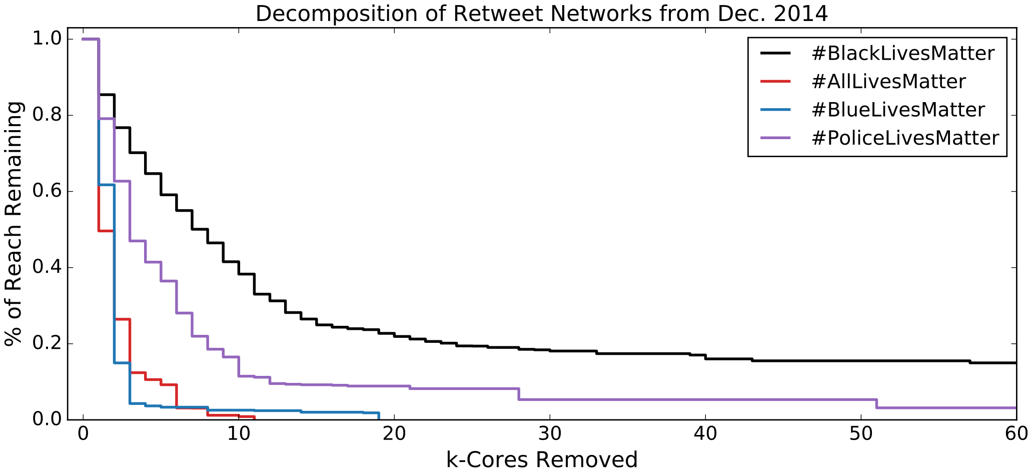


Figure 2.5: Percentage of reach remaining versus the number of cores removed. Both #AllLivesMatter and #BlueLivesMatter drop drastically in terms of reach and expire with less cores removed than #BlackLivesMatter. This suggests that the reach of their messages was primarily driven by one-off, slacktivist users. On the other hand #BlackLivesMatter and even #PoliceLivesMatter exhibit more stable cores that dissemintate mesages.

We see that the reach of both #AllLivesMatter and #BlueLivesMatter drops drastically with the removal of the first few cores, suggesting that the majority of the reach of these hashtags comes from one-off, slacktivist users. The reach of #BlackLivesMatter drops more steadily with the removal of each core, settling in at a dense core that accounts for approximately 30% of the total reach. So, #BlackLivesMatter has proportionally less periphery users than the other Lives Matter hashtags. We see also that #PoliceLivesMatter declines faster than #BlackLivesMatter in terms of reach, but does not extinguish nearly as fast as its two counterpart hashtags. So, #PoliceLivesMatter had more engaged users than both #AllLivesMatter and #PoliceLivesMatter.

We are interested not just in the reach of these hashtags but also the diversity of their topics and conversations and how those relate to any slacktivist topology. To study this, we consider the full hashtag topic network (*not* the backbone of the

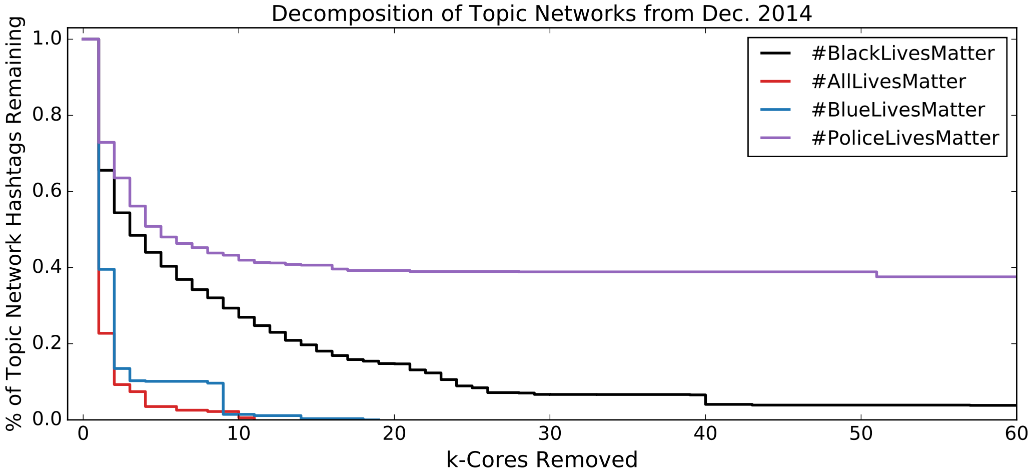


Figure 2.6: Percentage of hashtag topic network nodes remaining versus the number of cores removed.

topic network extracted using the disparity filter). As we decompose the retweet network, we also decompose the hashtag network, removing the nodes of the hashtag network as all of their corresponding users are removed from the retweet network. We measure how the density of hashtags and their connections vary as a function of the number of cores removed in Figures 2.6 and 2.7.

As in the reach networks, the #AllLivesMatter and #BlueLivesMatter hashtag topic networks decompose quickly as the retweet network is decomposed. This fits with our qualitative results about topic diversity from Section 2.3 and matches our later findings in Chapter 3. With #BlackLivesMatter, the hashtag topic network decomposes steadily, honing in on a small subset of topics as we reach the core of the social movement. The #PoliceLivesMatter network collapses onto its core topics more quickly than #BlackLivesMatter. Interestingly, in terms of topics, we see that #BlackLivesMatter lives between two extremes. It does not decompose quickly due to lack of support, but its core topics are supported by a denser subset of users,

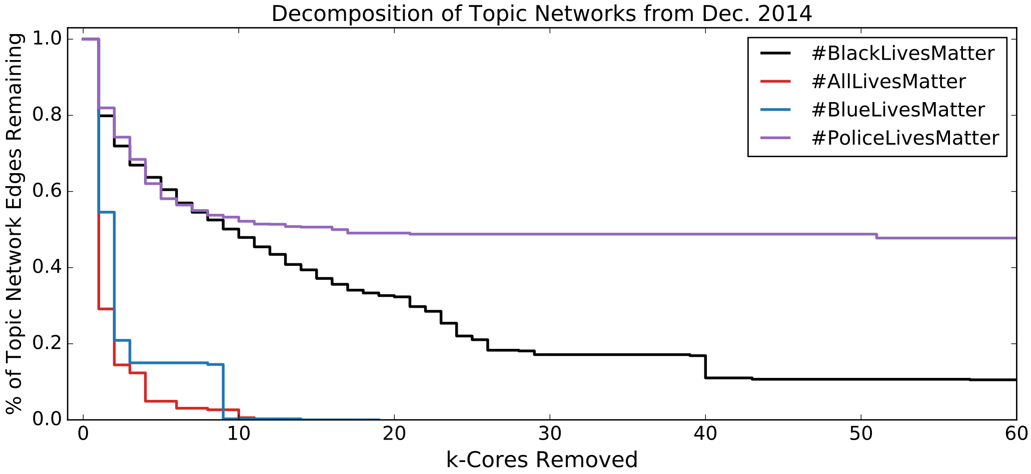


Figure 2.7: Percentage of hashtag topic network edges remaining versus the number of edges removed.

whereas 37% of the #PoliceLivesMatter topics are initiated by a single user.

Chapter 3

Divergent Discourse

Chapter abstract goes here.

3.1 Methods

3.1.1 Entropy and Diversity

Our divergence analysis relies on tools from information theory, so we describe these methods here and frame them in the context of the corpus. We later build upon these tools in Chapter 4 to develop a novel topic model. Given a text with n unique words where the ith word appears with probability p_i, the Shannon entropy H encodes "unpredictability" as

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{3.1}$$

Because Shannon's entropy describes the unpredictability of a body of text, we say that a text with higher Shannon's entropy is less predictable than a text with lower Shannon's entropy. It can then be useful to think of Shannon's entropy as a measure of diversity, where high entropy (unpredictability) implies high diversity. In this case, we refer to Shannon's entropy as the Shannon index.

Of the diversity indices, only the Shannon index gives equal weight to both common and rare events [50]. For this reason, we employ the Shannon index in our study of textual diversity. In addition, even for a fixed diversity index, care must be taken in comparing diversity measures to one another [50]. For example, suppose we have a text with n equally-likely words. The Shannon index of a text with $2n$ equally-likely words is not twice that of the first text, even though we would expect the second text to be twice as diverse. In order to make linear comparisons of diversity between texts, we convert the Shannon index to an effective diversity. The effective diversity D of a text T with respect to the Shannon index is given by

$$D = 2^H = 2^{\left(-\sum_{i=1}^{n} p_i \log_2 p_i\right)}. \tag{3.2}$$

The expression in Eq. 3.2 is also known as the perplexity of the text. The effective diversity gives the number of words D that would be needed to construct a text T' where each word has an equal probability of occurrence, and T and T' have the same entropy, i.e. $H(T) = H(T')$. Unlike the raw Shannon index, the effective diversity doubles in the situation of comparing texts with n and $2n$ equally-likely words, and, in general, allows us to correctly make statements about the ratio of diversity between two texts.

3.1.2 Jensen-Shannon Divergence

The Kullback-Leibler divergence builds upon the notion of entropy to assess the differences between two texts. Given two texts P and Q with a total of n unique words, the Kullback-Leibler divergence between P and Q is defined as

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i}, \tag{3.3}$$

where p_i and q_i are the probabilities of seeing word i in P and Q respectively. However, if there is a single word that appears in one text but not the other, this divergence will be infinitely large. Because such a situation is not unlikely in the context of Twitter, we instead leverage the Jensen-Shannon divergence (JSD) [51], a smoothed version of the Kullback-Leibler divergence:

$$D_{JS}(P||Q) = \pi_1 D_{KL}(P||M) + \pi_2 D_{KL}(Q||M). \tag{3.4}$$

Here, M is the mixed distribution $M = \pi_1 P + \pi_2 Q$ where π_1 and π_2 are weights proportional to the sizes of P and Q such that $\pi_1 + \pi_2 = 1$. The Jensen-Shannon divergence has been previously used in textual analyses that range from the study of language evolution [52, 53] to the clustering of millions of documents [54].

The JSD has the useful property of being bounded between 0 and 1. When entropy is measured in bits (i.e. when logarithm base 2 is used), the JSD is 0 when the texts have exactly the same word distribution, and is 1 when neither text has a single word in common. Furthermore, by the linearity of the JSD we can extract the contribution of an individual word to the overall divergence. The contribution of word i to the JSD is given by

$$D_{JS,i}(P||Q) = -m_i \log_2 m_i + \pi_1 p_i \log_2 p_i + \pi_2 q_i \log_2 q_i, \tag{3.5}$$

where m_i is the probability of seeing word i in M. The contribution from word i is 0 if and only if $p_i = q_i$. Therefore, if the contribution is nonzero, we can label the contribution to the divergence from word i as coming from text P or Q by determining which of p_i or q_i is larger.

3.2 Word-Level Divergence

For each of the eight time periods, the collections of #BlackLivesMatter and #AllLivesMatter tweets are each represented as bags of words where user handles, links, punctuation, stop words, the retweet indicator "RT," and the two hashtags themselves are removed. We then calculate the Jensen-Shannon divergence between these two groups of text, and rank words by percent contribution to the total divergence. We present the results of applying the JSD to each of the weeks of interest in Figures 3.1–3.8.

All contributions on the JSD word shifts are positive, where a bar to the left indicates the word was more common in #AllLivesMatter and a bar to the right indicates the word was more common in #BlackLivesMatter. The bars of the JSD word shift are also shaded according to the diversity of language surrounding each word. For each word w, we consider all tweets containing w in the given hashtag. From these tweets, we calculate the Shannon index of the underlying word distribution with the word w and hashtag removed. A high Shannon index indicates a high diversity of words which, in the context of Twitter, implies that the word w was used in a variety of different tweets. On the other hand, a low Shannon index indicates that the word w originates from a few popular retweets. We emphasize that here we are using the Shannon index not to compare diversities between words, but to simply determine if a word was used diversely or not. By using Figure 3.5 as a baseline (a period where #AllLivesMatter was dominated by one retweet), we determine a rule of thumb that a word is not used diversely if its Shannon index is less than approximately 3 bits.

By inspection of Figure 3.1, we find that #ferguson and #fergusondecision, both hashtags relevant to the non-indictment of Darren Wilson for the death of Michael Brown, contribute to the divergence of #BlackLivesMatter from #AllLivesMatter. Similarly, in Figure 3.6 #freddiegray emerges as a divergent hashtag during the Baltimore protests due to #BlackLivesMatter. In each of these periods, #BlackLivesMatter diverges from #AllLivesMatter by talking proportionally more about the relevant deaths of Black Americans. Similar divergences appear in the other periods as well, as evidenced by the appearance of #ericgarner, #walterscott, and #sandrabland in Figures 3.2–3.5, and 3.8.

During important protest periods, the conversation within #AllLivesMatter diversifies itself around the lives of law enforcement officers. As shown in Figure 3.6, during the Baltimore protests in which #baltimoreuprising and #baltimore were used significantly in #BlackLivesMatter, users of #AllLivesMatter responded with diverse usage of #policelivesmatter and #bluelivesmatter. Similarly, Figure 3.3 shows that upon the death of the two NYPD officers, words such as “officers,” “ramos,” “liu,” and “prayers” appeared in a variety of #AllLivesMatter tweets. In addition, pro-law enforcement hashtags such as #policelivesmatter, #nypd, #nypdlivesmatter, and #bluelivesmatter all contribute to the divergence of #AllLivesMatter from #BlackLivesMatter. Such divergence comes at the same time

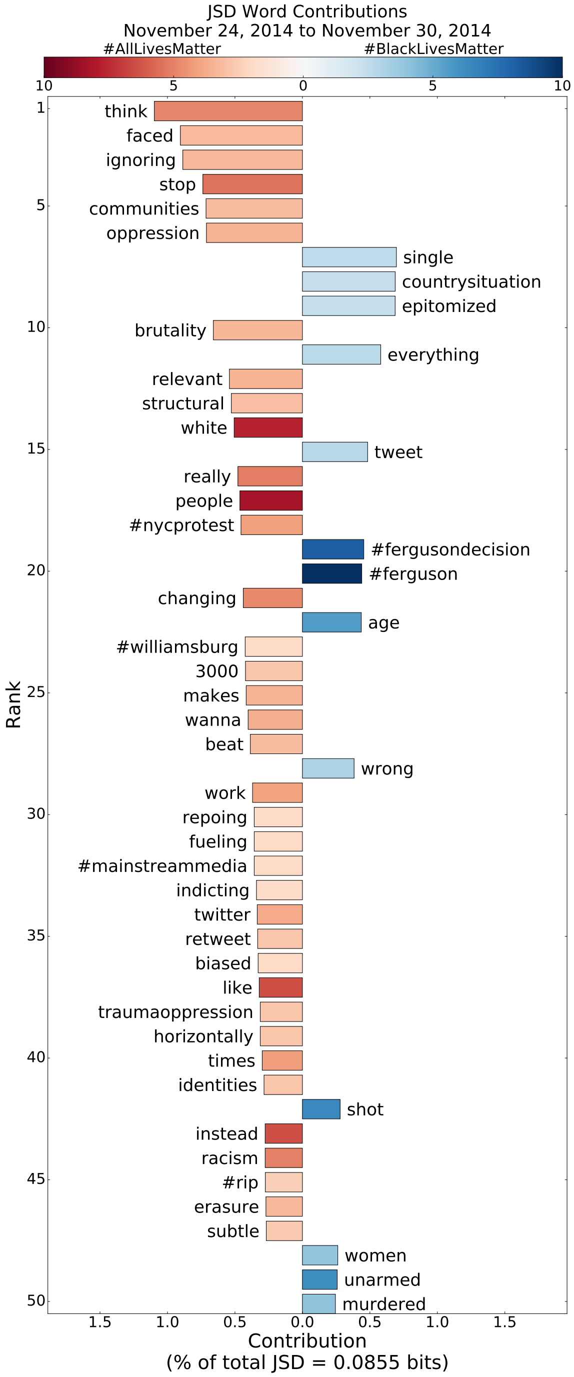


Figure 3.1: Jensen-Shannon divergence word shift for the week following the non-indictment of Darren Wilson in the death of Michael Brown.

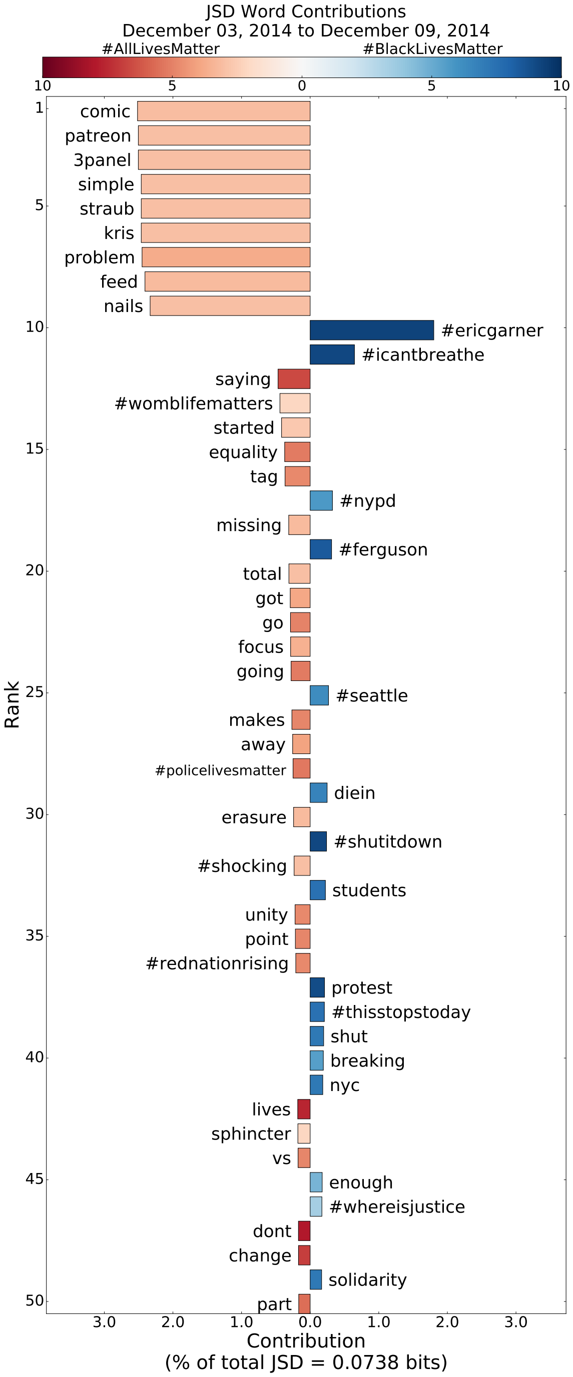
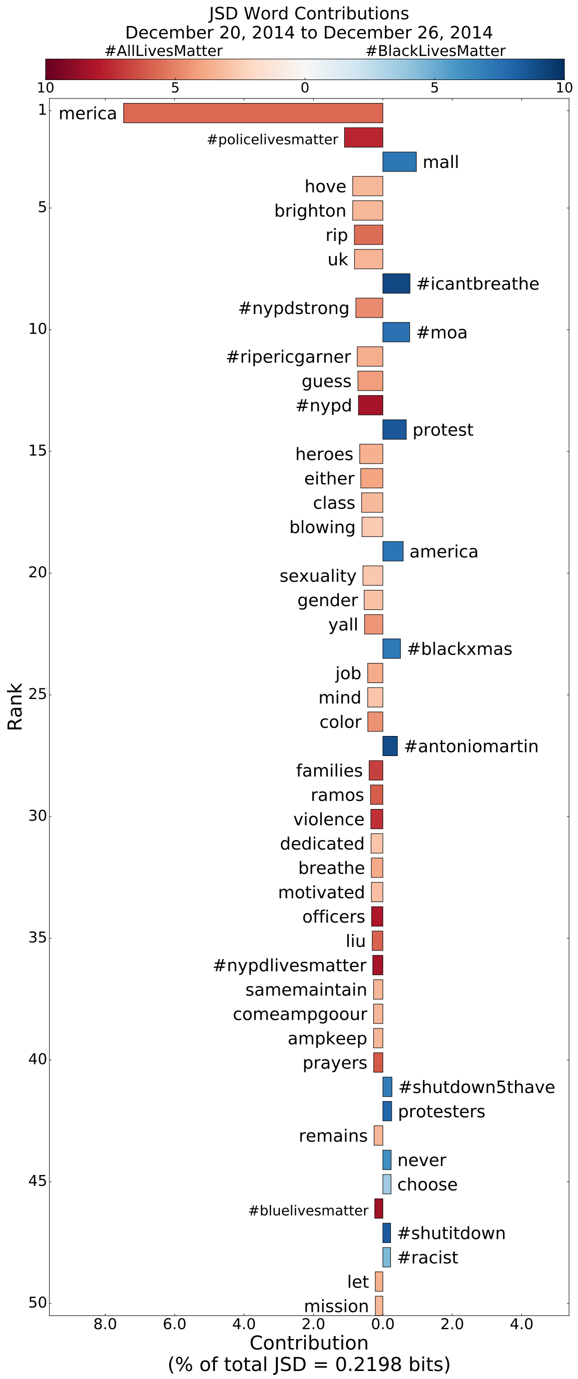


Figure 3.2: Jensen-Shannon divergence word shift for the week following the non-indictment of Daniel Pantaleo in the death of Eric Garner.



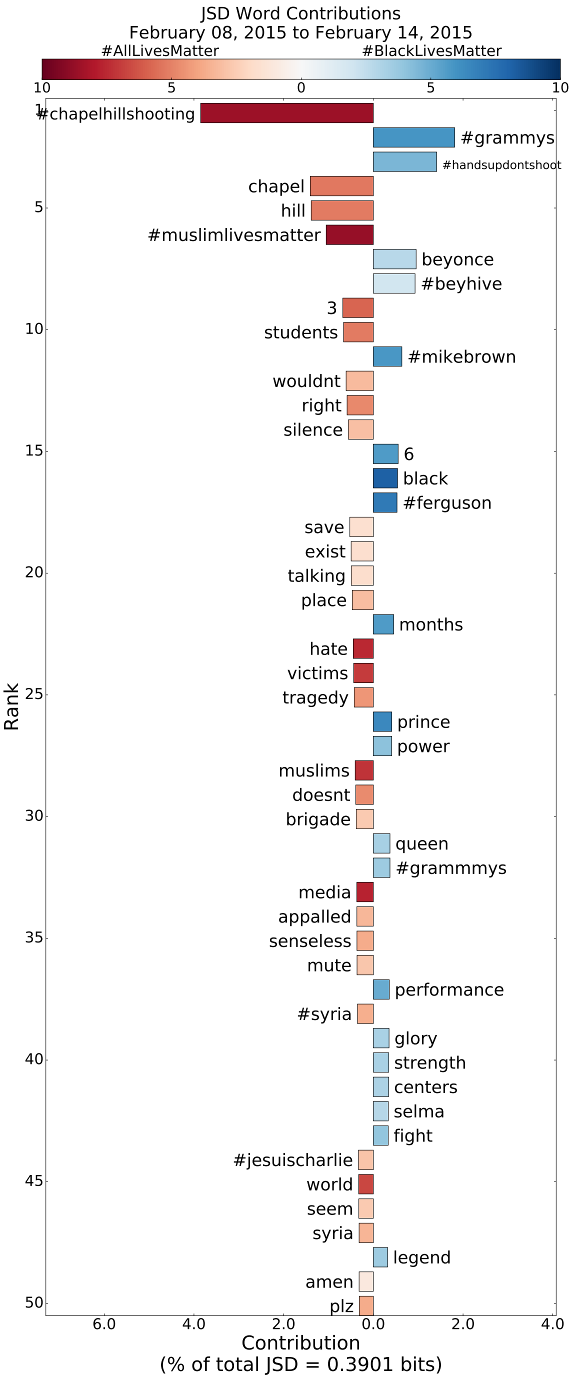


Figure 3.3: Jensen-Shannon divergence word shift for the week following the deaths of New York City police officers Wenjian Liu and Rafael Ramos.

Figure 3.4: Jensen-Shannon divergence word shift for the week following the 2015 Grammy Awards and the Chapel Hill shooting.

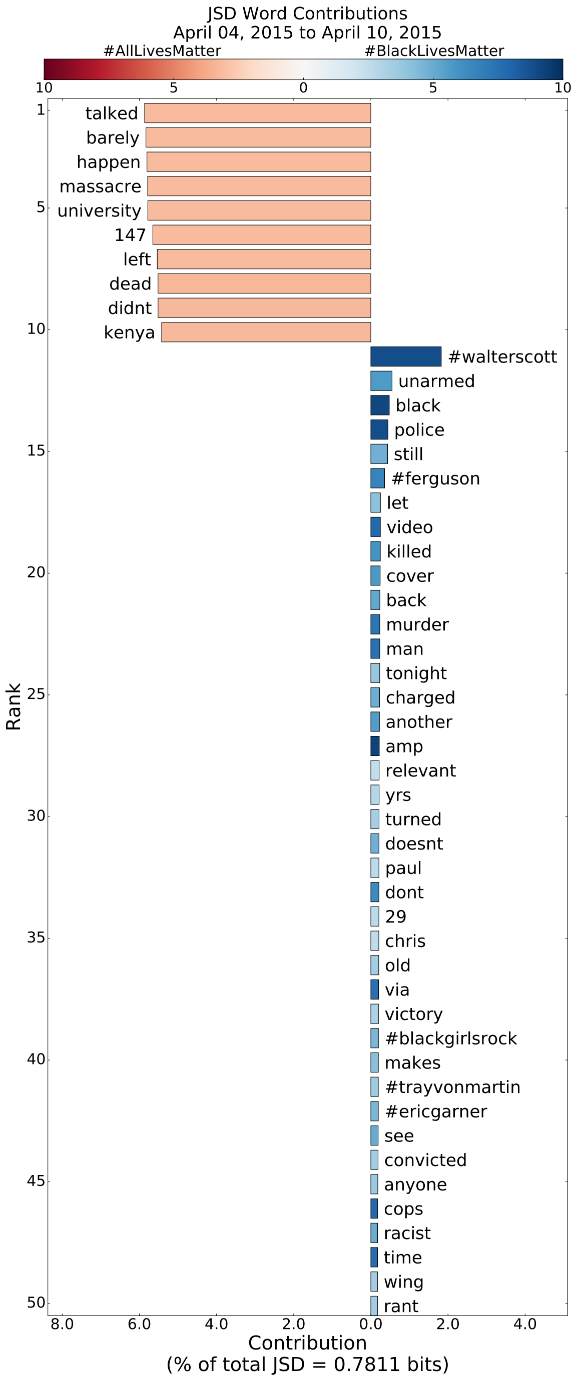


Figure 3.5: Jensen-Shannon divergence word shift for the week following the death of Walter Scott.

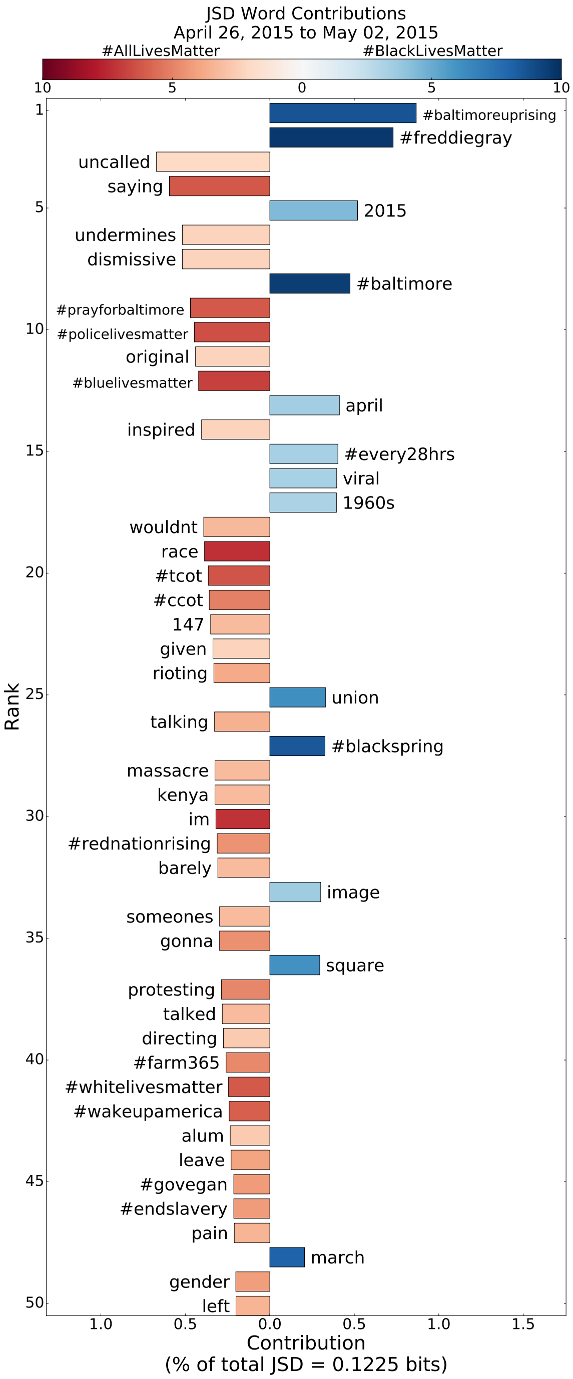


Figure 3.6: Jensen-Shannon divergence word shift for the week encapsulating the peak of the Baltimore protests surrounding the death of Freddie Gray.

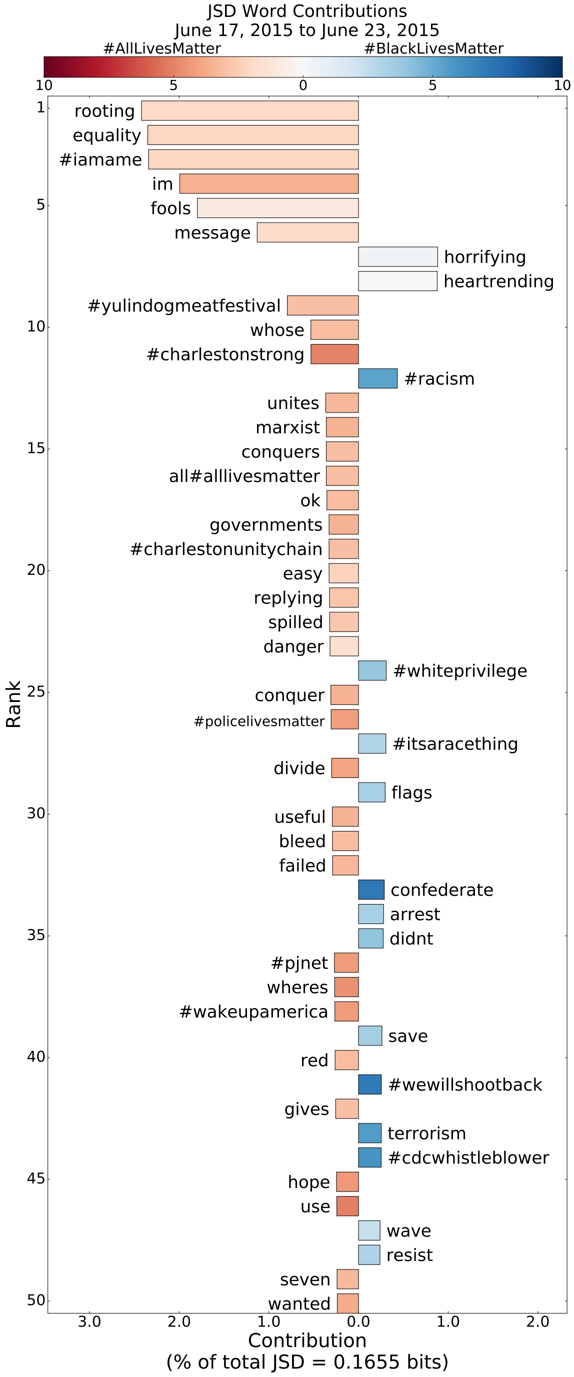


Figure 3.7: Jensen-Shannon divergence word shift for the week following the Charleston Church shooting.

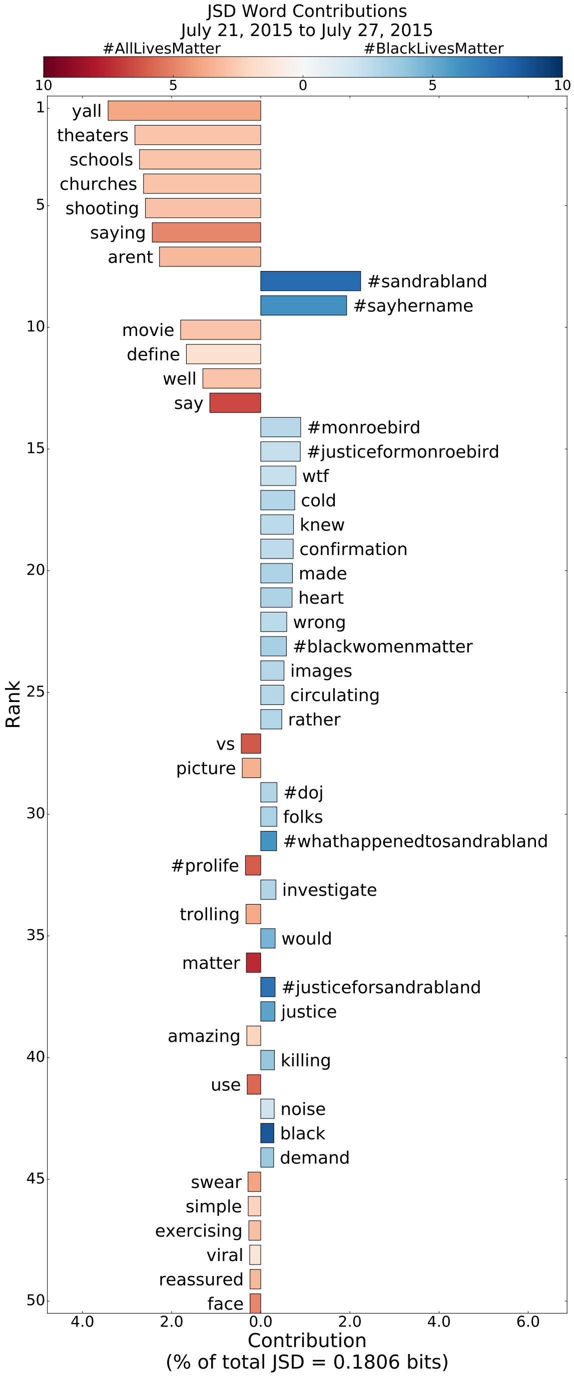


Figure 3.8: Jensen-Shannon divergence word shift for the week encapsulating the outrage over the death of Sandra Bland.

that the hashtags #moa and #blackxmas and words “protest,” “mall,” and “america” were prevalent in #BlackLivesMatter due to Christmas protests, specifically at the Mall of America in Bloomington, Minnesota. So, in the midst of political protests by Black Lives Matter advocates, we see a law-enforcement-aligned response from #AllLivesMatter.

Although the notions of “Black Lives Matter” and “Police Lives Matter” are not necessarily mutually exclusive [30], we see that the conversations within #AllLivesMatter often frame Black protesters versus law enforcement with an“us versus them” mentality. This framing echoes the ways in which media outlets have historically framed the tension between Black protesters and law enforcement [55–57], where police officers and protesters are seen as “enemy combatants” [29] and such movements appear to “jeopardize law enforcement lives” [30]. So, by facilitating this opposition, #AllLivesMatter becomes the center of upholding historically contentious views in the midst of what some consider the New Civil Rights Movement.

During the period following the non-indictment of Darren Wilson, there are some words, such as “oppression,” “structural,” and “brutality,” that seem to suggest engagement from #AllLivesMatter with the issues being discussed within #BlackLivesMatter, such as structural racism and police brutality. Since the diversities of these words are low, we can inspect popular retweets containing these words to understand how they were used. Doing so, we find that the words actually emerge in #AllLivesMatter due to hijacking [37], the adoption of a hashtag to mock or criticize it. That is, these words appear not because of discussion of structural oppression and police brutality by #AllLivesMatter advocates, but because

#BlackLivesMatter supporters are specifically critiquing the fact such discussions are not occurring within #AllLivesMatter. (We have chosen to not provide direct references to these tweets so as to protect the identity of the original tweeter.) Similarly, a “3-panel comic” strip criticizing the notion of “All Lives Matter” circulated through #AllLivesMatter following the death of Eric Garner (Figure 3.2), and after the Chapel Hill and Charleston Church shootings, #BlackLivesMatter proponents leveraged #AllLivesMatter to question why believers of the phrase were not more vocal (Figures 3.4 and 3.7). We note that we are able to pick up on these instances of hijacking by inspecting words with high divergence, but low diversity (meaning the divergence comes almost entirely from the few retweets containing the word). This hijacking drives a significant portion of the divergence of #AllLivesMatter from #BlackLivesMatter in many of these periods.

As shown with the topic networks, we also uncovered hijacking of #BlackLivesMatter by #AllLivesMatter advocates. Such hijacking of #BlackLivesMatter is similar to the content injection described by Conover et al. [31], where one group adopts the hashtag of politically opposed group in order to inject their ideological beliefs. Content injection of this type has also between found in the work of Egyptian political polarization [35]. However, a significant portion #AllLivesMatter hijacking by #BlackLivesMatter supporters is not simple content injection. Rather, advocates of #BlackLivesMatter often use #AllLivesMatter to directly interrogate the stance of “All Lives Matter” and the worldview implied by that phrase. Furthermore, as seen by the topic networks and word shifts, such discussions have largely been regulated to #AllLivesMatter, allowing #BlackLivesMatter to exhibit diverse conversations about a variety of topics.

Although past research has expressed concern that #AllLivesMatter would derail from the movement started by #BlackLivesMatter [27, 29, 30, 58], our data-driven approach has allowed us to uncover that #BlackLivesMatter has countered #AllLivesMatter content injection. Our findings suggest that a protest movement can maintain its conversational momentum by forcing opposing opinions to be a central part of a counter-protest's discussions, rather than its own.

Finally, although we have found that the divergences between #BlackLivesMatter and #AllLivesMatter result partially from proportionally higher discussion of Black deaths in #BlackLivesMatter, it is important to note that #AllLivesMatter is not completely devoid of discussion about these deaths. For instance, #ripericgarner is prominent within #AllLivesMatter following the death of Eric Garner, #iamame ("I am African Methodist Episcopal") contributes more to #AllLivesMatter following the Charleston Church shooting, and the names of several Black Americans appear in the #AllLivesMatter topic networks. However, many of these signs of solidarity are associated with low diversity. In light of this, it is also important to note that there is a lack of discussion of other deaths within #AllLivesMatter. That is, in examining several of the main periods where #AllLivesMatter spikes, only the Chapel Hill shooting period shows discussion of non-Black deaths.

3.3 Conversational Diversity

Having quantified both the word-level divergences and the large-scale topic networks, we now measure the informational diversity of #BlackLivesMatter and

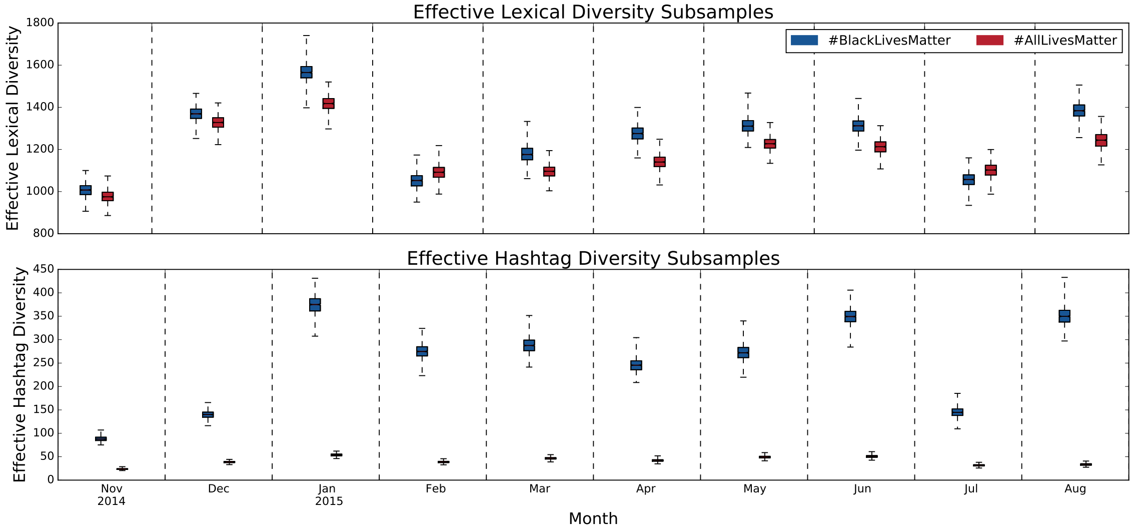


Figure 3.9: To control for how volume affects the effective diversity of #BlackLivesMatter and #AllLivesMatter, we break the time scale down into months and subsample 2000 tweets from each hashtag 1000 times. The plot shows notched box plots depicting the distributions of these subsamples for effective lexical and hashtag diversity. The notches are small on all the boxes, indicating that the mean diversities are significantly different at the 95% confidence level across all time periods.

#AllLivesMatter more precisely. We do this through two approaches. First, we measure "lexical diversity," the diversity of words other than hashtags. Second, we measure the hashtag diversity. We measure these diversities using the effective diversity described in Eqn. 3.2 in Section 3.1.1. Furthermore, to account for the different volume of #BlackLivesMatter and #AllLivesMatter tweets, we break the time scale down into months and subsample 2000 tweets from each hashtag 1000 times, calculating the effective diversities each time. The results are shown in Figure 3.9.

The lexical diversity of #BlackLivesMatter is larger than #AllLivesMatter in eight of the ten months with an average lexical diversity that is 5% more than that of #AllLivesMatter. Interestingly, the two cases where #AllLivesMatter has higher lexical diversity are in the two periods when there were large non-police-involved shootings of people of color and #AllLivesMatter was used as a hashtag of solidarity. However, the more striking differences are in terms of hashtag diversity. On average, the hashtag diversity of #BlackLivesMatter is six times that of #AllLivesMatter. This is in line with our network analysis where we found expansive #BlackLivesMatter topic networks and tightly clustered, less diverse #AllLivesMatter topic networks.

The low hashtag diversity of #AllLivesMatter is relatively constant. One could imagine that the lack of diversity in the topics of #AllLivesMatter is a result of a focused conversation that does not deviate from its main message. However, as we have demonstrated through the JSD word shifts and topic networks, the conversation of #AllLivesMatter does change and evolve with respect to the different time periods. We see mentions of the major deaths and events of these periods within #AllLivesMatter, even if they do not rank as highly in terms of topic centrality. So, even though both protest hashtags have overlap on many of the major topics, the diversity of topics found within #BlackLivesMatter far exceeds that of #AllLivesMatter, even when accounting for volume.

Chapter 4

Topic Modeling with Minimal Domain Knowledge

Chapter abstract goes here. Need to mention LDA (maybe just use a variation of paper abstract)

4.1 Methods

4.1.1 Correlation Explanation (CorEx)

Correlation Explanatino (CorEx) is an information-theoretic approach to topic modeling, bypassing the traditional generative model assumed by Latent Dirichlet Allocation. Here, we largely adopt the notation used by Ver Steeg and Galstyan in their original presentation of the model [59]. Let X be a discrete random variable that takes on a finite number of values. Furthermore, if we have n such random variables, let X_G denote a subcollection of them, where $G \subseteq \{1, \dots, n\}$. The entropy of X is written as $H(X)$ and the mutual information of two random variables X_1 and X_2 is given by $I(X_1 : X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$.

The total correlation, or multivariate mutual information, of a group of random variables X_G is expressed as

$$TC(X_G) = \sum_{i \in G} H(X_i) - H(X_G) \tag{4.1}$$

$$= D_{KL}\left(p(X_G) || \prod_{i \in G} p(X_i)\right). \tag{4.2}$$

We see that Eqn. 4.1 does not quantify “correlation” in the modern sense of the word, and so it can be helpful to conceptualize total correlation as a measure of total dependence. Indeed, Eqn. 4.2 shows that total correlation can be expressed using the Kullback-Leibler Divergence and, therefore, it is zero if and only if the joint distribution of X_G factorizes, or, in other words, there is no dependence between the random variables.

The total correlation can be written when conditioning on another random variable Y, $TC(X_G \mid Y) = \sum_{i \in G} H(X_i \mid Y) - H(X_G \mid Y)$. So, we can consider the reduction in the total correlation when conditioning on Y.

$$TC(X_G; Y) = TC(X_G) - TC(X_G \mid Y) \tag{4.3}$$

$$= \sum_{i \in G} I(X_i : Y) - I(X_G : Y) \tag{4.4}$$

This measures how much Y explains the dependencies in X_G. The quantity expressed in Eqn. 4.3 acts as a lower bound of $TC(X_G)$ [60], as readily verified by noting that $TC(X_G)$ and $TC(X_G|Y)$ are always non-negative. Also note, the joint distribution of X_G factorizes conditional on Y if and only if $T(X_G \mid Y) = 0$. If this is the case, then $TC(X_G; Y)$ is maximized.

In the context of topic modeling, X_G represents a group of words and Y

represents a topic. Since we are always interested in grouping multiple sets of words into multiple topics, we will denote the latent topics as $Y_1, \ldots Y_m$ and their corresponding groups of words as X_{G_j} for $j = 1, \ldots, m$ respectively. The CorEx topic model seeks to maximally explain the dependencies of words in documents through latent topics by maximizing $TC(X; Y_1, \ldots, Y_m)$. Instead, we maximize the following lower bound on this expression:

$$\max_{G_j, p(y_j|x_{G_j})} \sum_{j=1}^{m} TC(X_{G_j}; Y_j). \tag{4.5}$$

This optimization is subject to the constraint that the groups, G_j, do not overlap and the conditional distribution is normalized. The solution to this objective can be efficiently approximated, despite the search occurring over an exponentially large probability space [59].

The latent factors, Y_j, are optimized to be informative about dependencies in the data and do not require generative modeling assumptions. Note that the discovered factors, Y, can be used as inputs to construct new latent factors, Z, and so on leading to a hierarchy of topics. Although this extension is quite natural, we focus our analysis on the first level of topic representations for easier interpretation and evaluation.

4.1.2 Anchoring and the Information Bottleneck

The information bottleneck formulates a trade-off between compressing data X into a representation Y, and preserving the information in X that is relevant to Z (typically labels in a supervised learning task) [61, 62]. More formally, the information bottleneck is expressed as

$$\max_{p(y|x)} \beta I(Z:Y) - I(X:Y), \tag{4.6}$$

where β is a parameter controlling the trade-off between compressing X and preserving information about Z.

To see the connection with CorEx, we rewrite the objective of Eqn. 4.5 using Eqn. 4.4 as follows,

$$\max_{G_j, p(y_j|x_{G_j})} \sum_{j=1}^{m} \sum_{i \in G_j} I(X_i : Y_j) - \sum_{j=1}^{m} I(X_{G_j} : Y_j). \tag{4.7}$$

by following the derivation of Ver Steeg and Galstyan [59] and introducing indicator variables $\alpha_{i,j}$ which are equal to 1 if and only if word X_i appears in topic Y_j (i.e. $i \in G_j$).

$$\max_{\alpha_{i,j}, p(y_j|x)} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{i,j} I(X_i : Y_j) - I(X : Y_j) \right) \tag{4.8}$$

Note that the constraint on non-overlapping groups now becomes a constraint on α. Comparing the objective to Eqn. 4.6, we see that we have exactly the same compression term for each latent factor, $I(X : Y_j)$, but the relevance variables now correspond to $Z \equiv X_i$. Inspired by the success of the bottleneck, we suggest that if we want to learn representations that are more relevant to specific keywords, we can simply anchor a word X_i to topic Y_j, by constraining our optimization so that $\alpha_{i,j} = \beta_{i,j}$, where $\beta_{i,j} \geq 1$ controls the anchor strength. Otherwise, the updates on α remain the same as in Ver Steeg and Galstyan's original presentation [59]. This schema is a natural extension of the CorEx objective and it is flexible, allowing for multiple words to be anchored to one topic, for one word to be anchored to multiple topics, or for any combination of these anchoring strategies. Furthermore, it combines supervised and unsupervised learning by allowing us to leave some topics without anchors.

4.1.3 Related Work

With respect to integrating domain knowledge into topic models, we have drawn inspiration from Arora et al., who used anchor words in the context of non-negative matrix factorization [63]. Using an assumption of separability, these anchor words act as high precision markers of particular topics and, thus, help discern the topics from one another. Although the original algorithm proposed by Arora et. al and subsequent improvements to the algorithm find these anchor words automatically [64, 65], recent adaptations allow manual insertion of anchor words and other metadata [66, 67]. Our work is similar to the latter, where we treat anchor words as fuzzy logic markers and embed them into the topic model in a semi-supervised fashion. In this sense, our work is closest to Halpern et al., who have also made use of domain expertise and semi-supervised anchored words in devising topic models [68, 69].

There is an adjacent line of work that has focused on incorporating word-level information into LDA-based models. Andrezejewski and Zhu have presented two flavors of such models. One allows specification of Must-Link and Cannot-Link relationships between words that help partition otherwise muddled topics [70]. The other model makes use of "z-labels," words that are known to pertain to a specific topics and that are restricted to appearing in some subset of all the possible topics [71]. Similarly, Jagarlamudi et. al proposed SeededLDA, a model that seeds words into given topics and guides, but does not force, these topics towards these integrated words [72]. While we also seek to guide our model towards topics containing user-provided words, our model naturally extends to incorporating such information, while the LDA-based models require involved and careful construction of new assumptions. Thus, our framework is more lightweight and flexible than LDA-based models.

Mathematically, CorEx topic models most closely resemble topic models based on latent tree reconstruction [73]. In Chen et. al.'s analysis, their own latent tree approach and CorEx both report significantly better perplexity than hierarchical topic models based on the hierarchical Dirichlet process and the Chinese restaurant process. CorEx has also been investigated as a way to find "surprising" documents [74].

4.2 Data and Evaluation

Our first data set consists of 504,000 humanitarian assistance and disaster relief (HA/DR) articles collected from ReliefWeb, an HA/DR news article aggregator sponsored by the United Nations. Of these articles, about 111,000 of them are in English and contain a label indicating at least one of 21 disaster types, such as Flood, Earthquake, or Wild Fire. To mitigate overwhelming label imbalances, we both restrict the documents to those with one label, and randomly subsample 2000 articles from each of the largest disaster type labels. This leaves us with a corpus of 18,943 articles.

These articles are accompanied by an HA/DR lexicon of approximately 34,000 words and phrases. The lexicon was curated by first gathering seed terms from HA/DR domain experts and CrisisLex, resulting in approximately 40-60 terms per disaster type. This term list was then expanded through the use of several word2vec models per each set of seeds words, and then filtered by removing names, places, non-ASCII characters, terms with fewer than three characters, and words deemed too “semantically distant” from the seeds words by the word2vec models. Finally, the extracted terms were audited using CrowdFlower, where users rated the relevance of the terms on a Likert scale. Low relevance terms were dropped from the lexicon. Of these terms 11,891 appear in the HA/DR articles.

Our second set of data consists of deidentified clinical discharge summaries from the Informatics for Integrating Biology and the Bedside (i2b2) 2008 Obesity Challenge. These summaries are labeled by clinical experts with conditions frequently associated with obesity, such as Coronary Artery Disease, Depression, and Obstructive Sleep Apnea. For these documents, we leverage a text pipeline that extracts common medical terms and phrases [75, 76]. There are 4,114 such terms that appear in the i2b2 clinical health notes. For both sets of data, we use their respective lexicons to parse the documents.

It is well-known that traditional methods for evaluating topic models, such as perplexity and held-out log-likelihood do not necessarily correlate with human evaluation of semantic topic quality [77]. Therefore, we measure the semantic quality of the topic models using Mimno et. al’s UMass automatic topic coherence score [78]. This measure has been shown to correlate well with human evaluation of topic coherence. Suppose there are n topics, and that the k most probable words of topic t are given by the list $(w_1^t, \ldots, w_k^t)$. Then the coherence of topic t is given by

$$\sum_{i=2}^{k} \sum_{j=1}^{i} \log \frac{D(w_i^t, w_j^t) + 1}{D(w_j^t)} \tag{4.9}$$

where $D(w_i^t)$ is the number of documents in which word w_i appears, and $D(w_i^t, w_j^t)$ is the number of documents in which w_i and w_j appear together.

Second, in the case of the disaster relief documents, we make use of the HA/DR lexicon word labels to report the purity of the topic word lists, the highest fractional count of the word labels. For example, given a topic list with k words, the purity of a list with words all of the same label is 1, while that of a list with words all different labels is $1/k$. Since the HA/DR lexicon labels are the result of expert knowledge and crowd-sourcing, the purity provides us with a measure of semantic topic consistency similar to word intrusion tests [77, 79].

Finally, we evaluate the models in terms of document classification, where the feature set of each document is its topic distribution. The classification is carried out using multiclass logistic regression as implemented by the Scikit-Learn library [80], where one binary regression is trained for each label and the label with the highest probability of appearing is selected. While more sophisticated machine learning algorithms may produce better predictive scores, their complex frameworks have the potential to obfuscate differences between topic models. We also leverage the interpretability of logistic regression in our analysis of anchored CorEx. We perform all document classification tasks using a 60/40 split for training and testing.

4.3 Comparison to Latent Dirichlet Allocation

CorEx takes binarized documents as input for its topic model, so we compare it to LDA giving LDA two different inputs: binarized document-word counts and standard document-word counts. In doing these comparisons, we use the Gensim implementation of LDA [81]. The results of comparing CorEx to LDA as a function of the number of topics are presented in Figure 4.1.

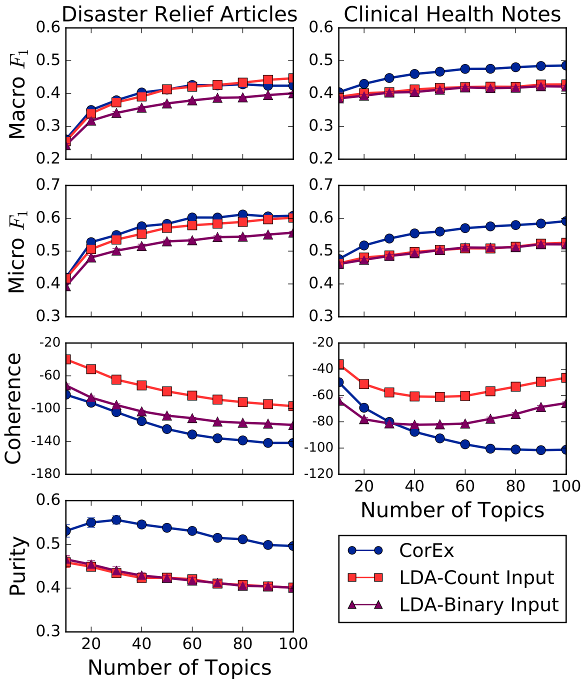


Figure 4.1: Baseline comparison of CorEx to LDA with respect to document classification and topic quality on disaster relief articles and clinical health notes as the number of topics vary. Points are the average of 30 runs of a topic model. Confidence intervals are plotted but are so small that they are not distinguishable. CorEx uses binarized documents, so we compare CorEx to LDA with binarized input and standard count input.

On the disaster relief articles, we see that CorEx is competitive with LDA in terms of document classification, and even outperforms LDA in terms of document classification on the clinical health notes. This is despite the fact that CorEx leverages only binary word counts, and LDA uses regular count data. So, with less information than LDA, CorEx produces topics that are as good as or better than the topics produced by LDA when used for document classification.

Inspecting the last two rows of Figure 4.1, we find that LDA performs better than CorEx in terms of topic coherence, while CorEx performs better than LDA in terms of topic purity. While this appears to yield seemingly conflicting information about the semantic quality of these topic models, it is important to acknowledge that the UMass topic coherence measures assumes that the topic words are the most probable words per each topic. CorEx does not output the most probable words, but rather the words of highest mutual information with the topic. This provides a possible explanation for why CorEx does not perform as well as LDA in terms of coherence, but significantly outperforms in terms of purity. Although topic coherence correlates well with human evaluation of semantic quality, it appears important to apply the measure only *within* models and not *across* models if the topic words are ordered according to different criteria.

4.4 Effect of Anchor Words

In analyzing anchored CorEx, we wish to systematically test the effect of anchor words given the domain-specific lexicons. To do so, we follow the approach used by Jagarlamudi et. al: for each label in a data set, we find the words that have the highest mutual information, or information gain, with the label [72]. For word w and label L, this is computed as

$$I(L : w) = H(L) - H(L \mid w), \tag{4.10}$$

where for each document of label L we consider if the word w appears or not.

To discern the effects of anchoring words to CorEx and simulate domain knowledge injection, we devise the following experiment: first, we determine the top five anchor words for each document label using the methodology described in Section 4.3. Second, for each document label, we run an anchored CorEx topic model with that label's anchor words anchored to exactly one topic. We compare this anchored topic model to an unsupervised CorEx topic model using the same random seeds, thus creating a matched pair where the only difference is the treatment of anchor words. Finally, this matched pairs process is repeated 30 times, yielding a distribution for each metric over each label.

We use 50 topics when modeling the ReliefWeb articles and 30 topics when modeling the i2b2 clinical health notes. These values were chosen by observing diminishing returns to the total correlation explained by additional topics. In Figure 4.2 we show how the results of this experiment vary as a function of the anchoring parameter β for each disaster and disease type in the two data sets. We examine a more detailed cross section of these results in Fig 4.3, where we set $\beta = 5$ for the clinical health notes and set $\beta = 10$ for the disaster relief articles.

A priori we do not know that anchoring will cause the anchor words to appear at the top of topics. So, we first measure how the topic overlap, the proportion of the top ten mutual information words that appear within the top ten words of the

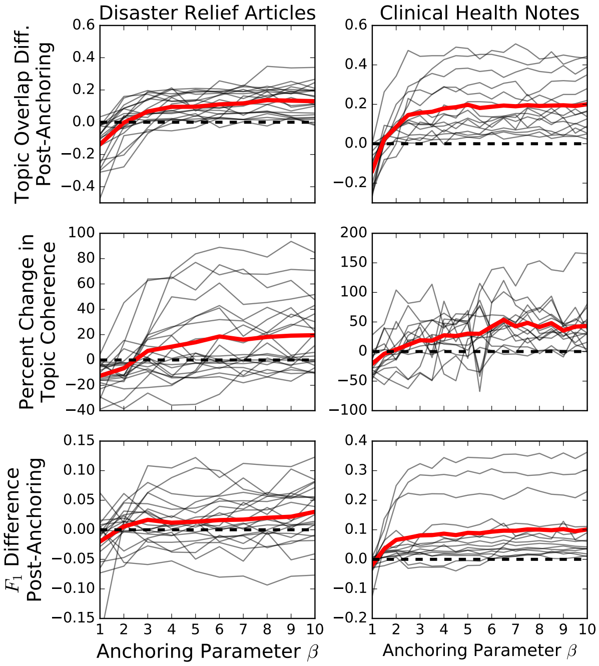


Figure 4.2: Effect of anchoring words to a single topic for one document label at a time as a function of the anchoring parameter β. Light gray lines indicate the trajectory of the metric for a given disaster or disease label. Thick red lines indicate the pointwise average across all labels for fixed value of β.

topics, changes before and after anchoring. From Figure 4.2 we see that as β increases, more of these relevant words consistently appear within the topics. For the disaster relief articles, many disaster types see about two more words introduced, while in the clinical health notes the overlap increases by up to four words. Analyzing the cross section in Figure 4.3, we see many of these gains come from disaster and disease types that appeared less in the topics pre-anchoring. Thus, we can sway the topic model towards less dominant themes through anchoring. Document labels that were already well represented are those where the topic overlap changes the least.

Next, we examine whether these anchored topics are more coherent topics. To

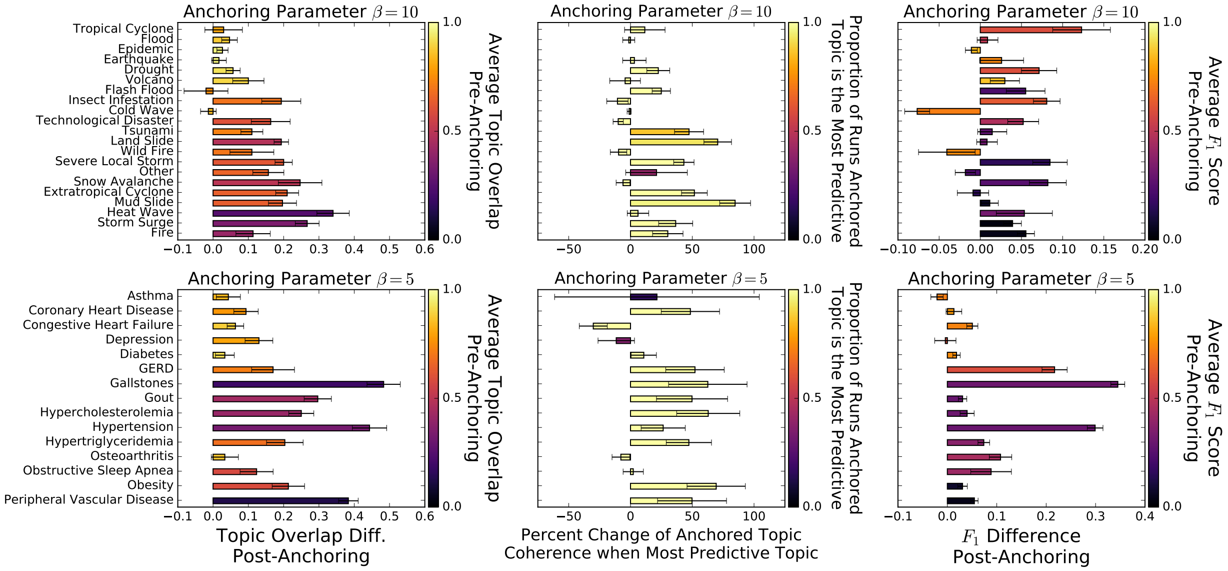


Figure 4.3: Cross-section results of the anchoring metrics from fixing $\beta = 5$ for the clinical health notes, and $\beta = 10$ for the disaster relief articles. Disaster and disease types are sorted by frequency, with the most frequent document labels appearing at the top. Error bars indicate 95% confidence intervals. The color bars provide baselines for each metric: topic overlap pre-anchoring, proportion of topic model runs where the anchored topic was the most predictive topic, and F_1 score pre-anchoring.

do so, we compare the coherence of the anchored topic with that of the most predictive topic pre-anchoring, the topic with the largest corresponding coefficient in magnitude of the logistic regression, when the anchored topic itself is most predictive. From Figure 4.2, we see these results have more variance, but largely the anchored topics are more coherent. In some cases, the coherence is 1.5 to 2 times that of pre-anchoring. Furthermore, by Figure 4.3, we find that the anchored topics are, indeed, often the most predictive topics for each document label. Similar to topic overlap, the labels that see the least improvement are those that appear the most and are already well-represented in the topic model.

Finally, we find that the anchored, more coherent topics can lead to modest gains in document classification. For the disaster relief articles, Figure 4.2 shows that there are mixed results in terms of F_1 score improvement, with some disaster types performing consistently better, and others performing consistently worse. The results are more consistent for the clinical health notes, where there is an average increase of about 0.1 in the F_1 score, and some disease types see an increase of up to 0.3 in F_1. Given that we are only anchoring 5 words to the topic model, these are significant gains in predictive power.

Unlike the gains in topic overlap and coherence, the F_1 score increases do not simply correlate with which document labels appeared most frequently. For example, we see in Figure 4.3 that Tropical Cyclone exhibits the largest increase in predictive performance, even though it is also one of the most frequently appearing document labels. Similarly, some of the major gains in F_1 for the disease types, and major losses in F_1 for the disaster types, do not come from the most or least frequent document labels. Thus, if using anchored CorEx for document classification, it is important to examine how the anchoring affects prediction for individual document labels.

We hypothesize that the results of topic overlap, topic coherence, and F_1 score are more muted and have higher variance on the disaster relief articles because there is higher lexical overlap between disaster types than the disease types in the clinical health notes. For example, documents discussing Floods and Flash Foods share many common themes, as do documents discussing Landslides and Mudslides. So again, we emphasize that in applying anchored CorEx, the user should pay attention to how the topics change with the introduction of anchoring, and that the user should experiment with different values of the anchoring parameter β to see how these topics are affected.

Bibliography

[1] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

[2] Michael D Conover, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. The digital evolution of Occupy Wall Street. *PloS ONE*, 8(5):e64679, 2013.

[3] Neal Caren and Sarah Gaby. Occupy online: Facebook and the spread of Occupy Wall Street. *Available at SSRN 1943168*, 2011.

[4] Kevin M DeLuca, Sean Lawson, and Ye Sun. Occupy Wall Street on the public screens of social media: The many framings of the birth of a protest movement. *Communication, Culture & Critique*, 5(4):483–509, 2012.

[5] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Mazaid. Opening closed regimes: What was the role of social media during the Arab Spring? *Available at SSRN 2595096*, 2011.

[6] Zizi Papacharissi and Maria de Fatima Oliveira. Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication*, 62(2):266–282, 2012.

[7] Eva Anduiza, Camilo Cristancho, and José M Sabucedo. Mobilization through online social networks: The political protest of the Indignados in Spain. *Information, Communication & Society*, 17(6):750–764, 2014.

[8] Sandra González-Bailón and Ning Wang. Networked discontent: The anatomy of protest campaigns in social media. *Social networks*, 44:95–104, 2016.

[9] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movement communication network. *PLoS ONE*, 8(3):e55957, 2013.

[10] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1(197), 2011.

[11] 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.

[12] Jennifer Larson, Jonathan Nagler, Jonathan Ronen, and Joshua Tucker. Social networks and protest participation: Evidence from 93 million twitter users. In *Political Networks Workshops & Conference*, 2016.

[13] Henrik Serup Christensen. Political activities on the Internet: Slacktivism or political participation by other means? *First Monday*, 16(2), 2011.

[14] Pablo Barberá, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. The critical periphery in the growth of social protests. *PLoS ONE*, 10(11):e0143611, 2015.

[15] Javier Borge-Holthoefer, Nicola Perra, Bruno Gonçalves, Sandra González-Bailón, Alex Arenas, Yamir Moreno, and Alessandro Vespignani. The dynamics of information-driven coordination phenomena: A transfer entropy analysis. *Science Advances*, 2(4), 2016.

[16] Hong Qi, Pedro Manrique, Daniela Johnson, Elvira Restrepo, and Neil F Johnson. Open source data reveals connection between online and on-street protest activity. *EPJ Data Science*, 5(1):1, 2016.

[17] Alicia Garza. A herstory of the #blacklivesmatter movement. *The Feminist Wire* (2014). http://www.thefeministwire.com/2014/10/blacklivesmatter-2/. Accessed 20 June 2016.

[18] Fredrick C Harris. The next civil rights movement? *Dissent*, 62(3):34–40, 2015.

[19] Jay Caspian Kang. 'our demand is simple: Stop killing us': How a group of Black social media activists built the nation's first 21st-century civil rights movement. *New York Times* (2015). http://www.nytimes.com/2015/05/10/magazine/our-demand-is-simple-stop-killing-us.html. Accessed 20 June 2016.

[20] Opal Tometi and Gerald Lenoir. Black Lives Matter is not a civil rights movement. *Time* (2015). http://time.com/4144655/international-human-rights-day-black-lives-matter/. Accessed 20 June 2016.

[21] Deen Goodwin Freelon, Charlton D McIlwain, and Meredith D Clark. Beyond the hashtags: #ferguson, #blacklivesmatter, and the online struggle for offline justice. *Available at SSRN*, 2016.

[22] Sarah J Jackson and Brooke Foucault Welles. #ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication & Society*, pages 1–22, 2015.

[23] Yarimar Bonilla and Jonathan Rosa. #ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the united states. *American Ethnologist*, 42(1):4–17, 2015.

[24] Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. Characterizing the demographics behind the #blacklivesmatter movement. *arXiv preprint arXiv:1512.05671*, 2015.

[25] Deen Freelon, Charlton McIlwain, and Meredith Clark. Quantifying the power and consequences of social media protest. *New Media & Society*, page 1461444816676646, 2016.

[26] Marlon Twyman, Brian C Keegan, and Aaron Shaw. Black lives matter in Wikipedia: Collaboration and collective memory around online social movements. *arXiv preprint arXiv:1611.01257*, 2016.

[27] Mark Orbe. #alllivesmatter as post-racial rhetorical strategy. *Journal of Contemporary Rhetoric*, 5(3/4):90–98, 2015.

[28] Terry Husband Jr. ”i don’t see color”: Challenging assumptions about discussing race with young children. *Early Childhood Education Journal*, 39(6):365–371, 2012.

[29] Russell Rickford. Black lives matter toward a modern practice of mass struggle. In *New Labor Forum*, volume 25, pages 34–42. SAGE Publications, 2016.

[30] Catherine L. Langford and Montené Speight. #blacklivesmatter: Epistemic positioning, challenges, and possibilities. *Journal of Contemporary Rhetoric*, 5(3/4):78–89, 2015.

[31] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. *ICWSM*, 133:89–96, 2011.

[32] Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Partisan asymmetries in online political activity. *EPJ Data Science*, 1(1):1–19, 2012.

[33] Leticia Bode, Alexander Hanna, Junghwan Yang, and Dhavan V Shah. Candidate networks, citizen clusters, and political expression strategic hashtag use in the 2010 midterms. *The ANNALS of the American Academy of Political and Social Science*, 659(1):149–165, 2015.

[34] Mark Tremayne. Anatomy of protest in the digital era: A network analysis of twitter and occupy wall street. *Social Movement Studies*, 13(1):110–126, 2014.

[35] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. Content and network dynamics behind egyptian political polarization on

twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 700–711. ACM, 2015.

[36] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 290–297. ACM, 2013.

[37] Sarah J Jackson and Brooke Foucault Welles. Hijacking #myNYPD: Social media dissent and networked counterpublics. *Journal of Communication*, 65(6):932–952, 2015.

[38] Erin McClam. Ferguson cop Darren Wilson not indicted in shooting of Michael Brown. *NBC News* (2014). http://www.nbcnews.com/storyline/michael-brown-shooting/ferguson-cop-darren-wilson-not-indicted-shooting-michael-brown-n255391. Accessed 20 June 2016.

[39] Ray Sanchez and Shimon Prokupecz. Protests after N.Y. cop not indicted in chokehold death; Feds reviewing case. *CNN* (2014). http://www.cnn.com/2014/12/03/justice/new-york-grand-jury-chokehold/. Accessed 20 June 2016.

[40] Benjamin Muller and Al Baker. 2 N.Y.P.D. officers killed in Brooklyn ambush; Suspect commits suicide. *New York Times* (2014). http://www.nytimes.com/2014/12/21/nyregion/two-police-officers-shot-in-their-patrol-car-in-brooklyn.html. Accessed 20 June 2016.

[41] Saeed Ahmed and Catherine E. Shoichet. 3 students shot to death in apartment near UNC Chapel Hill. *CNN* (2015). http://www.cnn.com/2015/02/11/us/chapel-hill-shooting/. Accessed 20 June 2016.

[42] Erika Ramirez. Grammys 2015: Pharrell Williams, Beyonce, Prince pay tribute to Black Lives Matter movement. *Billboard* (2015). http://www.billboard.com/articles/events/grammys-2015/6465687/grammys-2015-pharrell-williams-beyonce-prince-black-lives-matter-hands-up-dont-shoot. Accessed 20 June 2016.

[43] Christina Elmore and David MacDougall. Man shot and killed by North Charleston police officer after traffic stop; sled investigating. *The Post and Courier* (2015). http://www.postandcourier.com/article/20150404/PC16/150409635/1180/. Accessed 2016/06/20.

[44] David A. Graham. The mysterious death of Freddie Gray. *The Atlantic* (2015). http://www.theatlantic.com/politics/archive/2015/04/the-mysterious-death-of-freddie-gray/391119/. Accessed 20 June 2016.

[45] Tessa Berenson. Everything we know about the Charleston shooting. *Time* (2015). http://time.com/3926112/charleston-shooting-latest/. Accessed 20 June 2016.

[46] Jon Schuppe. The death of Sandra Bland: What we know so far. *NBC News* (2015). http://www.nbcnews.com/news/us-news/death-sandra-bland-what-we-

know-so-far-n396036. Accessed 20 June 2016.

[47] Marc A Smith, Lee Rainie, Ben Shneiderman, and Itai Himelboim. Mapping Twitter topic networks: From polarized crowds to community clusters. *Pew Research Center*, 20, 2014.

[48] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704. ACM, 2011.

[49] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[50] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[51] Jianhua Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[52] 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.

[53] Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Is language evolution grinding to a halt?: Exploring the life and death of words in english fiction. *arXiv preprint arXiv:1503.03512*, 2015.

[54] Kevin W Boyack, David Newman, Russell J Duhon, Richard Klavans, Michael Patek, Joseph R Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, 6(3):e18029, 2011.

[55] Teun A Van Dijk. Race, riots and the press an analysis of editorials in the british press about the 1985 disorders. *International Communication Gazette*, 43(3):229–253, 1989.

[56] Teun A Van Dijk. New (s) racism: A discourse analytical approach. *Ethnic minorities and the media*, pages 33–49, 2000.

[57] Jaime Ann Banks Lackey. *Framing Social Protest: How Local Newspapers Covered the University of Georgia Desegregation Protests in January, 1961.* 2005.

[58] Nikita Carney. All lives matter, but so does race black lives matter and the evolving role of social media. *Humanity & Society*, 40(2):180–199, 2016.

[59] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pages 577–585, 2014.

[60] Greg Ver Steeg and Aram Galstyan. Maximally informative hierarchical representations of high-dimensional data. *in depth*, 13:14, 2015.

[61] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[62] Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 152–161. Morgan Kaufmann Publishers Inc., 2001.

[63] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.

[64] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288, 2013.

[65] Moontae Lee and David Mimno. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of Empirical Methods in Natural Language Processing*. Citeseer, 2014.

[66] Thang Nguyen, Yuening Hu, and Jordan L Boyd-Graber. Anchors regularized: Adding robustness and extensibility to scalable topic-modeling algorithms. In *ACL (1)*, pages 359–369, 2014.

[67] Thang Nguyen, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. Is your anchor going up or down? fast and accurate supervised topic models. In *North American Chapter of the Association for Computational Linguistics*. Citeseer, 2015.

[68] Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. Using anchors to estimate clinical state without labeled data. In *AMIA Annual*

Symposium Proceedings, volume 2014, page 606. American Medical Informatics Association, 2014.

[69] Yoni Halpern, Steven Horng, and David Sontag. Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299*, 2015.

[70] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

[71] David Andrzejewski and Xiaojin Zhu. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48. Association for Computational Linguistics, 2009.

[72] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[73] Peixian Chen, Nevin L Zhang, Leonard KM Poon, and Zhourong Chen. Progressive em for latent tree models and hierarchical topic detection. *arXiv preprint arXiv:1508.00973*, 2015.

[74] Nathan Hodas, Greg Ver Steeg, Joshua Harrison, Satish Chikkagoudar, Eric Bell, and Courtney Corley. Disentangling the lexicons of disaster response in

twitter. In *The 3rd International Workshop on Social Web for Disaster Management (SWDM'15)*, 2015.

[75] Manhong Dai, Nigam H Shah, Wei Xuan, Mark A Musen, Stanley J Watson, Brian D Athey, Fan Meng, et al. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, 21, 2008.

[76] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

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

[78] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[79] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539, 2014.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn:

Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[81] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, May 2010.