Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

US Presidential Election: What Engaged People on Facebook

Milad Kharratzadeh,[1] **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, comments, and shares. Using topic modeling with Linear Dirichlet Allocation (LDA) on the Facebook posts, we identify the top 10 topics related to each candidate and then assess the audience engagement for these topics across 10 different news organizations. We use Hierarchical Bayesian Models (HBMs) to analyze the data, which allow us to partially pool the information across different sources.

Introduction

A recent survey shows that 69% of Americans use social media (Pew Research Center 2017). 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.

During an 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. We study the 2016 presidential election campaign and explore the topics that attracted voters' attention. We examine the Facebook posts created by 10 major news organizations 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?". Our work is a first step to understand the voter views on Facebook during the 2016 campaign.

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. 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.

First, we identify the common topics in the Facebook posts on each candidate. For the topic modeling, we preprocess 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.

Topic Modeling

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. 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. 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

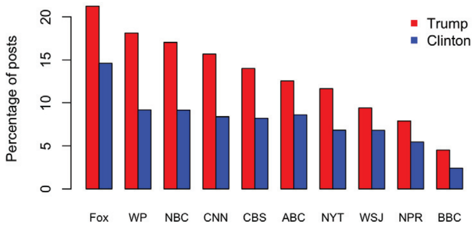


Figure 1: Percentage of posts containing the words Trump or Clinton. Ordered according to percentage of Trump posts.

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. 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 2, 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).

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) which allow us to partially pool the information across different sources. 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

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.

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:

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

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, we set the following weak hyper-priors:

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

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:

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

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

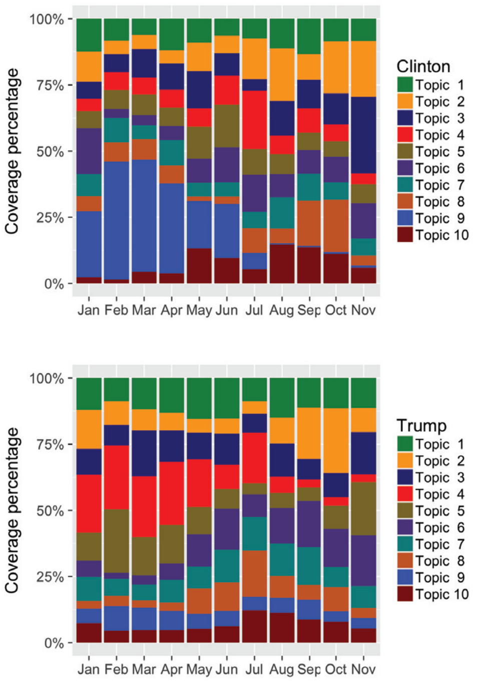


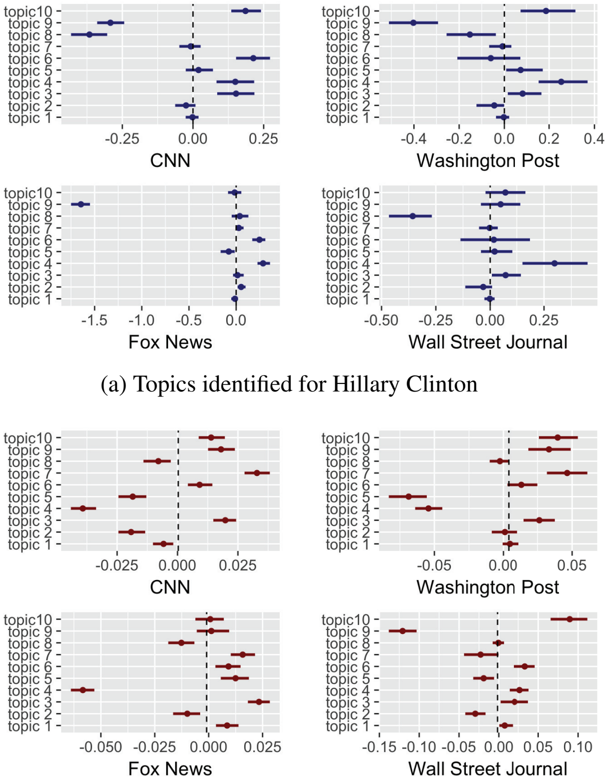
Figure 2: Evolution of the topics over the months leading to the election. The graphs show the portion of the posts in each month corresponding to each topic.

(MCMC) sampling method and uses a Hamiltonian dynamics model to efficiently explore the parameter space[1].

We examine $C_{m,t}$ 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 demean 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 and Trump are shown in Figure 3. Due to limited space, we only show results for a subset of news corporations[1].

Clinton. 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 (DNC) received the highest user engagement for the rest (NYT, Fox, WaPo, and WSJ). Democratic primary race and the debates 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 WaPo.

Trump. 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. Topics 7 (Trump's foreign policies) and 3 (rallies) raised significant engagement in most news organizations. 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.



(a) Topics identified for Hillary Clinton

(b) Topics identified for Donald Trump

Figure 3: The median and 50% Bayesian intervals for $C_{m,t}$. $\exp(C_{m,t})$ is the percentage increase in the engagements compared to average.

References

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. *J. Machine Learning Research* 3(1):993–1022.

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

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.

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

[1] All the code and data for our work, as well as an extended version of this paper with more results, are available at https://github.com/milkha/FBElec16