Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Tarrek Shaban July 20, 2017

Trumping the Polls: Event Analysis During the 2016 Election

by

Tarrek A. Shaban

Jinho D. Choi
Adviser

Department of Mathematics and Computer Science

Jinho D. Choi

Adviser

Alan Abromowitz

Committee Member

James Lu

Committee Member

2017

Trumping the Polls: Event Analysis During the 2016 Election

By

Tarrek A. Shaban

Jinho D. Choi

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2017

Abstract

Trumping the Polls: Event Analysis During the 2016 Election
By Tarrek A. Shaban

Since its introduction in 2006, Twitter has grown into an integral venue for political discourse. With this in mind, it is not surprising that Twitter and other social media services have played an important role in shaping the political debate during the 2016 presidential election. The dynamics of social media provide a unique opportunity to detect and interpret the pivotal events and scandals of the candidates quantitatively. This paper examines several text-based analysis to determine which topics have a lasting impact on the election for the two main candidates, Clinton and Trump. About 135.5 million tweets are collected over the six weeks prior to the election. From these tweets, topic clustering, keyword extraction, and tweeter analysis are performed to better understand the impact of the events occurred during this period. This analysis builds upon a social science foundation to provide another avenue for scholars to use in discerning how events detected from social media show the impacts of campaigns as well as campaign the election.

Trumping the Polls: Event Analysis During the 2016 Election

By

Tarrek A. Shaban

Jinho D. Choi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2017

Acknowledgements

I wish to thank my three committee members for their generous grant of time and expertise. This thesis would not have been possible without the ample advice of Dr. Alan Abromowitz and Dr. James Lu. However, I am indebted to my advisor, Dr. Jinho Choi, for his endless encouragement, direction, and patience over the past year and a half. I would like to acknowledge Lindsey Hexter for her assistance in coding. Similarly, Mathew Sperling who helped shape the thesis through many late-night conversations regarding this work and for his assistance in editing the final manuscript. Moreover, each of the faculty members I have had the privilege to interact with while at Emory deserve recognition for this thesis is the culmination of my course of study in both Computer Science and Political Science.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During 2016 alone, *The New York Times* published over 800 articles reporting on the events surrounding Hillary Clinton and Donald Trump. Many of these articles, through their air of urgency, conveyed the impression that the election might hinge on the events covered therein. Political pundits endlessly pontificated on how each development would have transformed the campaign.

It has been shown, however, that relatively few events *actually* change the course of the election [52]. Political scientists have proven their ability to accurately predict the two-party national vote even months before the election happens [20, 21, 25]. The 2016 U.S. presidential election was not an exception to this [25]; although Hillary Clinton and Donald Trump each faced an onslaught of potentially debilitating scandals [51, 32], the popular vote was still predicted accurately before many of these events transpired [25].

Thus, a critical question for both observers to and the participants in campaigns is: Which events hold genuine sway over the outcome of the election? This is addressed by applying natural language processing (NLP) and information retrieval (IR) techniques to dataset of over 135.5 million tweets collected during the 2016 presidential election campaign. To that end, several notable events are interpreted, providing quantitative analysis to a practice which has been primarily qualitative.

The assumption underlying this approach is that there is meaningful political discourse occurring on social media [60, 57, 48]. Donald Trump's efficacious use of the platform proved pivotal in his bid for the White House [40]. Indeed, Trump's victory over Clinton has cemented his legacy as the first Twitter president. Moreover, social media allows anyone, anywhere to engage in conversations and debates of the sort which could have been reserved for the dinner tables, water coolers or other intimate social events [9].

Because of this, social scientists, journalists and businesses alike view social media as measurable conduits of public opinion. For example, O'Connor et al (2010) found that sentiment derived from Twitter is a good predictor of presidential approval rating [45]; Ozdikis et al (2012) used Twitter to detect new trends to provide a competitive advantage for businesses [47]; and Petrovi

et al (2013) determined that journalists can treat Twitter like a hyper-local newswire service [49].

In this paper, news articles published during the election are first clustered using vector space models and keywords are extracted from each of the resulting clusters (section 3.2). Various scores are run for each Tweet to quantitatively evaluate (Section 3.3.1). Additional conversation keywords are extracted from our Twitter corpus to represent the topics of discussions on social media using a novel variation of TF-IDF (section 3.3.2). Event keywords are then used to identify which tweets are good representatives of those discussions. Both the quantitative and qualitative analyses on the trends and topics of the election by this work (section 4.1) provide a foundation for future exploratory work. The datasets collected for this research are publicly available for further work.[1]

[1] `http://nlp.mathcs.emory.edu/election-2016`

Chapter 2

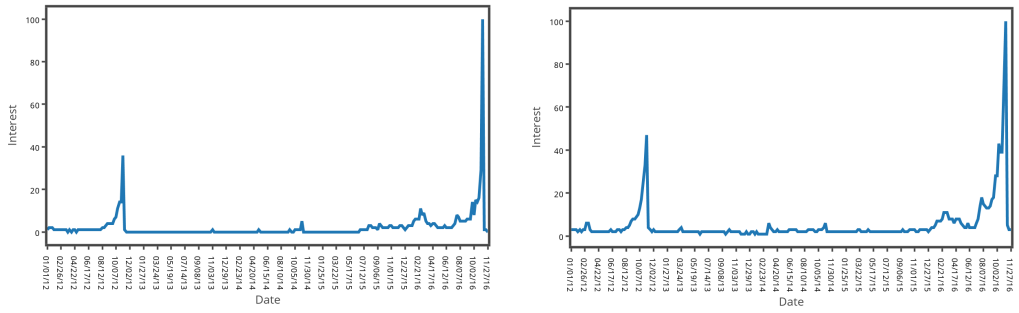
Background

2.1 Predicting Presidential Elections

There was an unusual level of interest in opinion polling during the 2016 presidential election. Frank Bruni, a notable *New York Times* columnist, wrote a piece in January of 2016 titled "Our Insane Addiction to Polls" [17]. Figure 2.1 demonstrates his observation by highlighting the 60% increase in the number of searches for the terms `polls` and `FiveThirtyEight` from the 2012 election to the end of the 2016 election.[1] Indeed, everyone from the news media to the candidates themselves obsessed over the latest poll numbers.

However, opinion polling is not the only method that can be used to predict the election. As mentioned, *supra*, scholars have been able to predict accurately the popular vote of the election using fundamentals models months

[1]The data used in figure 2.1 was obtained using Google Trends: `https://trends.google.com/`



(a) Polls (b) FiveThirtyEight

Figure 2.1: Searches for the terms `polls` and `FiveThirtyEight` on Google

before the election [20, 21, 25]. This section reviews both opinion polling (section 2.1.1) and fundamentals models (section 2.1.2) as methods to predict presidential elections. Then the performance of these various approaches during the 2016 election is reviewed (section 2.1.3).

2.1.1 Opinion Polling

The first scientific opinion polling was conducted by Gallup in 1936 [33]. Since then, there has been a considerable rise in the level of sophistication in which pollsters approach the task [33]. Though individual opinion polls are covered endlessly by the news media, aggregate polling models are a more accurate approach. During the 2016 presidential election, these models included ones produced by FiveThirtyEight and RealClear Politics.[2]

[2]FiveThirtyEight's Polls-Only Model: `https://projects.fivethirtyeight.com/2016-election-forecast/`; RealClear Politics: `https://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton-5491.html`

Polling models use polls produced by third-party organizations like Pew and Ipsos to predict the election with a higher degree of accuracy than any single poll alone. [3] This is because polling errors, which result in a skewed representation of the electorate, are common. For this reason, virtually all models based upon opinion polls weigh the results of any new poll based upon the historical record and practices of the organization (*e.g.* house effect) which sponsored the survey.[4] This approach can be applied to virtually anything that centers around opinion polling. For example, FiveThirtyEight has modeled President Trump's approval and disapproval ratings.[5]

Modeling opinion polls can accurately represent the current political situation facing the two candidates [39, 56]. However, this is only the case in the weeks leading up to the election [19]; the father out from the election, the less representative polls are. This is particularly true when attempting to assess the impact of events on the standing of the candidates. Wlezien and Erikson (2002) found that polling is far more volatile before the conventions than after the conventions. Moreover, polling is dependent on finding willing

[3]Pew Research Center: `http://www.pewresearch.org/`; Ipsos Game Changers Polls: `https://www.ipsos.com/en-us/news-and-polls/overview`

[4]A popular polling model during the 2012, 2014, and 2016 campaigns was produced by FiveThirtyEight; the organization publishes its pollster rankings online: `https://projects.fivethirtyeight.com/pollster-ratings/`

[5]See this site for FiveThirtyEight's presidential approval rating forecast: `https://projects.fivethirtyeight.com/trump-approval-ratings`

respondents to answer the phone when pollsters call. In the modern era, less than 1% of those called actually pick up.

2.1.2 Fundimentals Models

Fundamentals models are most often created by political scientists. These models are statistical approaches built around a theoretically and historically based presupposition of how the electorate behaves [42, 4, 23]. The best of such models are able to predict the results of the election months beforehand [8, 43, 25, 21, 20]. Each model relies on a unique combination of factors to predict the two-party popular vote. The most common historical-fundamentals used in predicting elections are: economic and ideological indicators, presidential approval ratings, and the predisposition of the electorate to a candidate. To better understand the composition and utility of historial-fundamentals models, consider the following three scholars' approaches which are summarized by figure 2.2. As is evident in figure 2.2, the accuracy of fundamentals models varies each year. This reflects the nature and intent of the models variables.

The first of these models is the Time-for-Change model (TFC) by Alan Abramowitz [4, 5, 2, 6, 7, 3]. Like most other fundamentals models, TFC, first

published to predict the 1988 election cycle, forecast the incumbents party's share of the two-party popular vote [4]. It uses three fundamentals to do so: the incumbent president's net approval rating (the percent of Americans who approve of his performance minus the percent who disapprove), the growth in the gross domestic product (GDP) in the second quarter of the election year, and if the current president is running for his second term or not [4]. Abramowitz's model has correctly predicted every election it has been used to forecast up to and inclusive of the 2012 race with an average error of 1.7 percentage points.

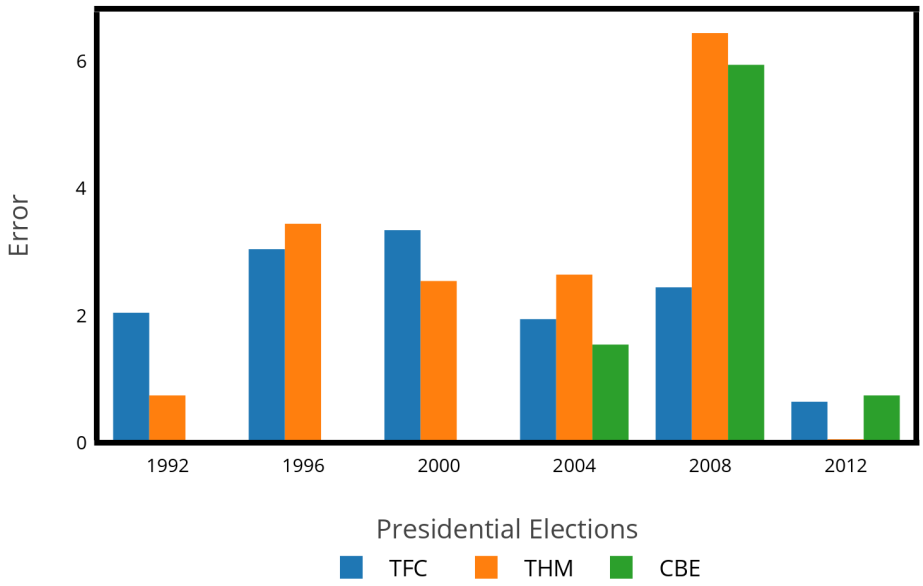


Figure 2.2: Accuracy of Fundamentals Models: 1992 to 2012

James Campbell offered two distinct models to forecast the 2016 pres-

idential election [22]. The trial heat and economy model (TEH), was first published by Campbell and Wink in 1990 [23]. TEH employs the incumbent party's share of the two-party popular vote at Labor Day based upon polls, and the real growth in the second quarter GDP [23, 22]. The second of the two models, the convention bump and economy model (CBE), was introduced in 2002 [24]. Similarly to TEH, this model uses both polling data and GDP growth [24]. The difference in the polling data is in two parts. As the name suggests, CBE takes into account the in-power party's support before the convention as measured by their share of the two-party vote and the change in support between then and after the second convention. Since 1992 and 2004 respectively, the TEH's average miss was by 2.4 percentage points and CBE's average miss was by 2.6 percentage points.

2.1.3 2016 Cycle Accuracy

Though Trump carried the white house, Clinton won the popular vote. The final results indicate that Clinton claimed 51.1% of the national two-party popular vote. First, consider the opinion polling at two points during the election: the day of the first debate (26 September) and the day before the election (7 November). FiveThirtyEight forecast that Clinton's support was

50.8% on the 26 September and 51.9% on 7 November. Figure 2.3 shows FiveThirtyEight's forecast from September 25th to November 7th. RealClear Politics' poll average showed Clinton at 51.3% on the 26 September and 51.8% on 7 November. These results are both is in line with the findings by Campbell and Wink [23]: as polling approaches the election day, the results are more accurate.

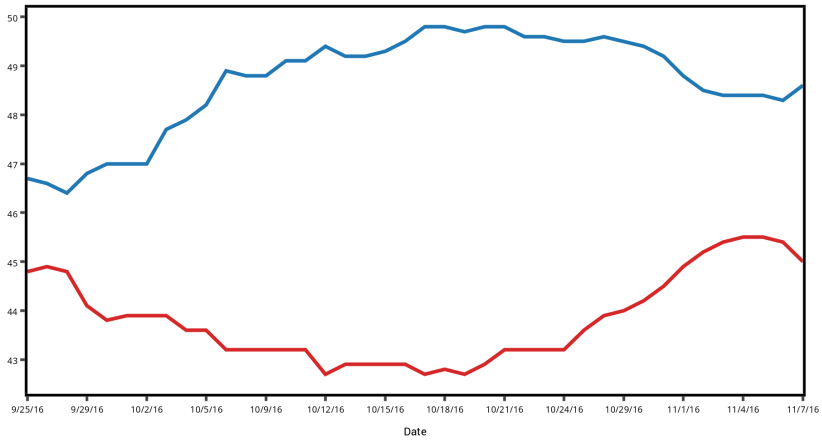


Figure 2.3: FiveThirtyEight's 2016 Polls-Only Model for Clinton and Trump

Compare the results of these polling models to the fundamentals models discussed in section 2.1.2. Abramowitz's TFC model predicted that Clinton would win 48.6% of the two-party vote. Campbell's TEH model predicted that Clinton would win 50.7% of the two-party vote. Campbell's CBE model predicted that Clinton would win 51.2% of the two-party vote. These represent three of the ten models included by The American Political Science

Association in a presidential election forecast symposium. Overall, the median of the forecasts issued by the symposium members predicted 51.1% share of the vote for Clinton, which actually was the result come election day.

2.2 Campaigns and Polling

There is ample evidence to suggest that historical-fundamentals models can accurately predict an election months beforehand [25]. Yet, there is a large public demand for daily polls [58]. Though there are various explanations for this demand [41, 31], our focus is on the utility of polling to campaigns. Since Kennedy's bid for the White House [34], polls have had a profound impact on the way elections are won [37, 36, 34]; they allow candidates to make decisions based upon how the public feels. Campaigns now can act upon polls as a source of public sentiment when purposing policies or deciding how to handle an event [36, 35]. Furthermore, such benefits are available to the victor of the election when they begin to govern [35].

However, this poses an implicit contradiction: How can social scientists accurately predict the outcome of an election so far ahead of time if a candidate's strategy, decided in part by using the information provided by polls, change the course of an election? Gelman and King resolve this by

suggesting that a campaign's primary function is to enlighten voters to their candidate and their policy preferences [29]. Scholars, using historical-fundamental approaches, predict what the electorate's preferences ought to be while the candidates work to enlighten voters to these preferences before election day [29, 10].

One critical assumption made by Gelman and King is that both campaigns operate in a balanced environment with similar resources and staff talent levels [29]. Thus, neither candidate is able to gain a perceivable advantage in voter influence that would drastically change the political scientists' early models. Assuming that two candidates are running balanced campaigns, only an event with a sizable impact on the electorate can cause a break from the election's deterministic outcome (*i.e.* the historical-fundamentals models' predictions) [59]. The paper adopts this concept of a "shock" from Wlezien and Erikson [59].

2.3 Corpus Considerations

Twitter is a social media outlet composed of short microblogs, each under 140 characters. Unlike Facebook, a user on Twitter can follow any other user's activity on the site [38]. Only if a user specifically opts-out of a public

profile is there a mechanic similar to friend requests on Facebook. Though this seems like a minute difference, it changes the dynamics of user interaction and behavior on the networks [18].

Processing statuses from Twitter also poses some challenges. The language used in tweets is informal and the structure of each post is often misconstructed. There are several platform specific features – re-tweets, at-replies, etc – which must be accounted for. [30]. For example, researchers must decide whether to remove the re-xtweet identifier RT, non-ASCII characters, and hashtags. Though each of these unique characteristics of a tweet might require extra effort on the part of the researcher, they also add potential value if used as meta-information [38, 54, 61]. For example, Conover et al (2011) found that re-tweeting is a politically polarizing activity but at-mentioning users in a tweet is not [27].

Sentiment may be hard to discern because a lot of information and likely various sentiments are expressed in a relatively compressed message [15]. However, there are some advantage in using twitter data. Tweets often include emoji chosen by users. Emoji can be thought of as a source of sentiment annotation [44]. The emoji serves as a visual representation of the user's intent which could help label messages as sarcastic or ironic, challenging feats

for NLP [13, 28]. Hashtags and user tagging could also provide an immediate characterization of topics or ideas discussed in the tweet [27].

Additionally, Twitter data are extremely accessible, since tweets can be pulled from the API much easier than, for example, trying to scrape comments and blogs from across the Internet to gauge public sentiment. Therefore, although Twitter data pose some challenges, they have recently been used as measures of public opinion due to their many advantages [57, 45].

2.4 Event Detection

As Twitter has matured over the past ten years, so have the algorithms used for automatic event detection [12]. There are ample situations when more efficient or reliable event detection would provide a competitive advantage [14, 12]. The presidential election, however, is not one. Campaigns, by design, control much of the national attention. Hundreds of news outlets around the country are manually identifying events. To this point, previous studies of event identification have found that events are often associated with a burst in news coverage [62]. The activity of the news media, in essence, builds a corpus to work with when analyzing tweets surrounding these events.

Chapter 3

Approach

In this chapter, the approach developed to quantitatively analyze the importance of the events which occurred during the final 40 days of the election is described. Section 3.1 details the collection of data from both Twitter and *The New York Times* to be processed and examined. Then, the news stories representing events collected from *The New York Times* is clustered in Section 3.2. Next, Section 3.3 describes a novel approach to extract representative keywords from the Twitter data obtained in Section 3.1.

3.1 Data Collection

In order to facilitate the necessary experiments, two unique datasets were required. The first, described in Section 3.1.1, is an anthology of the conversation which occurred on Twitter about either Hillary Clinton or Donald Trump from the first debate until election day. These statuses allow this

project to analyze the discussion on Twitter as a proxy for those which likely occurred around kitchen tables during the election.

The second, detailed in Section 3.1.2, contains information about every election related news article published by *The New York Times* during the same time period. This corpus, henceforth referred to as the news dataset, served as the primary ground truth of the events which occurred during the general election. The purpose was to reduce the bias of manually identifying which events occurred in fact. It is true that the choice of *The New York Times* as ground truth itself biased the resulting list of events. However, the potential that the Grey Lady would not report on an important event is lessened as the paper is commonly known as the United States' paper of record. [46]

3.1.1 Twitter Dataset

From September 25th, 2016 to November 8th, 2016 about 135.5 million tweets were collected for this project.[1] This was achieved by developing a client which took advantage of Twitter's streaming application programming interface (API).[2] The client watched for any public status update posted during this

[1]The client crashed three times during the course of collection, thus any tweets collected on those days were excluded from this project.

[2]Twitter Streaming API: `https://dev.twitter.com/streaming/overview`

time period which included any one of the following words: `hillary`, `clinton`, `hillaryclinton`, `donald`, `trump`, `realdonaldtrump`, `election`, and `debate`. It is true that these 8 words do not encompass every permutation of how those on Twitter might reference the two presidential candidates.

Yet, for the purposes of this research it was important to anthologize the least skewed portrait of the online conversations about Clinton and Trump. Therefore, the decision was made to only collect the formal names (*i.e.* their first and last names) that are used to reference either candidate. Because of how usernames are used on Twitter, as described *supra*, both candidates usernames were also included in collection. Additionally, the terms `election` and `debate` were also chosen as signal words for an important tweet to capture by the streaming client. The client only kept a status with either of these words when it appeared in conjunction with one of the other signal words, `she`, or `he`. The purpose of this is to capture tweets indirectly referencing either candidate.

After the client compiled each new status into the corpus, it was classified as referencing Clinton or referencing Trump. For the experiments using this data to be reliable indicators of the electorate's attitude, this step was crucial. Because of this, criteria for inclusion into the subset of either candidates

tweets was rigorously constructed. Damerau-Levenshtein distance was used to perform approximate string matching between a pre-determined set of terms for each canidate to determine if the tweet exclusively references one of the candidates. Through experimentation, an edit distance of 2 was chosen as the cutoff for matches. The resulting subset discussing Clinton included 35.7 million tweets and the subset discussing Trump included 51.9 million tweets.

3.1.2 News Dataset

Articles published by *The New York Times* during time period mentioned above were gathered after the election concluded using the news organization's public API. [3] For each article added to this dataset, the following information about the article was included: the headline, an abstract written by *The New York Times*, the author(s), the publication date. Only those articles which were identified by the Grey Lady's editorial team as news and part of the "politics" subsection of the paper were included.

3.2 Event Clustering

Articles in the news dataset about events are rarely one time occurrences, particularly important ones. Instead, there are stories published over time

[3] *New York Times* API: `https://developer.nytimes.com/archive_api.json`

discussing the same event. For that reason, the primary hurdle to analyze each event during the election is to track how these stories are discussed over time. Complicating things, however, is the fact that a relatively small vocabulary is used repeatedly in all of the articles in the news dataset; across the 760 articles, only 6,037 unique words are occurred. This causes a concern as two articles including the same topic words may discuss entirely different events. As an example, take the following two articles from the news dataset, both articles use the topic words such as *Email* and *F.B.I.* although they talk about two separate events:

Emails Warrant No New Action Against Hillary Clinton, F.B.I. Director Says F.B.I.'s Email Disclosure Broke a Pattern Followed Even This Summer

In order to counter this issue, a word embedding model, trained by FastText [16] on the Twitter dataset, is used. The word embeddings from this model allows us to represent and compare these articles using the average vectors [26]. However, averaging the full headline and abstract of an article begets substantial semantic loses.

It is clear that most terms in the headline and the abstract are keywords and possibly cause the average vector representing the passage to become diluted. Thus, intelligent pruning of unnecessary words in the passage could

provide a more accurate vector for clustering. To this end, a variant of term frequency - inverse document frequency (TF-IDF) is used to understand the relative importance of each word. The variant, term frequency - *probabilistic* inverse document frequency (TF-pIDF), has a unique property: terms which appear across many documents likely end with low negative weights:[4]

$$\text{TF}-p\text{IDF} = \text{TF}_t \times \log(\frac{N - n_t}{n_t})$$

A TF-pIDF score is calculated for every word in every article and only those terms under a certain threshold are used to generate the embedding of the article. The vector representation of each article is then calculated by averaging the word vectors for the selected terms which made the TF-pIDF cutoff in either the headline or abstract of the article. These average vectors are then clustered into nine groups using the k-means++ algorithm [11]. Note that we also experimented with the agglomerative clustering algorithm; nonetheless, k-means++ consistently produced more reliable results when manually inspected. Table 3.1 lists the nine clusters produced by this algorithm and the top-3 most frequent words from each cluster:

The topic labels (the topics column in Table 3.1) are manually generated

[4] N: the total # of documents, n_t: the # of documents containing the term t.

Table 3.1: Results of the event clustering on the news dataset. The topics are qualitatively analyzed and the top-3 most frequent words are selected from each cluster. The total column shows the number articles in each cluster, and the purity column shows the number of the articles discussing about the indicated topics over the total (in %).

ID	Topics	Top-3 Words	Total	Purity
C1	Debates	*debate*, *presidential*, *first*	64	95.31
C2	FBI and Emails	*emails*, *campaign*, *fbi*	81	87.65
C3	Policies	*tax*, *obama*, *bill*	107	91.59
C4	Supreme Court	*court*, *supreme*, *justices*	33	87.88
C5	Campaigning	*campaign*, *obama*, *presidential*	124	85.48
C6	Voting and Polls	*voters*, *states*, *polls*	80	47.50
C7	Congress	*senate*, *republicans*, *house*	81	88.89
C8	Trump and Women	*says*, *debate*, *women*	70	60.00
C9	Election	*presidential*, *campaign*, *women*	120	98.33

by looking at the top terms found in the cluster and verifying the findings with a review of the articles belong to that cluster. When it could be determined that multiple sub-topics are present in a cluster, the broadest topic label is chosen. For example, C8 includes, among others, articles discussing "Trump's interactions with women", "Accusations about Trump by women", and "Bill Clinton's affairs".

The quality of these clusters, purity, is also manually evaluated. The above described clustering algorithm yields the total purity score of 83.55%, which is very promising. It is clear, upon the qualitative analysis of the

clusters, that smaller clusters generally agree more precise and focused on a single event. This observation matches with the findings after calculating the gap statistic for the dataset [55]. Based upon these results, the ideal number of clusters would be 23. Yet, the results are varied when clustering into a high number of clusters. Some of the resulting clusters are more precise, but often they tend to cluster based upon one word or a phrase.

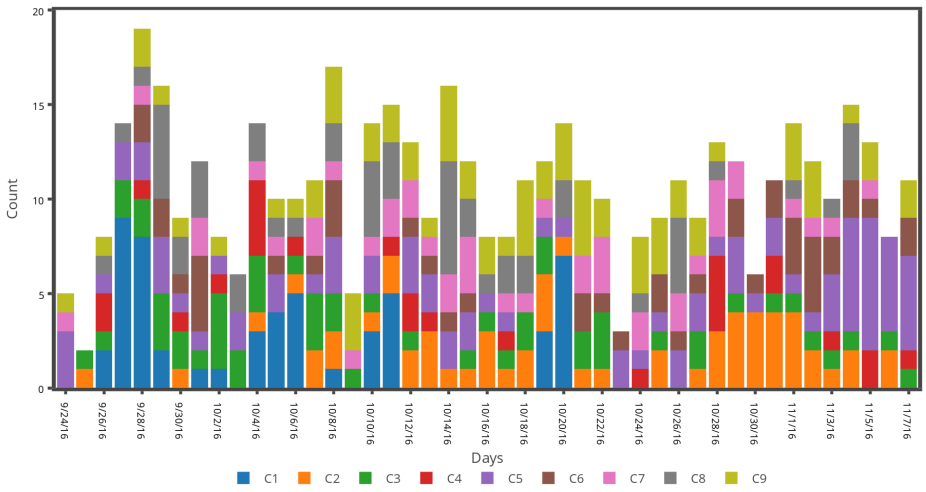


Figure 3.1: The number of articles per day for each of the clusters detailed by Table 3.1.

Figure 3.1 shows a broad overview of the topics in each cluster over the time period. For example, C1 appears only within the days proximate to the presidential and vice-presidential debates. Further, there are peaks in activity the day after each debate. This makes sense as the debates are at night and one would expect a flood of news coverage the following day. As

expected, clusters broader in scope are clearly less indicative of particular events occurrence. C7 demonstrates this; articles discussing "congress" and "congressional elections" are scattered. On the other hand, C4 about the "Supreme Court" is highly indicative of important events, which are shown in the very specific dates.

3.3 Extracting Meaning from Tweets

As demonstrated in Section 3.2, it is clear that the news clusters can provide a convenient shorthand for which events occurred on a given day. However, it would be a mistake to rely solely on the keywords which are extracted from each cluster to identify tweets referring to each event. The domain from which of these keywords have been extracted from is one, particularly at the highest levels where *The New York Times* resides, where word choices matter. Journalists vernacular reflects this fact.

Moreover, there are many events which occur that a reputable media source simply would not report on (e.g. fake news and hyper-local news) [9]. During the election the Grey Lady published, on average, 17.5 political articles a day. It is not unreasonable to assume that more than that number of events occur on any given day. Indeed, there are certainly important conversations

which would have surfaced on Twitter which would did not appear in the news dataset. Therefore, Section 3.3.2 details the approach used to extract meaningful conversation keywords from the twitter corpus each day. Before extracting the conversation keywords, Section 3.3.1 reviews the steps taken to pre-process and score each status in the Twitter corpus.

3.3.1 Processing the Tweets

Before any scores can be processed, the tweets collected in Section 3.1.1 are decoded and re-encoded in ASCII for compatibility. Then, the statuses are tokenized and divided by date and candidate referenced therein. This allows the statuses to be scored for sentiment and win/loss synsets. To perform sentiment analysis on the tweets, the convolutional neural network model with lexicon embeddings by Shin et al. is utilized, which has shown the state-of-the-art accuracy on tweets [53]; this model classifies each tweet as positive, negative or neutral. Because election rhetoric often focuses around which candidate is winning or losing, it only feels natural to also count how many tweets include words similar to "win" and "lose". This concept is inspired by the use of emoji for determining the sentiment of a tweet [44]. To better account for variations in terminology, a set of win/loss words is

extracted from the WordNet synsets [1]. Each tweet is then matched against this set to count how many tweets per candidate include win/loss terms.

3.3.2 Keyword Extraction

The traditional approach to determine the importance of words in a corpus of documents is TF-IDF [50]. However, this approach fails when applied to the present project due to the unique composition of the Twitter dataset. Specifically, many words in the Twitter dataset, like "hillary" and "trump", appear with a high relative frequency. Exasperating this problem is that most words appear once every day, which renders that inverse document frequency (IDF) as an inadequate method to meaningfully weight words.

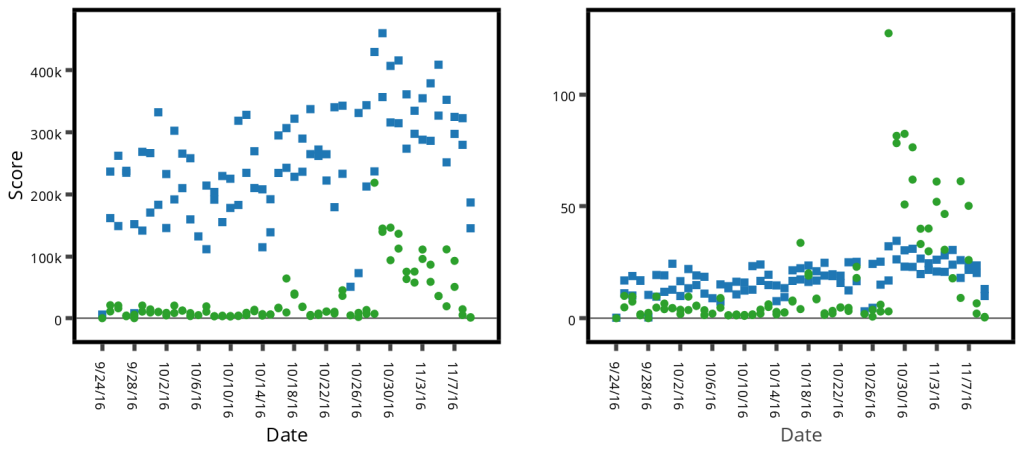


Figure 3.2: (a) TF-IDF scores for the terms "hillary" (blue squares) and "fbi" (green circles). (b) The same graph but measured with wTF.

In order to demonstrate these issues, Figure 3.2(a) shows the TF-IDF scores for "hillary" and "fbi." As is apparent, the scores for "hillary" dominates those for "fbi", even on the Oct. 28th when it was reveled that the FBI director sent a letter to Congress. There was a near unanimous consensus that this event was significant as it threw Clinton's bid for the White House into disarray. Because the goal is for words to only spike on days they are uniquely important, another measurement is needed to draw clearer trends.

this paper consider three desiderata for what a good algorithm would ensure: (1) terms ought to appear with a higher score when they are elevated relative to their average rate of usage, (2) a low average rate of usage should not hinder a terms score, and (3) a term with a low relative usage frequency and high proportional rate of usage should not be given a higher score than a term with a exponentially higher specific frequency of usage, but similar proportional rate of usage.

Therefore, this paper proposes a new weighting scheme to ascertain important keywords. This scheme, weighted logarithmic term frequency (wTF), works as it normalizes the count of each term t by $a\text{TF}_t$ the average frequency of the term t across all D days. wTF is calculated for each term t on each

day d as follows:

$$w\mathrm{TF}_{t,d} = \log_2 \frac{\mathrm{TF}_{t,d}}{a\mathrm{TF}_t} \qquad \text{where} \qquad a\mathrm{TF}_t = \frac{\sum_{n=1}^{D} \mathrm{TF}_{t,d_n}}{D}$$

Figure 3.2(b) demonstrates the advantage of wTF to calculate the keyword scores over TF-IDF; "hillary" shows no clear trend whereas "fbi" shows clear spikes on certain days. This new scoring mechanism allows useful keywords to be selected from tweets each day. However, an additional component is required before the selected keywords are meaningful. Because wTF treats the growth as a scalar, a term that is used, on average, 10 times a day and is mentioned 10^3 times suddenly would have a similar wTF as a term which, on average, is used 10^3 times a day seeing a spike to 10^5 times.

With this in mind, a filter is required before running wTF to remove words with a low count. This hyperparameter allows one to tune the algorithm in terms of sensitivity to outside noise. The tweets are divided into separate documents by date, and the top-10 keywords based upon their wTF scores are selected from each day. Stop words, variations on the candidates names, and any word which appears more than 80% and less than the max term each day are excluded from the keyword selection. Inclusion of this filter resolves the third requirement laid out above. Qualitative review of the chosen

conversation keywords indicates that they are meaningful and aligned with the events.

Chapter 4

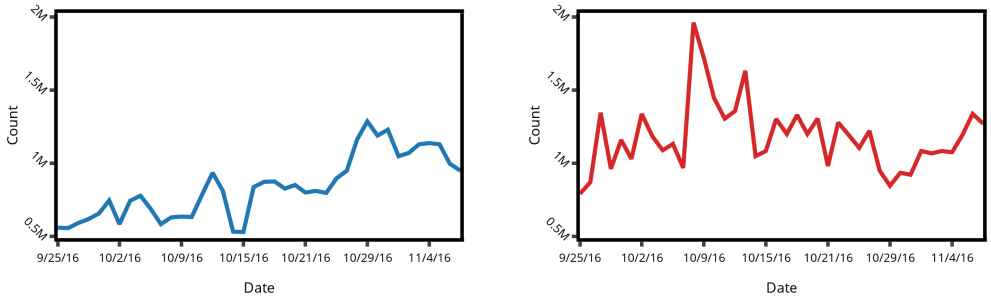
Experiments

4.1 Exploring Top-Level Trends

Before applying the methods developed in chapter 3 to analyze events, it is important to first identify the top level trends for each candidate. To do this, each of the scores described in section 3.3.1 will be evaluated for trends. Each unique score, when taken in aggregate for a group of tweets, can be thought of as a way of describing an aspect of the conversation on Twitter. This is why analyzing and comparing each of the scores for both Clinton and Trump can provide meaningful high level trends, as previous studies have determined. Thus, this section will explore the volume of activity surrounding each candidate (section 4.1.1), the sentiment surrounding each candidate (section 4.1.2), and the quantity of Win/Loss terms used in conversation on Twitter about each candidate (section 4.1.3).

4.1.1 Volume of Tweets

Figure 4.1 illustrates the raw number of tweets each day regarding either Clinton or Trump. What is immediately obvious when comparing Figure 4.1(a) and Figure 4.1(b) is that people, on average, talked more about Trump than Clinton on Twitter. However, there is an interesting trend in volume of conversation on Twitter referencing Clinton. Though, on average, Clinton is discussed less than Trump on Twitter, there is a noticeable positive linear increase over time; when a linear regression model is fit to Clinton's raw counts over time. The R^2 of this line of best fit is 0.69.

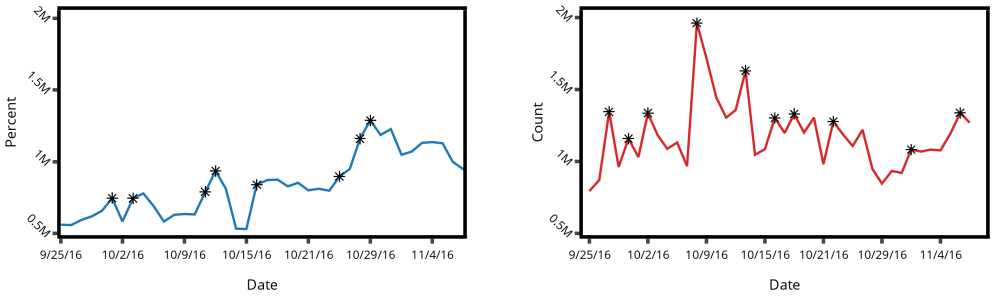


(a) Clinton (b) Trump

Figure 4.1: Clinton and Trump's Raw Counts

Exploring this trend further, not once before the 26th of October did the number of Tweets referencing Clinton surpass the number referencing Trump. Then, Clinton's volume exceeded Trump's on seven of the twelve days of the campaign. In fact Clinton's increased Twitter activity in the last twelve

days of the election is enough to increase the average count by about 16%. Moreover, the standard deviation increased by 71.75%.



(a) Clinton (b) Trump

Figure 4.2: Clinton and Trump's Raw Counts with Rapid Change Markers

This leaves open the question: Is the volume of tweets or the change in volume of tweets more significant to event importance? In order to address this question, it appears prudent to explore instances of rapid tweet volume change. Figure 4.2 includes the same volume information as Figure 4.1 except that a black marker is included if the increase in volume about the candidate exceeded 10% of the candidate's average volume. On eight occasions the the conversation around Clinton merited a marker. While the conversation around Trump earned ten markers.

A surprising observation is that the markers do not always appear on local maxima as would be expected. The first occasion where a local maximum fails to earn a marker is on October 4th for Clinton. Evaluating the events

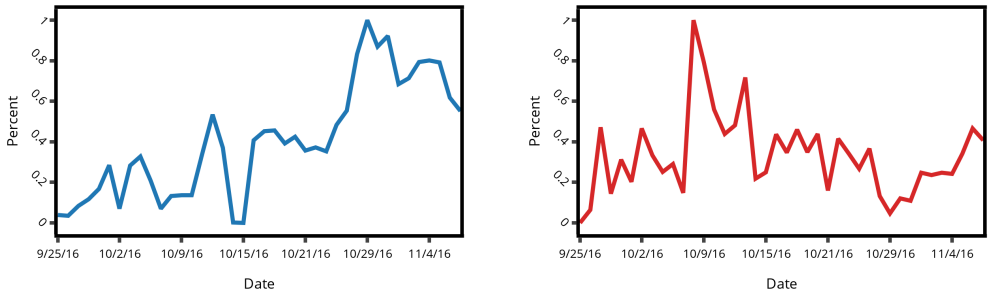
which triggered the rise in tweets on October 3rd and October 4th is easy when using the keywords extracted in section 3.3.2. The top-6 conversation keywords on October 3rd are as follows: `lebron`, `drone`, `strike`, `endorses`, `julian`, `assange`. Looking at the news dataset it is clear that `lebron` and `endorses` stems from this event: *LeBron James, Calling for Hope and Unity, Endorses Hillary Clinton.*

However, no other keyword from Twitter matched to an article published by *The New York Times* on October 3rd. As discussed in section 3.3, this is what drives the need for conversation keywords. Turning to a Google search for each keyword along with the date and candidates name quickly validates this suspicion of an institutional bias in coverage by the Grey Lady. After examining the first three pages of search results,[1] it became clear that `drone`, `strike`,`julian`, and `assange` all referred to the same event.

Every item on the first three pages of the search included one or more of these terms. Moreover on the first page of results, seven of the ten results were dated October 3rd. This is the headline from an article published by RT found on the first page of the search results: *Hillary Clinton considered drone attack on Julian Assange.* Indeed, the conversation keywords extracted the

[1]The following phrase was used for this evaluation: *3 October 2016 clinton drone.*

next day were: `guccifer`, `drone`, `obamacare`, `assange`, `julian`, `guy`. Five of these six terms are related to the event which surfaced on October 3rd. In addition to this, three keywords are identical to those on the previous day.



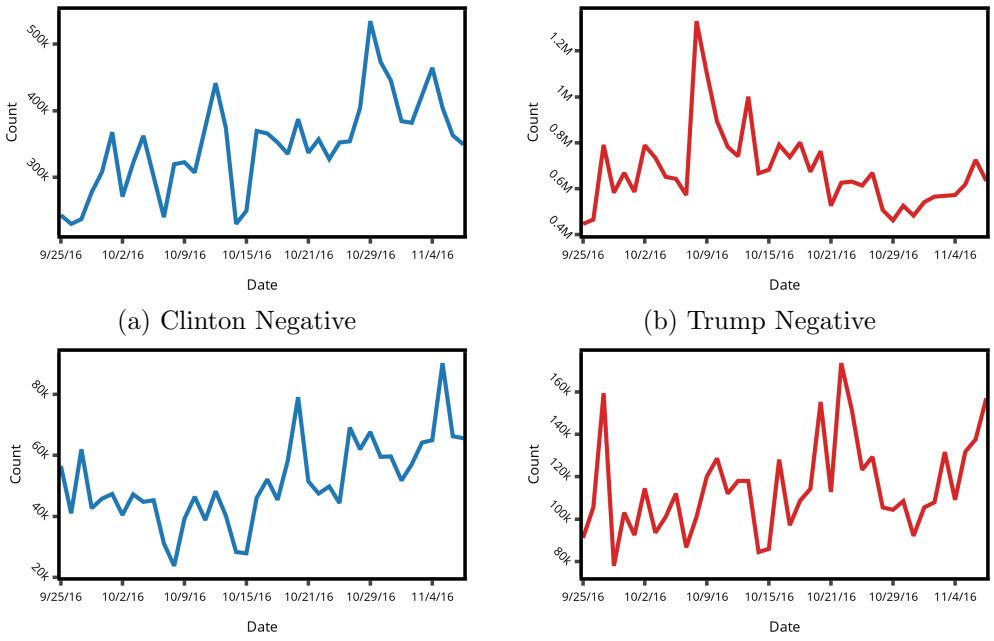
(a) Clinton (b) Trump

Figure 4.3: Clinton and Trump's Normalized Raw Counts

Similar results are seen when the second and third instances of this pattern are examined. Going back to the open question, this result suggests that the change in quantity of tweets about a candidate is in fact an important characteristic. Now, turn to considering the second trait: sheer count alone. In order to evaluate this notion, consider each of the candidates top three days by volume. For this, the normalized tweet count for each candidate, as shown in Figure 4.3, is useful. Clinton's top three days are October 29th, October 31th, and October 30th; Trump's top three days are October 8th, October 9th, and October 13th.

4.1.2 Sentiment of Tweets

Figure 4.4(a) and Figure 4.4(b) shows the raw volume of Clinton and Trump's negative tweets each day. Notice that there is a conspicuous closeness between either of the candidate's sentiment volume and the counts reported in Figure 4.1. The correlation between negative sentiment and number of Tweets received by Clinton and Trump is 91.3% and 96.1% respectively.



(a) Clinton Negative

(b) Trump Negative

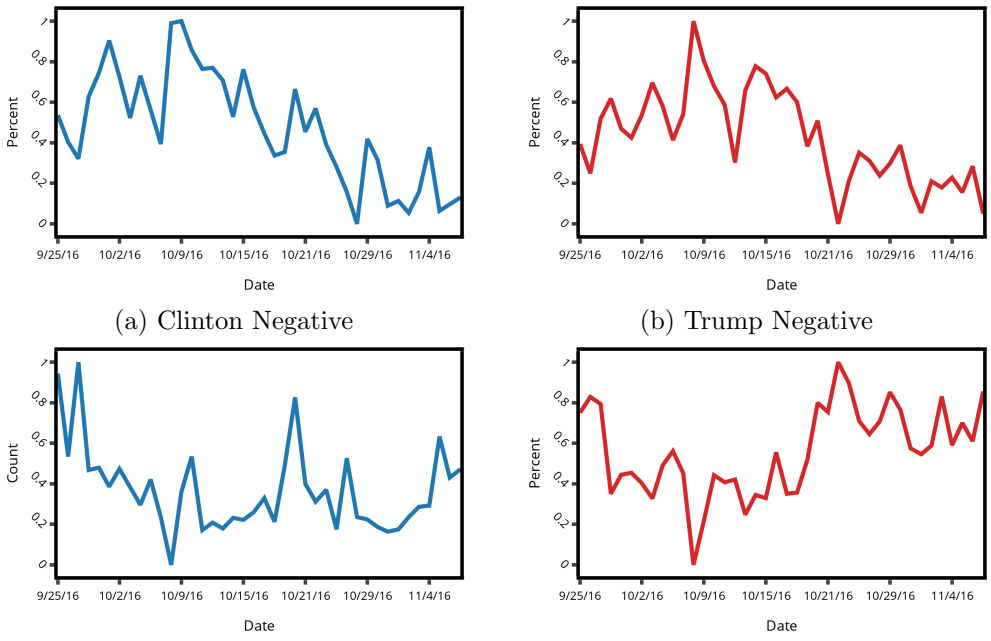
(c) Clinton Positive

(d) Trump Positive

Figure 4.4: The Sentiment of Tweets about Clinton and Trump

A similar correlation is not seen in the volume of positive statuses. This suggests that as the candidate is spoken about more on Twitter, the increased

conversation surrounding them is often far more negative than it is positive. It seems clear that the both negative sentiment and tweet volume are good approximations for important events when basing this distinction exclusively on keywords extracted from the Twitter dataset. If analysis is performed using the raw counts alone, however, it might lead to a possible conflation between volume and sentiment on a given day. Thus, analysis moving forward will be done on both the raw volume results and the sentiment data that has been normalized first. Figure 4.5 showcases both candidates normalized positive and negative sentiment.



(a) Clinton Negative

(b) Trump Negative

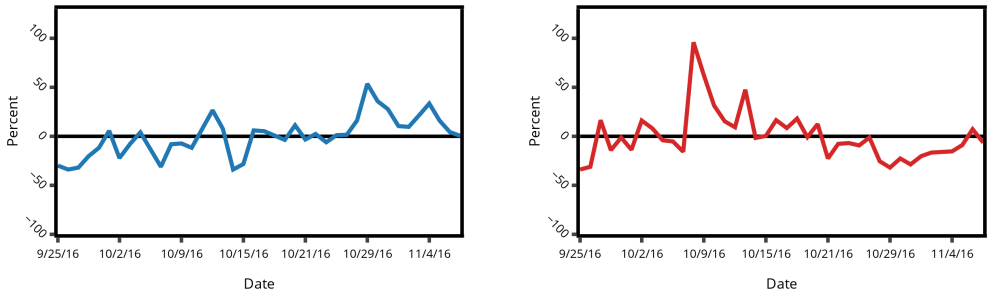
(c) Clinton Positive

(d) Trump Positive

Figure 4.5: The Normalized Sentiment of Tweets about Clinton and Trump

The values in Figures 4.5 (a) and (b) are calculated by dividing the number of tweets with a negative sentiment by the total number of tweets that day. Each of these percent value are then normalized by subtracting the minimum value across all days from and dividing by the max across all days minus the minimum value again. This gives figures which are an indication how negative the overall conversation compared to all other days.

Interestingly, Figure 4.5 indicates different trends than Figure 4.4 for Clinton but similar trends for Trump. Notice that Figure 4.4(a) shows that the most negative events are towards the end of the campaign.



(a) Clinton (b) Trump

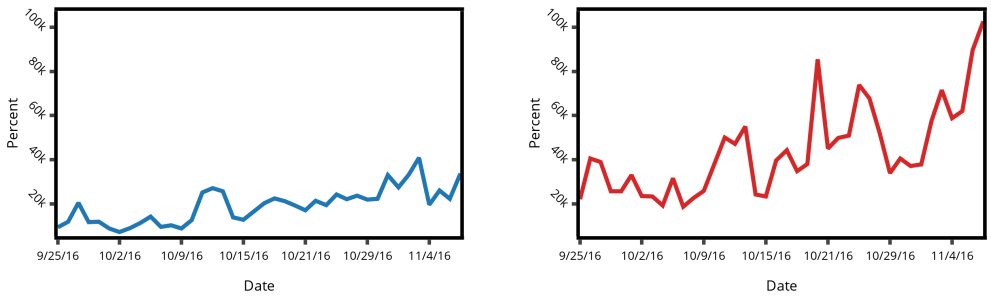
Figure 4.6: The Proportional Negative Sentiment of Tweets about Clinton and Trump

Yet, Figure 4.5(a) would suggest that those events were not the most negative. This is because Figure 4.5 is measuring the *proportion* of negative tweets each day. Though the total number of negative tweets about Clinton reached its peak towards the end of the campaign, there were also a far higher number

of neutral tweets then. Figure 4.6 helps to better visualize this trend by displaying the number of negative tweets over the average number of negative tweets per day.

4.1.3 Win/Loss Synsets of Tweets

Often election rhetoric and political discourse focuses on which candidate is winning or losing. For this reason, it only felt natural to also count how many tweets included words similar to "win" and "lose." This concept was inspired by the use of emoji in determine sentiment of a tweet [44]. To better account for variations in terminology, a list of win/loss words was generated based upon the WordNet synsets of "win" and "lose" [1]. Each tweet was matched against this list to count how many tweets per candidate included a win/loss term. Figure 4.7 visualizes the number of times win or any win synset was used in a tweet about Clinton or Trump.



(a) Clinton (b) Trump

Figure 4.7: Usage of Win Term per Day for each Canadate

To better understand to usefulness of win synsets, consider two peaks from Figure 4.7(a): 27 September, 12 October. The first peak, September 27th, begins its increase the pervious day, September 26th. This was the date of the first debate in which Clinton was unanimously declared the winner by critics. The second peak, October 12th, is a local peak following several days of consecutive increases. The increase began on October 10th, the day after the second presidential debate. Clinton, again, was declared the winner of this debate as well. What is most interesting about this increase is that there is a sharp fall-off on October 14. This decrease could be a sign that the news of Clinton's debate win has been replaced by another story in discussion on Twitter. It is clear that, after examining the trends for both Trump and Clinton, win and loss terms require context to be useful in any analysis.

4.2 Event Scoring

To better understand the usefulness of the news clusters developed in section 3.2, consider Figure 4.8. This figure allows us to explore the relationship between the keywords extracted from the news corpus and the trends established about events so far. As described above, the tweets used in Figure 4.8 are extracted from the Twitter dataset based solely upon the extracted article

keywords.

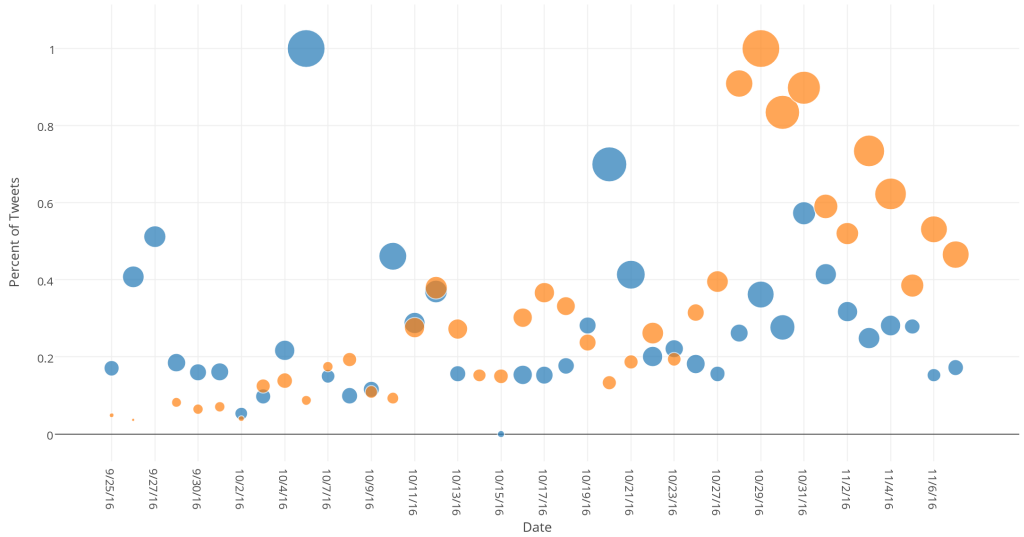


Figure 4.8: The percentage of tweets identified as a member of the clusters C1 (blue) and C2 (orange) from Table 3.1. The size of each point is proportional to the normalized count of tweets that are classified as negative.

Both the Debate cluster (C1) and the FBI cluster (C2) show promising results. Starting with the C1, the most interesting high level observation is that the points spike when there is a debate. This affirms that the keywords extracted in Section 3.2 are meaningful. Additionally, the same trend which is detected when looking at Figure 4.8 occurs here: negative sentiment increases as the tweet count increases. These observations are not exclusive to C1; C2 also shares these traits. Instead of peaking at debates, however, it peaks for the first time on Oct. 12th and then again on Oct. 29th. The events and

keywords on, both of these dates are discussed in the analysis of Figure 4.3.

Again, the tweets associated with C2 are the most negative when the count is the highest. Finally, compare the article counts for both C1 and C2 in Figure 3.1 with the peaks in Figure 4.3. For both clusters, when the number of published article peaks, so do their respective tweet counts. This is rather intuitive. When an event occurs and news spreads, it is expected that there would be more discussion around the topic. Moreover, when multiple stories per day are published within the same cluster, it might signal that a major event occurred that needs ample coverage.

Yet, an assumption underlying this discussion has not, so far, been addressed: the rise in negative sentiment around a major event is a reliable indicator of how individuals are responding to the event. Admittedly, it is not possible to establish this point within the parameters of the current study. However, there are a few indications that this assumption is warranted, the most convincing of which is that the increases in negative sentiment fall at points of intuitive voter displeasure. When discussion keywords which indicate a scandal that plagued one of the candidates surface, there is generally a corresponding rise in negative sentiment. Moreover, the jumps in sentiment are compartmentalized among the article clusters. When the tweets matched

with C1's keywords indicate a jump in negative sentiment on Oct. 5th, C2's negative sentiment volume did not follow.

Chapter 5

Discussion

Over the course of the 2016 election, a common question posed by the press went something like this: "will event X impact candidate Y's prospects for victory?" As an illustration, *The Guardian* ran a story weeks before the election titled: "Will Hillary Clinton lose the election because of the FBI email investigation." A Google search[1] for this exact phrase returned over 1.5 million results. Each and every major event prompted waves of similar questions from pundits.

Unfortunately, these type of questions have no quantitatively verifiable answer today. What is known, however, is that Clinton lost the election but won the popular vote. Because of this, the events and scandals involving Clinton are more important to investigate than those surrounding Trump. Therefore, this section will discuss the major events that followed the Democratic standard bearer during the 2016 election. Understanding the prominence and

[1]The search referenced by this proposal was performed on March 19, 2017.

importance of these threads is critical to effectively analyzing the outcome of the election. This paper does not purport to provided a wholistic methodology through this exploration, but instead show the feasibility and importance of future analysis.

5.1 Presidential Election Debates

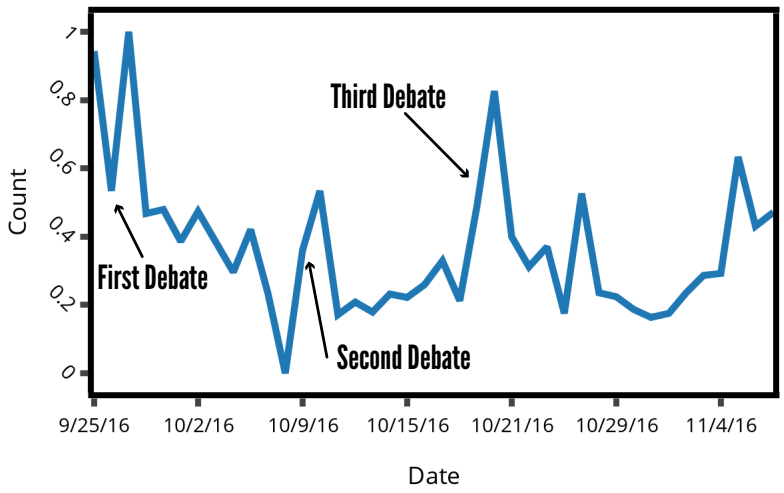


Figure 5.1: Clinton's Normalized Positive Sentiment Annotated with Debate Dates

To begin this analysis of the impact of the debates on Clinton, look first at figure 5.1, which shows the percent positive sentiment on Twitter from figure 4.5(c) with annotations pointing to the dates of the debates. Clinton receives bumps in positive sentiment as a result of each debate. A review of

the news corpus indicates that this aligns with her performance in the debate. As an example, the following headlines are representative of the coverage following the first debate:

Debate Takeaways: Hillary Clinton Digs In and Prevails

Suburban Women Find Little to Like in Donald Trumps Debate Performance

After a Disappointing Debate, Donald Trump Goes on the Attack

Commentators Give Hillary Clinton Edge in Debate

The debates account for three of five of Clinton's positive sentiment peaks during the collection period. Moreover, the first debate corresponded to an inflection point in her lead over trump in the popular vote polling.

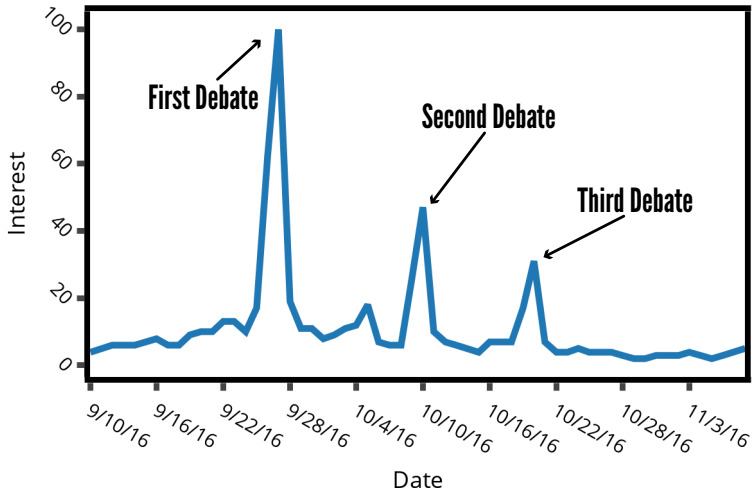


Figure 5.2: The Public's interest in the Presidential Debates

Though each of the debates causes Clinton's positive sentiment to spike, the lasting impact of each is in question. Conversation keywords from section 3.3.2 surrounding the the first debate appear in the top-6 terms for three days following the debate. However, for each of the following debates, related conversation keywords only appear in the top-6 terms the day immediately following the debate. This, along with the fact that the first event corresponds with the global maximum for Clinton's positive sentiment, suggests that the first debate was far more important to the election than either of the other debates. Figure 5.2 provides a second source of information, specifically results retrieved from Google Trend, to verify this.

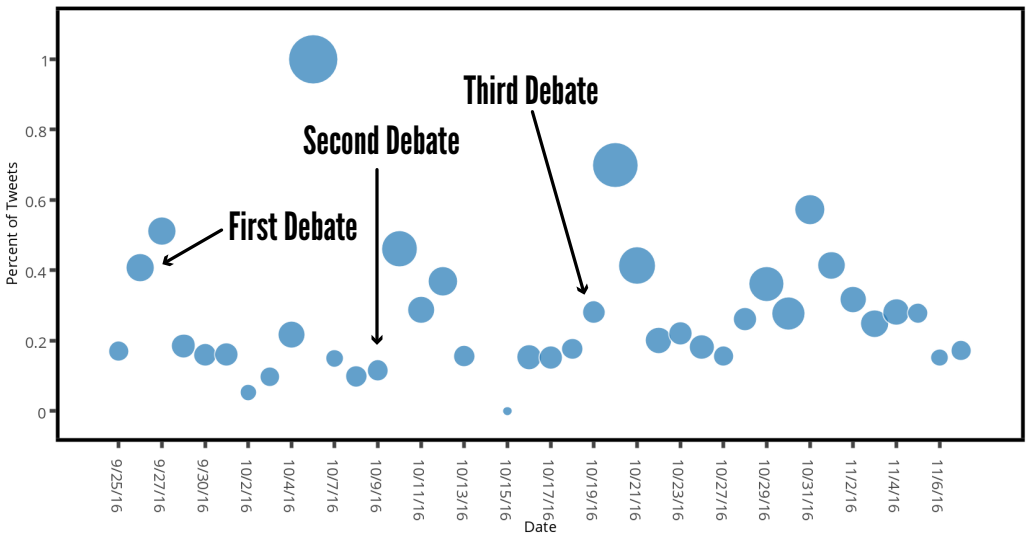


Figure 5.3: Information Regarding Tweets Linked to C_1, the Debate Cluster

Compare this result to data collected using the debate news cluster, C_1. Figure 5.3 displays the data from figure 4.8 alongside annotation of the debate dates. This figure is consistent with the trends in figure 5.1 and figure 5.2. Notice that, again, the first debate is discussed at a higher volume on Twitter than the second or third debates. Furthermore, all three debate data points, and the day following each, have a general low negative sentiment. This is except for the one outlier in these results: the vice-presidential debate on October 4th. One potential explanation for why this is the case is that Clinton's running mate, Tim Kaine, is not thought to have won his debate.

5.2 Clinton's Email Scandals

During the presidential election campaign, there were two primary email related scandals: Clinton's F.B.I. Investigation and the dumps of emails by Wikileaks. Both of these scandals were clustered into the same news event cluster, C_2.

Wikileaks, as the first of these, is a constant source of conversation keywords for Clinton. In fact, more related terms to Wikileaks appeared as keywords for Clinton than any other event. In section 4.1.2 October 3rd and 4th were analyzed, and it was determined that the volume of tweets on both of

these days is likely to be due to Wikileaks related events. Figure 5.4 visualizes the normalized negative sentiment discussed in section 4.1.2 alongside markers on which one of the conversation keywords was related to Wikileaks; On more than 50% of the days where tweets were collected, a Wikileaks related term was one of the top-6 keywords.

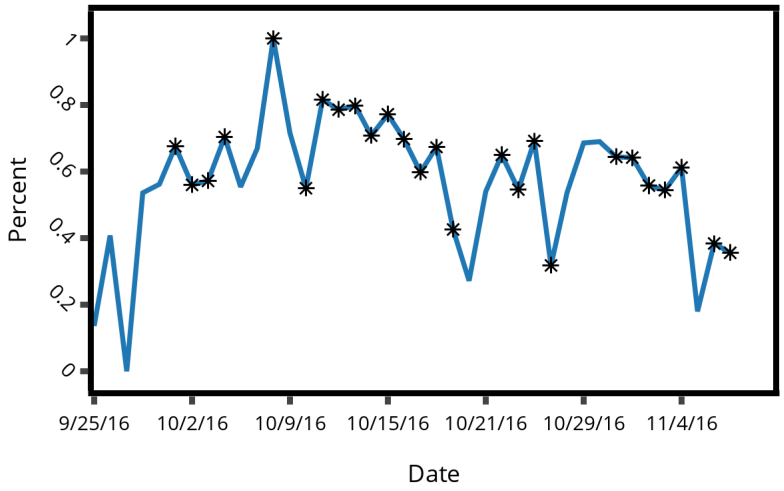


Figure 5.4: Normalized Negative Sentiment with Markers

Based upon a review of *The New York Times*, only three articles referred to Wikileaks by name, and only a handful more reported on the events connected to the related keywords. This, again, is a result of The Times' reporting bias. To be clear, many of the events indicated by the Wikileaks related conversation keywords were, essentially, fake news. As such, these events were, without exception, negative for Clinton.

Identifying the leaks which were most harmful to Clinton may be attempted using the data collected from Twitter. One option to accomplish this is to look to days where the negative sentiment surrounding Clinton reaches a peak and a Wikileaks related keyword is extracted from the twitter corpus. The challenge to this methodology is the potential for false positives.

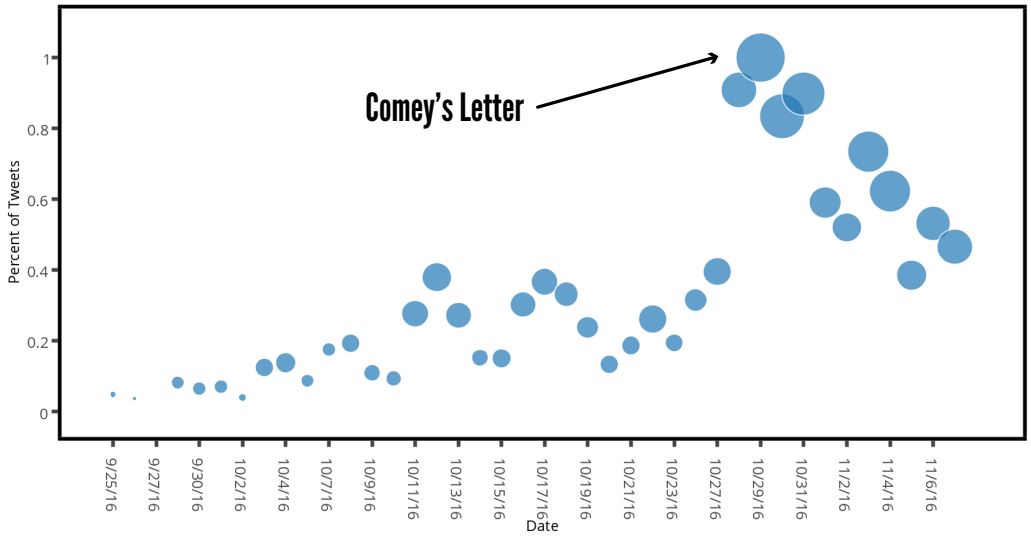


Figure 5.5: Information Regarding Tweets Linked to C_2

However, the scores for the subset of Tweets which relates to C_2 provides an indication of which mattered most. Leading up to Comey's letter sent to Congress the morning of October 28th, there were three distinct peaks: October 12th, October 17th, and October 27th. Interestingly enough, the Wikileaks related events on these three days were veridical news. Moreover,

these stories were wildly reported on by mainstream media outlets instead of the more fringe sites which comment on the fake news Wikileaks dumps. This is likely to have exposed more discerning individuals to the events.

The largest spike in figure 5.5 occurred on October 28th. As already mentioned, this coincided with Jim Comey's letter to Congress regarding the Clinton email investigation. Based upon the results in figure 5.5, this was likely the most important event in this cluster. Referencing figure 4.1, it is also clear that this event propagated the most conversation about Clinton –another indicator to the significance of the event.

Chapter 6

Conclusion

There is a competitive advantage in being able to distinguish important events. For example, a campaign might use such information to guide their communications decisions, particularly when responding to fake news online. Using Twitter to access all available data requires very little overhead and provides valuable insight. Something just as valuable as understanding where the conversation is currently, is how the conversation changes to stimuli. Using the analysis techniques demonstrated in this paper, it becomes a trivial endeavor to measure this.

This paper has two primary purposes: to explore the utility of Twitter in analyzing events during an election, and to determine which events during the 2016 election were the most important. To this end, sentiment analysis, win/loss counts and tweet volume were used to determine a signal in the over 135.5 million tweets which were collected. One finding while exploring the

utility of Twitter is that having a variety of datasets to better understand the events which occur during the election is key. In this research, only *The New York Times* was used. This limited the predictive power of the event clusters as they were not wholly representative of the conversations in the Twitter corpus. The keyword identification performed by this paper permits the extraction of useful conversation topics from tweets given any day using the weighted logarithmic term frequency.

Bibliography

[1] *About WordNet*. Princeton University, http://wordnet.princeton.edu, 2010.

[2] Alan Abramowitz. The time for change model and the 2000 election. *American Politics Research*, 29(3):279–282, 2001. doi: 10.1177/1532673X01293004. URL http://dx.doi.org/10.1177/1532673X01293004.

[3] Alan Abramowitz. Forecasting in a polarized era: The time for change model and the 2012 presidential election. *PS: Political Science and Politics*, 45(4):618–619, 2012.

[4] Alan I. Abramowitz. An improved model for predicting presidential election outcomes. *PS: Political Science and Politics*, 21(4):843–847, 1988. ISSN 10490965, 15375935. URL http://www.jstor.org/stable/420023.

[5] Alan I. Abramowitz. Bill and al's excellent adventure. *American Politics Quarterly*, 24(4):434–442, 1996. doi: 10.1177/1532673X9602400403. URL `http://journals.sagepub.com/doi/abs/10.1177/1532673X9602400403`.

[6] Alan I Abramowitz. When good forecasts go bad: The time-for-change model and the 2004 presidential election. *PS: Political Science & Politics*, 37(4):745–746, 2004.

[7] Alan I Abramowitz. Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science and Politics*, 41(4):691–695, 2008.

[8] Lorien C Abroms and R Craig Lefebvre. Obama's wired campaign: Lessons for public health communication. *Deighton & Kornfeld Journal of Health Communication*, 14:415–423, 2008. ISSN 1081-0730. doi: 10.1080/10810730903033000.

[9] Hunt Allcott. Social media and fake news in the 2016 election. 2017.

[10] Kevin Arceneaux. Do campaigns help voters learn? a cross-national analysis. *British Journal of Political Science*, 36(1):159173, 2006. doi: 10.1017/S0007123406000081.

[11] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[12] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015. ISSN 1467-8640. doi: 10.1111/coin.12017. URL `http://dx.doi.org/10.1111/coin.12017`.

[13] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics, 2013.

[14] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11(2011):438–441, 2011.

[15] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *International Conference on Discovery Science*, pages 1–15, 2010.

[16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[17] Frank Bruni. Our insane addiction to polls. `https://www.nytimes.com/2016/01/24/opinion/campaign-stops/our-insane-addiction-to-polls.html`, January 2016. Accessed: 2017-07-05.

[18] Francesco Buccafurri, Gianluca Lax, Serena Nicolazzo, and Antonino Nocera. Comparing twitter and facebook user behavior: Privacy and other aspects. *Computers in Human Behavior*, 52:87–95, 2015.

[19] James E. Campbell. Forecasting the presidential vote in the states. *American Journal of Political Science*, 36(2):386–407, 1992. ISSN 00925853, 15405907. URL `http://www.jstor.org/stable/2111483`.

[20] James E Campbell. The 2008 campaign and the forecasts derailed. *PS: Political Science and Politics*, 42(1):19–20, 2009.

[21] James E Campbell. Recap: Forecasting the 2012 election. *PS: Political Science and Politics*, 46(1):37, 2013.

[22] James E Campbell. The trial-heat and seats-in-trouble forecasts of the 2016 presidential and congressional elections. *PS: Political Science and Politics*, 49(4):664–668, 2016.

[23] James E. Campbell and Kenneth A. Wink. Trial-heat forecasts of the presidential vote. *American Politics Research*, 18(3):251–269, 1990. ISSN 1532-673X. doi: 10.1177/1532673X9001800301.

[24] James E. Campbell, Lynna L. Cherry, and Kenneth A. Wink. The convention bump. *American Politics Quarterly*, 20(3):287–307, 1992. doi: 10.1177/1532673X9202000302. URL `http://journals.sagepub.com/doi/abs/10.1177/1532673X9202000302`.

[25] James E. Campbell, Helmut Norpoth, Alan I. Abramowitz, Michael S. Lewis-Beck, Charles Tien, James E. Campbell, Robert S. Erikson, Christopher Wlezien, Brad Lockerbie, Thomas M. Holbrook, and et al. A recap of the 2016 election forecasts. *PS: Political Science and Politics*, 50(2):331338, 2017. doi: 10.1017/S1049096516002766.

[26] Michal Campr and Karel Ježek. Comparing semantic models for evaluating automatic document summarization. In *International Conference on Text, Speech, and Dialogue*, pages 252–260. Springer, 2015.

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

[28] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics, 2010.

[29] Andrew Gelman and Gary King. Why are american presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23(04):409–451, 1993.

[30] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[31] Louis Harris. Election polling and research. *The Public Opinion Quarterly*,

21(1):108–116, 1957. ISSN 0033362X, 15375331. URL http://www.jstor.org/stable/2746793.

[32] Gavin Hewitt. Trump faithful undeterred by polls and scandals. http://www.bbc.com/news/election-us-2016-37762510, October 2016. Accessed: 2017-07-03.

[33] D. Sunshine Hillygus. The evolution of election polling in the united states. *Public Opinion Quarterly*, 75(5):962–981, December 2011.

[34] Lawrence R. Jacobs and Robert Y. Shapiro. Issues, candidate image, and priming: The use of private polls in kennedy's 1960 presidential campaign. *The American Political Science Review*, 88(3):527–540, 1994. ISSN 00030554, 15375943. URL http://www.jstor.org/stable/2944793.

[35] Lawrence R. Jacobs and Robert Y. Shapiro. The rise of presidential polling: The nixon white house in historical perspective. *The Public Opinion Quarterly*, 59(2):163–195, 1995. ISSN 0033362X, 15375331. URL http://www.jstor.org/stable/2749700.

[36] Lawrence R. Jacobs and Robert Y. Shapiro. Polling politics, media, and election campaigns. *Public Opinion Quarterly*, 69(5):635, 2005. doi: 10.1093/poq/nfi068. URL +http://dx.doi.org/10.1093/poq/nfi068.

[37] Robert King and Martin Schnitzer. Contemporary use of private political polling. *The Public Opinion Quarterly*, 32(3):431–436, 1968. ISSN 0033362X, 15375331. URL http://www.jstor.org/stable/2747647.

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

[39] Richard R Lau. An analysis of the accuracy of trial heat polls during the 1992 presidential election. *Public Opinion Quarterly*, 58(1):2–20, 1994.

[40] Lukas Malec, Antonin Pavlicek, and Ladislav Luc. Systematic analysus of social media in 2016 us elections. *System approaches 16*, page 72, 2016.

[41] Patrick R. Miller and Pamela Johnston Conover. Red and blue states of mind. *Political Research Quarterly*, 68(2):225–239, 2015. doi: 10.1177/1065912915577208. URL http://dx.doi.org/10.1177/1065912915577208.

[42] Helmut Norpoth. Of time and candidates. *American Politics Quarterly*,

24(4):443–467, 1996. doi: 10.1177/1532673X9602400404. URL http://journals.sagepub.com/doi/abs/10.1177/1532673X9602400404.

[43] Helmut Norpoth and Michael Bednarczuk. History and primary: The obama re-election. *APSA 2012 Annual Meeting Paper*, 2012.

[44] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

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

[46] The Editors of Encyclopædia Britannica. *The New York Times*. Encyclopædia Britannica, July 2017.

[47] Ozer Ozdikis, Pinar Senkul, and Halit Oguztuzun. Semantic expansion of tweet contents for enhanced event detection in twitter. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 20–24. IEEE Computer Society, 2012.

[48] Chang Sup Park. Does twitter motivate involvement in politics? tweeting,

opinion leadership, and political engagement. *Computers in Human Behavior*, 29(4):1641–1648, 2013.

[49] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. Can twitter replace newswire for breaking news? In *ICWSM*, 2013.

[50] Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[51] Ed Rogers. The clinton scandals will never stop. `https://www.washingtonpost.com/blogs/post-partisan/wp/2016/11/03/the-clinton-scandals-will-never-stop/`, November 2016. Accessed: 2017-07-02.

[52] Daron R. Shaw. A study of presidential campaign event effects from 1952 to 1992. *The Journal of Politics*, 61(2):387–422, 1999. ISSN 00223816, 14682508. URL `http://www.jstor.org/stable/2647509`.

[53] Bonggun Shin, Falgun H. Chokshi, Timothy Lee, and Jinho D. Choi. Classification of Radiology Reports Using Neural Attention Models. Technical report, Anchorage, AK, 2017. URL `http://www.ijcnn.org`.

[54] Stefan Stieglitz and Linh Dang-Xuan. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Science*, pages 3500–3509. IEEE, 2012.

[55] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411–423, 2001. ISSN 1467-9868. doi: 10.1111/1467-9868.00293. URL http://dx.doi.org/10.1111/1467-9868.00293.

[56] Michael W Traugott. The accuracy of the national pre-election polls in the 2004 presidential election. *Public Opinion Quarterly*, 69(5):642–654, 2005.

[57] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Election forecasts with twitter. *Social Science Computer Review*, 29(4):402–418, 2011. doi: 10.1177/0894439310386557. URL http://dx.doi.org/10.1177/0894439310386557.

[58] Darrell M. West. Polling effects in election campaigns. *Political Behavior*,

13(2):151–163, 1991. ISSN 01909320, 15736687. URL http://www.jstor.org/stable/586039.

[59] Christopher Wlezien and Robert S Erikson. The timeline of presidential election campaigns. *Journal of Politics*, 64(4):969–993, 2002.

[60] Magdalena E Wojcieszak and Diana C Mutz. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 59(1):40–56, 2009.

[61] Zhiheng Xu and Qing Yang. Analyzing user retweet behavior on twitter. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 46–50, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-0-7695-4799-2. doi: 10.1109/ASONAM.2012.18. URL http://dx.doi.org/10.1109/ASONAM.2012.18.

[62] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 28–36, New York, NY, USA,

1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.290953. URL http://doi.acm.org/10.1145/290941.290953.