US Presidential Election: What Engaged People on Facebook

Milad Kharratzadeh[1] and **Deniz Ustebay**[2]
[1]Columbia University, New York, USA
[2]McGill University, Montreal, Canada
milad.kharratzadeh@columbia.edu, deniz.ustebay@mail.mcgill.ca

Abstract

We study Facebook posts published by major news organizations in the 10-month period leading to the 2016 presidential election. Our goal is to explore the topics related to the two major party candidates, Hillary Clinton and Donald Trump, and identify the ones that engaged the Facebook users the most. The engagement is measured by the total number of reactions (like, love, sad, etc.), comments, and shares. Using topic modeling with Linear Dirichlet Allocation (LDA) on the Facebook posts and headlines, we identify the top 10 topics related to each candidate and then assess the audience engagement for these topics across the 10 different news organizations. We use Hierarchical Bayesian Models (HBMs) to analyze the data, which allows us to provide separate estimates for each news organization while partially pooling the information across different sources.

Introduction

A recent survey shows that 69% of Americans use social media (Pew Research Center 2017). Among social media websites, Facebook is the most widely used platform with 79% of online adults in the US using it. For many users, Facebook is part of their daily life; 76% of Facebook users visit the site at least once a day. As traditional news sources are losing their audiences to online media, more people are getting their news directly from social media (Gottfried and Shearer 2016).

During a presidential election campaign period many issues are discussed. One of the interesting questions for the post–election analysis is to identify the issues that were important to voters during the campaign and potentially caused them to choose one candidate over the other. In this paper, we study the 2016 presidential election campaign and explore the topics that attracted voters' attention. To this end, we examine the Facebook posts created by 10 major news organizations (TV, radio, newspaper) during the months leading to the election (Jan–Nov 2016) and analyze the user reactions to these posts. These 10 different sources are chosen to represent various editorial views and also audiences from different political views and demographics. We are interested in answering the following questions: (i) "What were the main topics covered by the major news organizations on Facebook during the campaign?", and (ii)"What were the main issues that the followers of each news outlet paid the most attention to?". The results of this analysis can be seen as a first step to understand the voter views on social media during the 2016 election campaign.

Our Contributions. In the first half of the paper, we strive to answer the first question. Our focus is on the two major party candidates, Hillary Clinton and Donald Trump. We use Latent Dirichlet Allocation (LDA), a generative topic modeling algorithm, to identify the top 10 topics for each candidate. As we show later, the discovered topics are coherent and consistent with the main themes of the campaign. We analyze the evolution of the topics over time (i.e., when each topic was covered the most). For example, we show that the topics related to the primary races were mainly covered in the first half of 2016 when the races were actually happening. We also study the distribution of topics across different media (i.e., which topics were covered more by a certain news organization). For instance, our results show that Fox News covered topics unfavorable to Clinton the most. In the second part, we study how user engagement differed for different topics and different news organizations. We use a Hierarchical Bayesian Model (HBM) to estimate the engagement distributions. HBMs enable us to provide separate estimates for each news outlet while partially pooling the information available from other outlets. We also study how the user engagement varied across different media and different topics. We show that the engagement for posts on Clinton differed more significantly over topics. In other words, the users engagement for posts on Trump did not depend much on the topic of the post.

Related Work. LDA-based topic models have been applied to social media data before, e.g., to model the relationship between the text of a blog post and the volume of comments (Yano and Smith 2010), characterizing microblogs (Ramage, Dumais, and Liebling 2010), identifying topics of Tweets (Hong and Davison 2010), and characterizing the relationship between gender, topic, and audience response on Facebook (Wang, Burke, and Kraut 2013). HBMs are also used in the context of social media analy-

sis, e.g., for anomaly detection (Yu, He, and Liu 2015), detecting user attributes (Rao et al. 2011), locating users from Tweets (Ahmed, Hong, and Smola 2013), and identifying Tweets that will become news headlines (Zhang, Liu, and Si 2014). Our work is novel in building a hierarchical Bayesian model on top of the LDA-based topic model to analyze user engagement. Moreover, our work provides unique insight into the coverage of the 2016 election on Facebook by major news organizations as well as the patterns of user engagement across various topics and outlets.

Data and Pre-processing

We use Facebook posts published by 10 major news organizations (ABC, BBC, CBS, CNN, Fox News, NBC, NPR, New York Times, Washington Post, and Wall Street Journal) in the 10-month period leading to the US presidential election (i.e, January 1, 2016 to November 8, 2016). The data is provided in (Martinchek 2016). For each post, we keep the text content as well as the headline of the shared link (if there is one). We do not include the content of the shared articles as many social media users only read the description and headline (Gabielkov et al. 2016). We also record the total number of reactions (like, love, sad, etc.), comments, and shares, which we call *engagement*.

In this study, we focus on the two major party candidates, Hillary Clinton and Donald Trump. Our goal is to analyze the Facebook posts about these two candidates and identify the topics that engaged the audience most across different news organizations.

As the first step, we identify the common topics in the Facebook posts on each candidate. For the topic modeling, we pre-process the data by transforming all letters to lower cases and removing punctuation, numbers, and stop-words. Moreover, we only keep the posts containing the words *clinton* or *trump*. In total we have 87757 posts in our dataset. Out of these, 3131 posts (3.6%) contain both *trump* and *clinton* in the description or headline, 3760 posts (4.3%) contain only *clinton*, and 8317 posts (9.5%) contain only *trump*. This shows that Trump was mentioned significantly more than Clinton. A breakdown of these numbers for the news organizations is shown in Figure 1. Trump was mentioned more by all of the news outlets while Fox had the most mentions of both candidates.

We provide a concrete analysis of the user engagement later in the paper, but here, we examine the average engagement per post for different news organizations. Figure 2 demonstrates the average engagement for all posts as well as the posts only on Trump or Clinton. Fox News has by far the largest mean engagement per post and posts on either candidate has larger engagement than all posts. However, this pattern does not exist among all news outlets. As we see in Figure 2, the difference between average engagements in different outlets is very large (orders of magnitude). Thus, for the rest of the paper, we work with the logarithm of the engagement. Working on the log-scale also helps us handle the extreme long tail of the distribution of engagements (with kurtosis in the order of thousands).

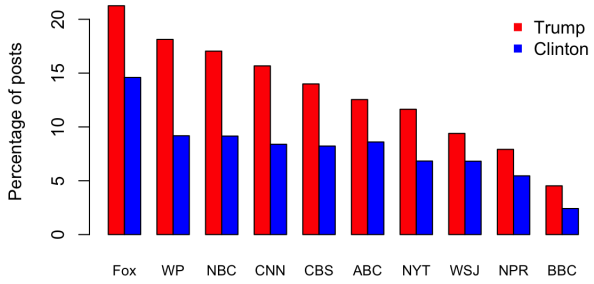


Figure 1: Percentage of posts containing the words Trump or Clinton. Ordered according to percentage of Trump posts.

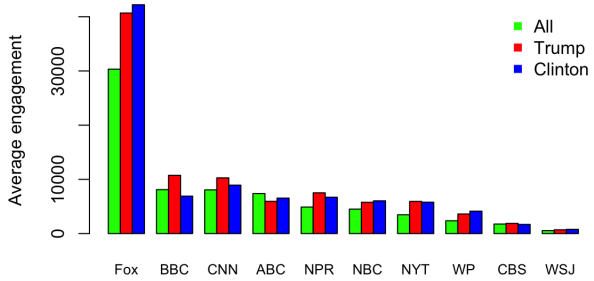


Figure 2: Average engagements for all posts and posts on Trump and Clinton. Ordered according to mean engagement for all posts.

Topic Modeling

Topic models are probabilistic frameworks in machine learning and natural language processing and are used to identify common themes and topics (Blei 2012). Topic models are based on term frequency occurrences in documents in a given corpus. Therefore, the ordering of the words in the documents does not matter. This is sometimes referred to as the exchangeability assumption for the words in the documents which is the basis of bag-of-words models. A complete review of topic models is beyond the scope of this paper.

Latent Dirichlet Allocation. In this work, we use Latent Dirichlet Allocation (LDA) which is one of the most widely used topic models (Blei, Ng, and Jordan 2003). LDA is a generative model where each document can address multiple topics, and the words in the documents are considered as samples from common words employed when discussing these topics. Assume that we have N documents, and the jth word in document i is represented by w_{ij}. These words are drawn from a fixed vocabulary of size V. The generative model can be described in three parts:

- First, each topic imposes a distribution over the vocabulary. Denoting the distribution of the words in topic k by

β_k (which is a V-dimensional vector), we have:

$$\beta_k \sim \text{Dirichlet}(\mu).$$

- Second, for each document, a vector of topic proportions is drawn:

$$\theta_i \sim \text{Dirichlet}(\alpha),$$

where θ_i is the K-dimensional vector of proportions for document i, assuming we have K topics in total.

- Third, for each word in document i, first, a topic is selected and then, a word is sampled from that topic:

$$z_{i,j} \sim \text{Multinomial}(\theta_i) \quad \& \quad w_{i,j} \sim \text{Multinomial}(\beta_{z_{i,j}}).$$

In LDA, it is assumed that we have a fixed number of topics and fixed word probabilities within each topic. In our work, documents correspond to Facebook posts. Since these posts are not very long and they generally address only one topic, we allocate only one topic to each post (the one with the highest proportion in θ_i). We fit LDA on two subsets of data: (1) for the posts about Trump from all ten news organizations and (2) for the posts about Clinton, again, from all news organizations. Here by 'posts about a candidate' we mean the posts that contain the name of the candidate at least once. Each time, we set the total number of topics to be 10. Therefore, we identify the top 10 topics surrounding each candidate on Facebook. We employ the *topicmodels* package in R to fit our data (Grün and Hornik 2011). We use a Gibbs sampler with 10000 iterations (4000 warm up) with a thinning factor of 20 and 10 initial points. The top 20 words for the estimated topics are shown in Tables 1 and 2. Below, we discuss the discovered topics for each candidate. In general, the topics are coherent and are consistent with the major themes of the campaigns.

Topics for Hillary Clinton

Topic 1 is on the relationship of Clinton and the Wall Street, including her earnings from the speeches. Topic 2 is about the use of a private email server when she was the secretary of state, the ongoing FBI and state department investigations, and the WikiLeaks release of emails. The third topic is about the poll results with emphasis on certain states (Florida, Ohio, etc.). Topic 4 is on the Democratic National Convention. Topic 5 is about Clinton becoming the first female candidate of a major party. The sixth topic is about the rallies of the Clinton campaign and the several prominent speakers including Michelle and Barack Obama, Bill Clinton, and Joe Biden. Topic 7 seems to be a mix of a few topics including Clinton's health issues and her policies on gun control. The presidential debate is the eighth topic. The ninth topic is on the Democratic primary race with Bernie Sanders and results in different states. Finally, topic 10 includes criticisms of Trump's policies on a number of issues (tax returns, his businesses, relation with Russia and Putin, economy, etc.).

Topics for Donald Trump

The first topic is on the anti-establishment nature of Trump's presidency and his arguments against both Democrats (e.g., White House) and Republicans (e.g., Paul Ryan). Topic 2 is about the presidential debates. In this topic, we can also see the words *women* and *sexual*; this may be due to the release of the Access Hollywood video (in which Trump made some lewd comments about women) right before the second presidential debate. The third topic is about Trump's rallies, the violence in them, and the protests outside those rallies. The fourth topic involves the Republican primary race and the speculations about the delegates and the National Convention. Topic 5 is about the Democratic primary race and especially Bernie Sanders. This may be because Trump (his potential to become the Republican nominee and his ideas and plans) was one of the main topics discussed during primaries even on the Democratic side. Topic 6 is on the polls. Topic 7 is on Trump's foreign policy including his relation with Russia, immigration, and ISIS. The eighth topic has two themes: vice president choices (Mike Pence from Indiana) and Trump's tax return. Topic 9 is about Trump's plan on building a wall on the border with Mexico and making them to pay for it. Topic 10 does not seem to have a single theme; we could say it is about some of Trump's controversial comments (the third word in the topic).

Topics Coverage Over Time

In Figure 3, we show the percentage of posts allocated to each topic for the months before the election (for November, only the first week is included). For Clinton, topic 9 (primary race with Sanders) was dominant in the early months but then receded after she became the presumptive nominee, and completely vanished after the convention in July. Similarly, topic 4 for Trump (GOP primary race) took a relatively large portion of news before July and a small portion afterwards. Clinton's second topic (the email scandal) got its most coverage during the final months with the peak in November. Her first topic (relationship with Wall Street) was fairly covered throughout the year and her fourth topic (Democratic National Convention) had a large peak in July, the month it happened. The topics on the debates (topic 8 for Clinton and 2 for Trump) had their most coverage in September and October, the months the debates took place. Clinton's third topic and Trump's sixth topic are on polls, and both got a lot of coverage in the first week of November (right before election). For Trump, topic 1 (anti-establishment) had a consistent presence all year. The same presence existed, though not as strongly, for topics 9 (the wall) and 10 (controversial comments).

Topics Coverage in Different News Outlets

In Figure 4, we show the portion of the posts allocated to each topic by each of the 10 news outlets. On Clinton's topics, Fox News had the largest coverage of topic 2 (emails), almost twice as the next network (CBS). They also cover topic 7 (Clinton's health issues and gun policies) significantly more than other outlets. The extensive coverage of these negative topics is in agreement with the right-wing nature of Fox News and their opposition to Clinton. On the other hand, Fox covered topics 8 (debates) and 9 (Democratic primary race) considerably less than others. It is also

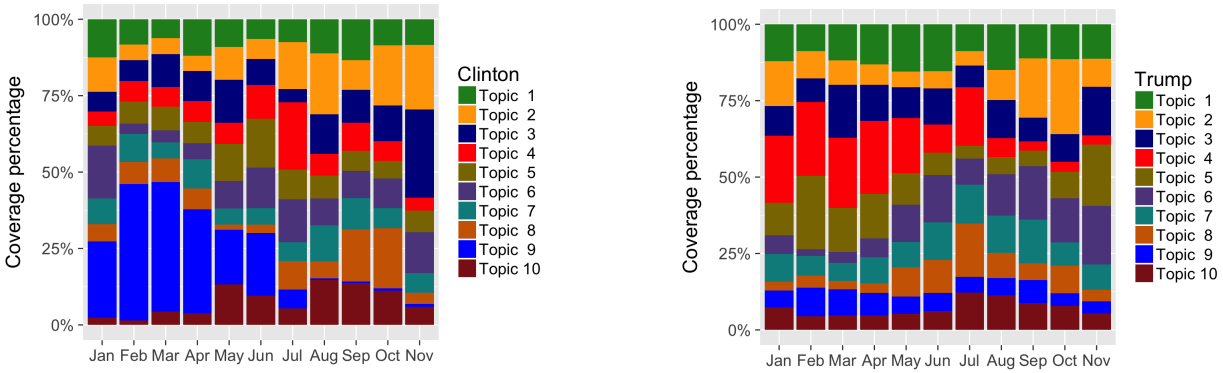


Figure 3: Evolution of the topics over the months leading to the election. The graphs show the portion of the posts in each month corresponding to each topic.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
clinton	state	election	clinton	clinton	clinton	clinton	debate	sanders	trump
hillary	fbi	poll	hillary	hillary	president	hillary	presidential	democratic	donald
campaign	email	voters	national	nominee	hillary	political	watch	bernie	comments
street	emails	race	supporters	party	obama	american	live	primary	tax
wall	private	states	convention	candidate	bill	americans	night	win	russia
behind	secretary	lead	speech	women	rally	country	candidates	nomination	business
million	server	vote	america	gop	barack	health	kaine	iowa	taxes
trail	investigation	support	dnc	democrats	next	issues	final	hampshire	slams
money	department	polls	attacks	white	campaigns	press	fact	nevada	putin
chief	top	florida	attack	republican	vice	policy	tim	victory	find
month	foundation	according	chelsea	woman	video	morning	running	cruz	economic
speeches	director	shows	benghazi	republicans	michelle	read	politics	projects	family
including	wikileaks	among	call	house	united	black	second	results	returns
taking	comey	leads	full	major	biden	gun	debates	super	comment
expected	released	still	spoke	history	pres	world	questions	south	past
presidency	general	days	asked	front	endorses	didn	sunday	sen	plans
pneumonia	release	latest	speaking	warren	husband	friends	monday	caucuses	decades
paid	interview	ohio	give	different	office	weekend	stage	close	bring
fundraising	james	points	half	event	group	better	claims	carolina	wrong
went	judge	republican	mother	presumptive	isis	matter	senator	delegates	remarks

Table 1: Top 20 words for the identified topics in the Facebook posts about Hillary Clinton. The words are ordered according to their probabilities in the corresponding topic. Please see the text for the explanation of the topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
trump	campaign	trump	republican	election	clinton	trump	trump	trump	trump
donald	presidential	donald	gop	primary	hillary	donald	donald	donald	donald
white	debate	rally	cruz	sanders	voters	president	pence	political	comments
house	trump	watch	ted	win	poll	obama	mike	wall	american
ryan	candidate	supporters	nominee	live	race	speech	tax	morning	questions
paul	women	video	convention	states	lead	america	running	read	americans
opinion	night	event	party	state	democratic	melania	attack	plan	interview
washington	candidates	protesters	support	bernie	bill	country	million	mexico	father
really	fact	police	republicans	vote	shows	immigration	vice	behind	ivanka
next	final	stage	national	florida	according	policy	mate	meet	words
change	second	anti	rubio	show	among	world	kaine	top	remarks
speaker	media	romney	nomination	polls	likely	united	university	press	recent
section	claims	woman	front	carolina	latest	attacks	business	money	asked
writes	manager	violence	john	iowa	leads	barack	indiana	christie	war
leaders	debates	story	marco	ohio	holds	putin	things	nbc	muslim
person	monday	outside	kasich	general	percent	history	taxes	chris	family
makes	third	supporter	bush	democrats	four	foreign	twitter	chief	whether
presidency	sexual	crowd	runner	results	points	isis	returns	personal	call
miss	thursday	full	presumptive	hampshire	emails	issues	tim	pay	question
public	different	mitt	delegates	super	fbi	michelle	nation	past	calling

Table 2: Top 20 words for the identified topics in the Facebook posts about Donald Trump. The words are ordered according to their probabilities in the corresponding topic. Please see the text for the explanation of the topics.

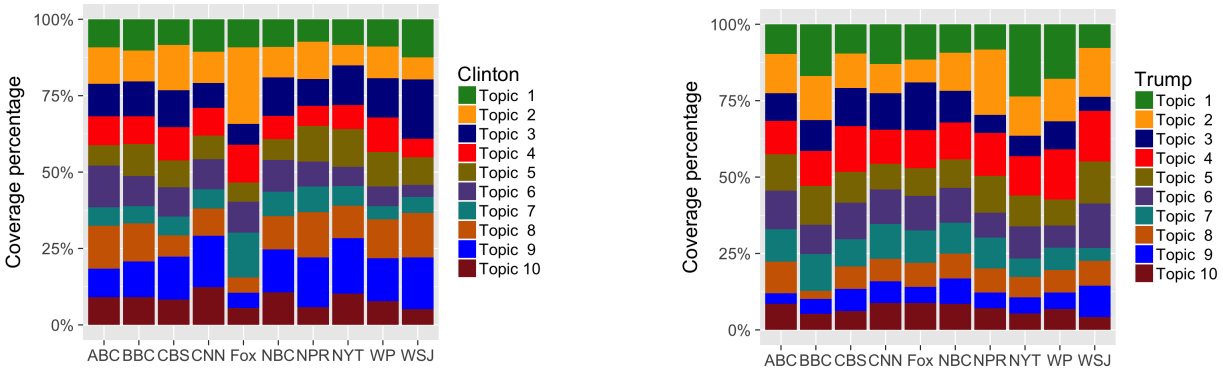


Figure 4: Coverage of the topics over the months leading to the election by different news organizations. The graphs show the portion of the posts corresponding to each topic.

interesting to note that Fox covered the poll results on Clinton (topic 3), which were favorable for her, the least. CNN, NBC, and NYT covered Clinton's criticisms of Trump the most. NYT and WSJ had the least coverage of Clinton's email scandal. On Trump's topics, NYT has the most extensive coverage of topic 1 (the anti-establishment nature of Trump's campaign). NPR covered the debates (topic 2) the most. Wall Street Journal covered Trump's proposal to build a wall (topic 9) and the poll results (topic 6) the most and Trump's rallies (topic 3) the least. Fox News had the most coverage of topic 4 (GOP primary race) and the least coverage of topic 2 (debates).

Analysis of Topics Engagements

In this section, we analyze the engagements of Facebook posts and identify topics that engaged the Facebook audience more. For our analysis, we use Hierarchical Bayesian Models (HBMs). We describe the details of our model below, but first elaborate on the importance of HBMs in our problem. We have 10 news organizations and 10 topics and would like to know how each topic engaged the audience of each news organization. We can do this in three different ways.

- **No Pooling.** We can calculate the mean engagement of each topic for each news outlet separately. However, these news organizations often share many of their followers. Even if we ignore common followers, a topic that is of great interest to group A, is probably of some interest to group B too. By carrying out the estimations completely separately, we would be not using the relatedness of our estimation parameters to improve the accuracy of our estimates and increase the statistical power of our analyses.
- **Complete Pooling.** We can also assume that the audience of all news organizations are exactly the same and a topic engages them all in the same way. Then we can calculate the mean engagement of topics by averaging over all news organizations. This assumption is not correct either; it is not sensible to assume the audiences of Fox News and NYT which have very different political views are engaged in the same way by different topics.
- **Partial Pooling.** HBMs allow us to avoid both complete and no pooling and do something in between. Unlike complete pooling, we allow the parameters for different news organizations to be different. However, unlike no pooling, we assume that these parameters are related with each other in a higher level (they are samples from the same distribution). Therefore, by assuming a hierarchical model, we partially pool the information across different news outlets. The amount of this partial pooling is determined by data.

Hierarchical Bayesian Model

We use the same model for Clinton's topics and for Trump's topics; the following model is independent of the candidate. Assume, we have N Facebook posts in total for M different news organizations, and identified T topics. In our problem, $M = T = 10$ and $N = 6891$ for Clinton and $N = 11448$ for Trump. Let us denote the engagement for post n, where $n = 1, \ldots, N$, by z_n and define:

$$y_n \quad = \quad \log(z_n) - \frac{\sum_{i=1}^{N} \mathbf{1}[m_i = m_n] \log(z_n)}{\sum_{i=1}^{N} \mathbf{1}[m_i = m_n]} \qquad (1)$$

where $m_n \in \{1, \ldots, M\}$ is the news organization ID for post n (i.e., the Facebook page of news organization on which the post appeared), and $\mathbf{1}$ represent the indicator function. We work with log-engagements because in the log domain, the engagement for different news outlets in the same range and the tails are less extreme. The second term on the right hand side is the average of the log-engagement for posts from the same news organization. In sum, we define y_n as the log-engagements which are de-meaned for each news organization.

Then, we define the following likelihood:

$$y_n \sim \mathcal{N}_+(C_{m_n,t_n}, \sigma_{m_n}), \qquad n = 1, \ldots, N \qquad (2)$$

where $t_n \in \{1, \ldots, T\}$ is the identified topic of post n, and $C_{m,t}$ is the likelihood mean for topic t in news organization m. The variation unexplained by our model is encoded by a page-specific standard deviation, σ_{m_n}. $\mathcal{N}_+$ denotes the positive normal distribution which is the ordinary normal distribution but with the probability of zero or less being zero (and

with the rest of the probability density function being scaled up to keep integral at 1). As mentioned above, we believe that $C_{m,t}$ are related for different m (i.e., engagement for topics across different news outlets are related). We model this as follows:

$$C_{m,t} \sim \mathcal{N}_+(0, \tau_t), \qquad m = 1, \ldots, M, \quad t = 1, \ldots, T, \tag{3}$$

where τ_t are hyper-parameters specifying how much partial pooling is done; $\tau_t = 0$ corresponds to complete pooling and $\tau_t = \infty$ corresponds to no pooling at all. We estimate these hyper-parameters from the data in a fully Bayesian manner (by allocating hyper-priors to them). Since, we are working in the log domain and we expect the mean engagements across all news organizations to be no larger than several thousands (see Figure 2), we set the following weak hyper-priors:

$$\tau_t \sim \text{Cauchy}_+(0, 5), \quad t = 1, \ldots, T, \tag{4}$$

where the choice of the half-Cauchy prior for variance is discussed in (Gelman 2006). Similarly, we set the following weak prior for the σ_m:

$$\sigma_m \sim \text{Cauchy}_+(0, 5), \quad m = 1, \ldots, M. \tag{5}$$

Inference

We use the R interface of Stan, a probabilistic programming language, to carry out the inference (Stan Development Team 2016). Stan employs Hamiltonian Monte Carlo (HMC) algorithm to sample from the posterior (Betancourt and Girolami 2015). HMC is a Markov Chain Monte Carlo (MCMC) sampling method and uses a Hamiltonian dynamics model to efficiently explore the parameter space. All the code and data for our work are available at `https://github.com/milkha/FBElec16`. We use 4 chains with 500 iterations (first half as warm up). Convergence is verified by examining the trace-plots and ensuring that Gelman-Rubin convergence criterion, $\widehat{R}$, is less than 1.1 (Gelman and Rubin 1992).

Results

In this part, we examine the produced samples from the posterior distributions for the parameters of our model. We start by $C_{m,t}$ which denotes the likelihood mean for the de-meaned log-engagement of topic t in news organization m. Since, we work in the log domain, $\exp(C_{m,t})$ can be interpreted as the percentage increase in the engagements of a post compared to the mean engagement (since we de-mean the log, the mean is the geometric mean). Therefore, $C_{m,t} = 0.1$ corresponds to a $\exp(0.1) \approx 10\%$ increase. The results for both Clinton (in blue) and Trump (in red) are shown in Figure 5. For each parameter, we show the median and the 50% Bayesian interval[1]. Below, we discuss the results for each candidate.

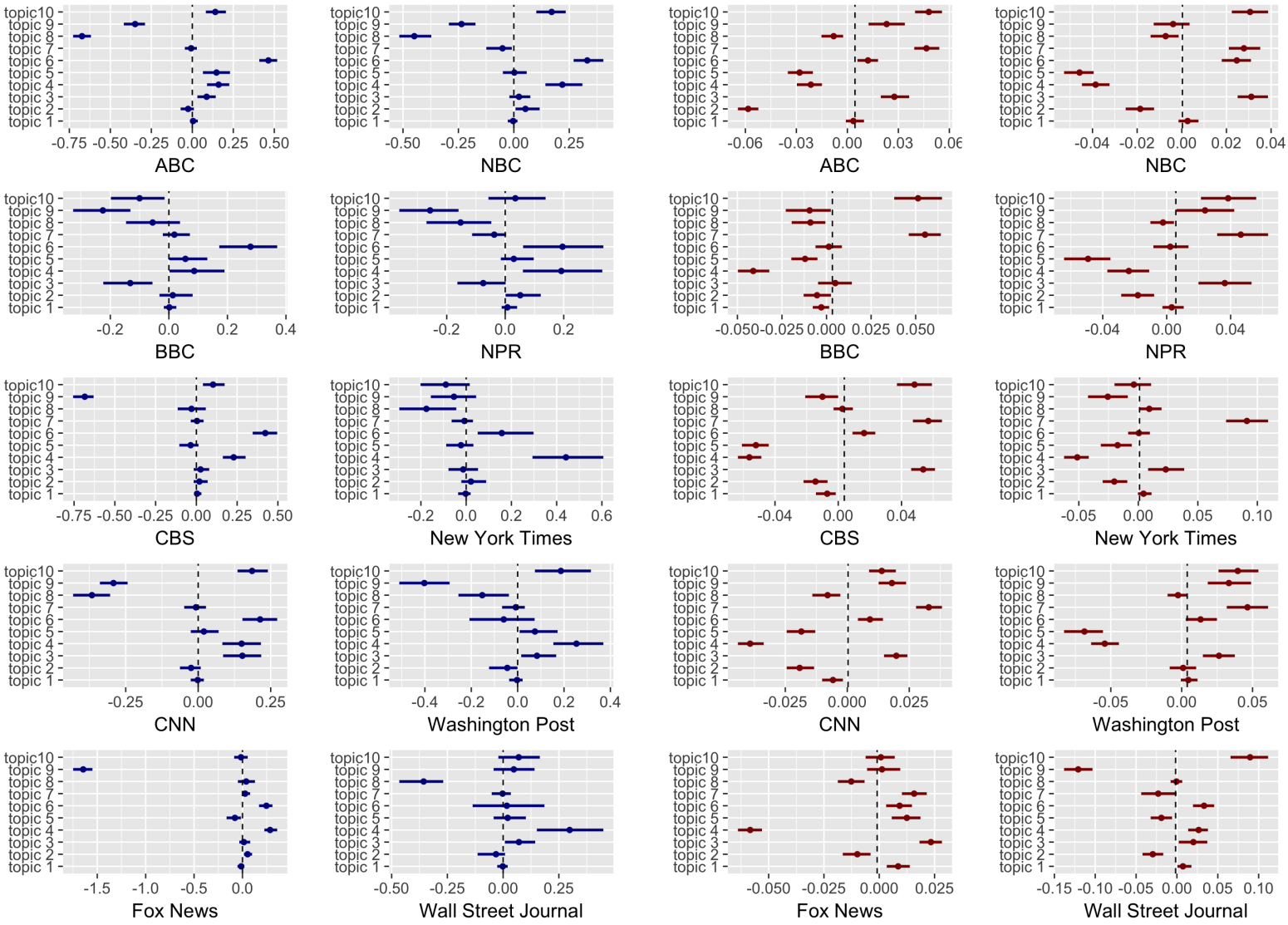
[1]The 50% interval, unlike the 95% interval, is more computationally stable and gives a better sense of where the parameters are (rather than offering an unrealistic near-certainty).

User Engagement for Hillary Clinton Topics. For several news outlets (ABC, NBC, BBC, NPR, CBS, and CNN) posts related to Clinton's rallies (topic 6) got the highest user engagement. Topic 4 (Democratic National Convention) received the highest user engagement for the rest (NYT, Fox, Washington Post, and WSJ). Democratic primary race (Topic 9) and the debates (topic 8) engaged the audience the least. For Fox News, Topic 9 got significantly less engagement than average (about 20%). This is somewhat expected that the right-wing audience of Fox do not care much about the primary race of the other party. Clinton's criticisms of Trump (Topic 10) also received a lot of attention in ABC, NBC, CBS, CNN, and Washington Post.

User Engagement for Donald Trump Topics. There are a few topics for Trump that engaged the Facebook users significantly more than average. Trump's controversial comments on a number of issues (topic 10) got more than average engagement in all outlets except Fox News and NYT. Topic 7 (Trump's foreign policies) raised significant engagement in all news organizations except Wall Street Journal. Trump's rallies (topic 3) received high attention in ABC, NBC, NPR, CBS, NYT, CNN, Washington Post, Fox News, and WSJ. Topics 4 and 5 (primary races) got the least attention almost all the time. We can examine the news organizations separately as well. The largest source of attention for NYT followers was–by far–Trump's foreign policies, whereas for WSJ followers, was his controversial comments. For Fox News, Trump's rallies got the most attention, whereas the primary race received significantly less engagement than average.

Variation of Engagements. Next, we examine the parameters σ_m and τ_t for the two candidates. These parameters measure the engagement variability in two different dimensions. σ_m models how the engagements vary across different topics inside the same news organization; larger σ_m indicates that different topics engaged the audience more differently. On the other hand, τ_t quantifies the amount of partial pooling across different news outlets for the same topic; larger τ_t means that the topic's engagements were more different across different outlets.

The inference results are shown in Tables 3 and 4. The results for the two candidates are very different. Both σ_m and τ_t are significantly smaller for Trump than Clinton. This indicates that (i) within each news organization, different Trump's topics engaged the audience more similarly than different Clinton's topics, and (ii) Trump's topics engaged people more similarly across different news outlets than Clinton's topics. These results show that the attention Trump gets is less dependent on the specific topic or media, and is more dependent on the fact that the post is about Trump. On the other hand, for Clinton, the engagement differs significantly more, both across the topics and across the news organizations. This agrees with the general themes of the two campaigns; Clinton had a more policy-based strategy focusing on different issues, whereas Trump had a self-centered campaign which was based more on his persona rather than any single policy.



(a) Topics identified for Hillary Clinton

(b) Topics identified for Donald Trump

Figure 5: The median and 50% Bayesian intervals for $C_{m,t}$ for different topics over different news organizations. Since, we work in the log domain, $\exp(C_{m,t})$ can be interpreted as the percentage increase in the engagements of a post compared to the mean engagement (since we de-mean the log, the mean is the geometric mean). Parameters greater than zero correspond to an increase in the engagement. For example, $C_{m,t} = 0.1$ corresponds to a 10% increase ($\exp(0.1) = 1.1$) and $C_{m,t} = -0.1$ corresponds to a 10% decrease ($\exp(-0.1) = 0.9$).

σ_m—-Clinton	mean	sd	25%	50%	75%
BBC	0.80	0.04	0.78	0.80	0.83
CNN	0.97	0.02	0.96	0.97	0.99
NBC	1.02	0.02	1.00	1.02	1.03
CBS	1.03	0.03	1.01	1.02	1.04
WSJ	1.05	0.04	1.02	1.05	1.07
Fox	1.11	0.03	1.09	1.11	1.13
ABC	1.13	0.02	1.11	1.13	1.14
WP	1.27	0.04	1.24	1.27	1.30
NYT	1.32	0.05	1.28	1.32	1.35
NPR	1.33	0.05	1.29	1.32	1.36

σ_m—Trump	mean	sd	25%	50%	75%
BBC	0.11	0.00	0.10	0.11	0.11
CNN	0.11	0.00	0.11	0.11	0.12
Fox	0.11	0.00	0.11	0.11	0.11
NBC	0.13	0.00	0.13	0.13	0.13
ABC	0.15	0.00	0.15	0.15	0.15
CBS	0.16	0.00	0.16	0.16	0.16
WP	0.17	0.00	0.17	0.17	0.18
WSJ	0.17	0.01	0.17	0.17	0.17
NYT	0.18	0.00	0.17	0.18	0.18
NPR	0.19	0.01	0.18	0.18	0.19

Table 3: Inference results for σ_m for topics identified for Clinton and Trump. Topics are sorted according to the mean.

τ_t—Clinton	mean	sd	25%	50%	75%
Topic 1	0.05	0.04	0.02	0.04	0.07
Topic 7	0.09	0.07	0.04	0.08	0.13
Topic 2	0.10	0.06	0.05	0.09	0.14
Topic 3	0.14	0.07	0.09	0.13	0.18
Topic 5	0.15	0.12	0.08	0.12	0.17
Topic 10	0.20	0.08	0.14	0.19	0.24
Topic 4	0.33	0.13	0.25	0.31	0.39
Topic 6	0.36	0.11	0.27	0.34	0.42
Topic 8	0.40	0.12	0.32	0.37	0.46
Topic 9	0.73	0.21	0.58	0.69	0.82

τ_t—Trump	mean	sd	25%	50%	75%
Topic 1	0.01	0.01	0.01	0.01	0.02
Topic 6	0.02	0.01	0.02	0.02	0.03
Topic 8	0.02	0.01	0.01	0.01	0.02
Topic 2	0.03	0.01	0.02	0.03	0.04
Topic 3	0.04	0.01	0.03	0.04	0.04
Topic 4	0.05	0.01	0.04	0.05	0.06
Topic 5	0.05	0.01	0.04	0.04	0.05
Topic 9	0.05	0.02	0.04	0.05	0.06
Topic 7	0.06	0.02	0.05	0.06	0.07
Topic 10	0.06	0.02	0.04	0.05	0.06

Table 4: Inference results for τ_t for topics identified for Clinton and Trump. Topics are sorted according to the mean.

For both candidates, BBC and CNN had the least variation of engagement across different topics, i.e., their followers react more similarly to different topics. Also for both candidates, NYT and NPR had the largest variation in the engagement across different topics, i.e., their followers' engagement is more topic-based. In terms of topics, Clinton's ninth topic (primary race with Sanders) had the largest variation across news organizations, i.e., it engaged the audience most differently in different outlets. On the other hand, topic 1 (her relationship with Wall Street) had the least variation, i.e., the followers of different outlets were more similarly engaged by this topic. For Trump, topic 1 (relationship with establishment) had the smallest and topic 10 (his controversial comments) had the largest variance in engagement across different outlets.

Discussion and Future Work

Social media, including Facebook, are replacing the traditional ways people used to get their information. More and more people rely on social media as the primary, and often only, source of their news. This provides an opportunity to study the users' behavior in the social media to identify the issues that are more important to them and to which they pay the most attention. In this paper, we took the first step to address these questions in the context of 2016 US presidential election. We analyzed the Facebook posts generated by 10 major news organizations in the months leading to the 2016 presidential election. We quantified the users' attention to different posts by a metric we called *engagement*, which is the sum of the comments, shares, and reactions (like, love, etc.) on each post. We analyzed both the contents of the posts and the patterns of user engagement.

Topics. The analysis of the contents of the posts showed Trump was covered significantly more than Clinton (Figure 1). We then applied LDA, a topic modeling algorithm, to the contents of the posts and identified 10 topics for either of the candidates (Tables 1 and 2). We also analyzed the distribution of these topics, both across the time (Figure 3) and across different news outlets (Figure 4). The discovered topics and their distributions were consistent with the general themes of the campaign. For instance, we observed that the topics related to primary races were covered more in early months of 2016 when the races were actually happening. We also observed that Fox News covered the topic unfavorable for Clinton significantly more than other outlets.

Engagements. In the second part of the paper, we strived to examine how different topics engaged users across different news organizations. We used a hierarchical Bayesian model to estimate the engagement distribution for topics over different news outlets. We observed that generally, posts related to rallies and controversial comments and criticisms gained the audience attention more whereas topics related to primary races grabbed the least attention (Figure 5). We also studied the variation of engagement across topics and across different outlets (Tables 3 and 4). The variation in both directions are significantly larger for Clinton. In other

words, the engagement for posts about Clinton are more sensitive to the topic and media. Whereas for Trump, the engagements are less dependent across the topics and media.

Methodology. We believe that hierarchical Bayesian models are very effective tools in social media analysis. HBMs' partial pooling of information across different sources increases the statistical power and provide us with great insight of how these different sources are related. For instance, in this paper, by analyzing the hyper-parameters σ_m and τ_t, we were able to draw useful conclusions about the patterns of engagement for different candidates. Moreover, analyzing HBMs is facilitated by Stan which is a user-friendly, effective, and fast tool for inference.

Future Directions. There are several future directions for this work. In this paper, we tried to only report the results of our theoretical analyses, without providing much political commentary. Our results can be further used by social/political scientists to draw conclusions and shed light on the how users are engaged in social media. We also believe that there are much more analysis that can be done on this data by social/political science, machine learning, and statistics communities[2]. For instance, one can do a sentiment analysis of the posts and identify how negatively or positively a news organization covers a story and how that affects the engagement. In this work, we allocated only one topic to each post. Another idea might be to keep all the topic weights and use that information as a predictor in modeling the engagement. One can also use a more sophisticated topic modeling algorithm where the number of topics is not set in advance and is learned from the data. Finally, our proposed framework can be used in other settings as well; for instance, instead of major news organizations, one can focus on Facebook pages disseminating the so-called "fake news". The topics most covered by these pages can be identified. Also, the analysis of user engagement for these fake news may provide useful insights into what draws people towards these alternative, often unreliable sources.

References

Ahmed, A.; Hong, L.; and Smola, A. J. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proc. Int. Conf. World Wide Web*, 25–36.

Betancourt, M., and Girolami, M. 2015. Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications* 79.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(1):993–1022.

Blei, D. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Gabielkov, M.; Ramachandran, A.; Chaintreau, A.; and Legout, A. 2016. Social Clicks: What and Who Gets Read on Twitter? In *ACM SIGMETRICS / IFIP Performance*.

[2]To facilitate these efforts, we provide the code and data of our work: `https://github.com/milkha/FBElec16`

Gelman, A., and Rubin, D. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 457–472.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1(3):515–534.

Gottfried, J., and Shearer, E. 2016. Social Media Fact Sheet. Pew Research Center.

Grün, B., and Hornik, K. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40(13):1–30.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in Twitter. In *Proc. Workshop on Social Media Analytics*, SOMA '10, 80–88.

Martinchek, P. 2016. 2012-2016 Facebook Posts. `https://data.world/martinchek/2012-2016-facebook-posts`. [Online; accessed 13-Jan-2017].

Pew Research Center. 2017. Social Media Fact Sheet. `http://www.pewinternet.org/fact-sheet/social-media/`. [Online; accessed 13-Jan-2017].

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proc. Int. Conf. Web and Social Media*.

Rao, D.; Paul, M.; Fink, C.; Yarowsky, D.; Oates, T.; and Coppersmith, G. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proc. Int. Conf. Weblogs and Social Media*.

Stan Development Team. 2016. RStan: the R interface to Stan. R package version 2.14.1.

Wang, Y.-C.; Burke, M.; and Kraut, R. E. 2013. Gender, topic, and audience response: An analysis of user-generated content on Facebook. In *Proc. Conf. Human Factors in Computing Systems*, 31–34.

Yano, T., and Smith, N. 2010. What's worthy of comment? Content and comment volume in political blogs. In *Proc. Int. Conf. Web and Social Media*.

Yu, R.; He, X.; and Liu, Y. 2015. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data* 10(2):18.

Zhang, D.; Liu, Y.; and Si, L. 2014. Which Tweets will be headlines? a hierarchical Bayesian model for bridging social media and traditional media. In *Proc. Workshop on Social Network Mining and Analysis*, 1–9.