USING SENTIMENT ANALYSIS ON REDDIT DATA TO PREDICT PRESIDENTIAL ELECTIONS AND POLLING

TURING HONORS THESIS

Ashvin Govil
Department of Computer Science
University of Texas at Austin
ashvion@utexas.edu

May 7, 2019

ABSTRACT

The perceived shortcomings of public opinion surveys in predicting the 2016 Presidential Election has boosted the search for new methods to measure public sentiment about presidential candidates. Using millions of comments taken from a political subforum on the social media website Reddit, we were able to construct a pair of bag-of-words sentiment models that measured the cumulative daily score of all comments containing one of the candidate's names. With a short delay to account for the time it takes to conduct polls, the time series data for each candidate has an inverse correlation with their polling averages in the 2016 Presidential Election (meaning that more discussion correlated with worse performance in polls). These results show the potential for using Reddit as a source for real-time political prediction that relies less on public polling. The correlation also suggests that online discussion around media events and news stories were almost always negative and played a fundamental role in determining shifts in public opinion during the 2016 election cycle. The data has several distinct advantages over polling, such as a much faster turnaround time, near zero operation costs, and drawing from a much larger population. Similar analysis of Reddit data using different keywords also shows potential for determining how certain scandals or news events (such as the "Comey letter") affected the polls and outcome of the election.

1 Introduction

The results of the 2016 US Presidential Election showed the world the shortcomings of relying on public polling to predict political outcomes. The polls in the swing states of Michigan, Wisconsin, and Pennsylvania proved inaccurate by large margins, leading to pundits and newspapers to inaccurately believe that Donald Trump had very little chance to become president. The inaccuracies largely stemmed from systemic biases in polling sampling methods [6]. In addition to inaccuracy, public polling by its nature takes large amounts of time and money to conduct. The last cycle combined with the perpetual drawbacks of polling show the need for new predictive methods in future elections.

Due to the rapidly increasing importance of the internet in politics, social media data holds potential to be used as a predictive tool. Websites like Facebook, Twitter, and Reddit are emerging as sources of political news for millions of Americans. People interact with these websites through comments and likes, or reactions, providing valuable insight for how people react to news stories in real-time. Over the last several years, using social media to make predictions in the real world has found many valuable uses [12]. Twitter has been used successfully to predict elections in the past [10], but until now, Reddit has remained elusive for political prediction.

Social media has also become an important facet of modern presidential campaigns. Both the Clinton and Trump campaigns extensively campaigned through every major social media website, including Reddit [8]. However, the media has largely focused its coverage on the role of Twitter and Facebook on the election, despite the fact that Reddit is the 6th most visited website in the US[2].

Some important challenges to extracting useful information from social media include choosing the right source, downloading a representative sample of data, and accounting for the demographic biases of various social media websites. Using Reddit solves many of these problems, due to the unique structure of the website and the voting system on comments and posts. Although Reddit gets less media attention than other social media websites, it carries a substantial following. A 2016 Pew survey found that 2.8% of US adults get their news from Reddit [4].

This paper will first overview our process for creating a linear regression model that can correlate social media activity with polling data, using a time-delay offset. This approach could lead to much faster analysis of the public's response to a news event (hours, rather than days) as well as more accurate election predictions. We used this approach to create a model that could predict the popular vote margin of the 2016 election within a fraction of a percentage using data from Reddit up through 2 weeks before election day. Finally, we will overview a process to create new forms of political analysis using similar methods of analyzing data on Reddit. Using social media data for political analysis could resolve long-standing public debates about which topics were especially important in deciding the 2016 election.

2 Data

The main data corpus for this study was ingested from Reddit, while public polling data was scraped from the polling aggregation website Real Clear Politics [1].

2.1 Definitions

The following terms related to Reddit will appear throughout this paper and are defined below.

- **Subreddit:** Reddit's version of a subforum for a specific topic. Subreddits contain links and posts by users, each of which contains comments by other users.
- **Upvote:** Reddit's equivalent of a like or favorite. Increases the score of a comment by 1. Usually used by users to express agreement or pleasure with a certain comment.
- **Downvote:** Similar to an upvote, but instead reduces the score of a comment by 1. Usually used by users to express disagreement or displeasure with a certain comment.
- **Comment Score:** Cumulative sum of upvotes and downvotes on a comment. Since the number of votes on comments is much higher than the number of comments, using the comment score as an input to a model increases the sample size of the model.
- **/r/politics:** The largest subreddit on Reddit focused on politics. This subreddit was chosen for this study due to its moderation policy and the quality of results.

2.2 Reddit Corpus

2.2.1 Why Reddit?

The basis of the data for this study was a corpus of 7.2 million Reddit comments posted during the 2016 election queried from the Reddit public API. Compared to other social media websites like Twitter or Facebook, Reddit stands out as a data source for a couple of reasons. Reddit is a centralized website, where people post articles on "subreddits" rather than on their own accounts. This means that discussion around a certain news story is far more concentrated on Reddit than on other social media websites such as Facebook or Twitter. Rather than a small amount of discussion spread across a large number of posts, Reddit features a very large amount of discussion spread across a small number of posts. Reddit also includes a more flexible voting system with positive and negative votes ("upvotes" and "downvotes"). This means that users who disagree affect the scores of comments, rather than just those who agree.

2.2.2 Reddit's Userbase

Since our data showed that there were far more upvotes on comments than there were comments, including the comment score allows us to include the opinion of a larger number of Reddit's users. Considering that as much as *2.8%* of the US population used Reddit to get news during the 2016 election [4], this means that the sample of the population this data draws from is far larger than any opinion poll even if only 10% of those users regularly comment or vote on comments. 0.28% of US adults would include many hundreds of thousands of voters, while most public polls rarely include more than a couple thousand voters.

2.2.3 Selecting a Subreddit

The subreddit `/r/politics` was selected for this analysis, since it contained the most consistent activity about the election. At first, this subreddit may appear to be a poor choice due to its notable liberal bias (Reddit's userbase was mostly college-aged white men during the 2016 election) [4]. However, the subreddit functions as a relatively open board for discussion, as the moderators only remove comments that violate rules of civil discussion, as opposed to other partisan subreddits that remove comments that oppose the political ideology of the subreddit. In addition, the moderators remove comments and posts that are not related to politics or spam, which effectively cleans the data before we even collect it.

In addition, even though most of the articles and comments had a liberal bias, conservative comments or comments critiquing Clinton were still frequent, especially during news events featuring negative stories about Clinton. And since the correlation to the polls in our model is a relative measurement rather than an absolute measurement due to the nature of linear regression, the absolute bias of Reddit's userbase does not matter as long as there is some amount of representation of both sides of the debate. In other words, the bias of /r/politics is constant throughout the election, and our model measures deviations from this baseline bias by correlating it to a relatively unbiased source (polls).

2.2.4 Example Post

Figure 1 is a labelled screenshot of a sample Reddit post and comment section from a notable news event of the 2016 election. This is to help the readers unfamiliar with Reddit better understand the user interface of Reddit.

- **A:** The news article that is the subject of discussion for this post. The number on the left represents the post score of the article, 42,200. This information is not used in our model.
- **B:** One of the top scoring comments of this post. Each comment also can be replied to. The comment text is used in our model.

- **C:** The comment score of this comment, as well as the date posted. The precise date and time of the post can be found by hovering over with a mouse, or through the API. The comment score and time posted are essential to our model.

Figure 1

2.2.5 Data Collection

Data from Reddit was ingested using the Reddit API through the Python Reddit API Wrapper (PRAW). All comments posted during the 2016 election season in a thread with at least 10 points were downloaded and added to a CSV file for analysis. First, a now-deprecated API call would

return all "Submission" objects from the `/r/politics` subreddit. Next, other API calls are used to find all comments for all of these submissions that have a positive score (more *upvotes* than *downvotes*. Finally, the comment text, date, and other metadata are stored in a CSV. At first, the data was processed directly from the CSV, but later the data was ingested into a Solr instance. This reduced active memory usage during later analysis from upwards of 1 GB to less than 100 MB without sacrificing query times. A few sample comments from the raw data are shown in *Table 1*.

Table 1

timestamp	*comment_text*	*comment_score*
7/31/2016 2:59	I've started to wonder if anything can affect Trump's numbers	626
7/18/2016 18:07	You could very easily argue that the stakes are higher now than The Apprentice.	39
...	...	...

2.3 Polling Data

The polling data used to correlate with Reddit data was the historical polling average data from Real Clear Politics (RCP) for the 2016 Presidential Election [1]. Real Clear Politics itself is an accumulator of many different polls measuring the presidential race. Each poll stays for one to two weeks on the average before being dropped or replaced by a newer version of the same poll. Averaging the polls smooths out the inaccuracies of individual polls, which often have large margins of error. Additionally, a separate moving average is later applied on top of the polling average (explained further in the Data Analysis section). Several candidate sources were considered for polling data, including other polling aggregators such as Huffington Post. However, RCP was ultimately used for the data, since it was simple to use and included all major national polls. After this research was initially conducted, FiveThirtyEight also released a polling aggregator[1] that would be suitable for this kind of research. Data from RCP was collected by pulling directly from the API that feeds the user interface of the website.

2.4 News Data

While news articles were not directly measured in this paper, all of the Reddit comments came from discussion about various news articles during the presidential campaign. A post-election Harvard study showed that coverage of both Hillary Clinton and Donald Trump was "overwhelmingly negative in tone" [11]. This negativity simplified our sentiment analysis model as explained in the next section. Due to the negative nature of the election, the sentiment of every comment mentioning a candidate is assumed to be negative.

[1] http://projects.fivethirtyeight.com/general-model/president_general_polls_2016.csv

3 Data Analysis

3.1 Preparing the Data

The first step in preparing our dataset from the raw Reddit comments data is applying a simple formula to the comments for each day in our data. The output is a time series dataset that maps each day to the cumulative score of all comments posted that day including one of the chosen keywords.

$$\begin{aligned}
\text{candidate_score_t} &= \sum_{i \in S_d} i_{comment_score} \\
t &= \text{The given date} \\
S_d &= \text{The set of all comments posted on date } t \text{ that contain at least one keyword in } k \\
comment_score &= \text{The cumulative score of a comment} \\
k &= \text{Set of keywords for the given candidate, listed below}
\end{aligned}$$

Table 2

Candidate	**Keywords (case insensitive)**
Hillary Clinton	Hillary, Clinton
Donald Trump	Donald, Trump

Although news cycles move quickly, it is rare that public opinion about candidates will swing multiple times within a single day. Collapsing the data to single days simplifies the model to discrete time steps without reducing the potential for useful output.

By including the candidates first and last names as separate keywords, we widen the net to capture as many comments discussing either of the candidates as possible. Online commenters often only use either the first or last name of the candidates when discussing them in the comments.

The resulting dataset contains three main variables, and another variable for visualization purposes:

- **date:** The time series variable representing the date of the data in other columns. Used to calibrate with polling numbers later.
- **clinton_score:** The first dependent variable representing the daily candidate word based on the keywords defined above for Hillary Clinton.
- **trump_score:** The second dependent variable representing the daily candidate word based on the keywords defined above for Donald Trump.

- **clinton_lead:** This variable is equal to **clinton_score** − **trump_score**. As a linear combination of the two variables, it does not add anything new to the model but is useful for visualization ("Difference").

3.1.1 Moving Average

Next, a moving average over k days is applied to smooth out both the Reddit data and the polling average. The notation below is borrowed from O'Connor et al.[10], who also used a moving average in their model using Twitter.

$$MA_t = \frac{1}{k} * (x_{t-k+1} + x_{t-k+2} + ... + x_t)$$

Smoothing the data helps eliminate daily noise in Reddit data and helps mitigate the inaccuracy and volatility of day-to-day polling averages. Due to random variation, the average of polls on a day-to-day often includes a fair amount of noise. However, taking a rolling average over the last several days mitigates this noise and provides a more accurate representation of polling during the election.

3.1.2 Transformed Data Example

After applying the transformations earlier in 3.1 as well as a moving average, the data is in complete time series form and ready for analysis and modeling. An example of what the transformed data looks like is displayed below in *Table 3*. The columns correspond to the variables described in section 3.1. Note that while the score of a comment will always be an integer value, the scores below are floating point numbers due to the moving average.

Table 3

date	*clinton_score*	*trump_score*	*clinton_lead*
2016-10-03	53697.250	118792.875	65095.625
2016-10-04	50320.250	120401.875	70081.625
2016-10-05	38972.750	99634.500	60661.750
2016-10-06	37250.875	93156.000	55905.125
...	...	...	...

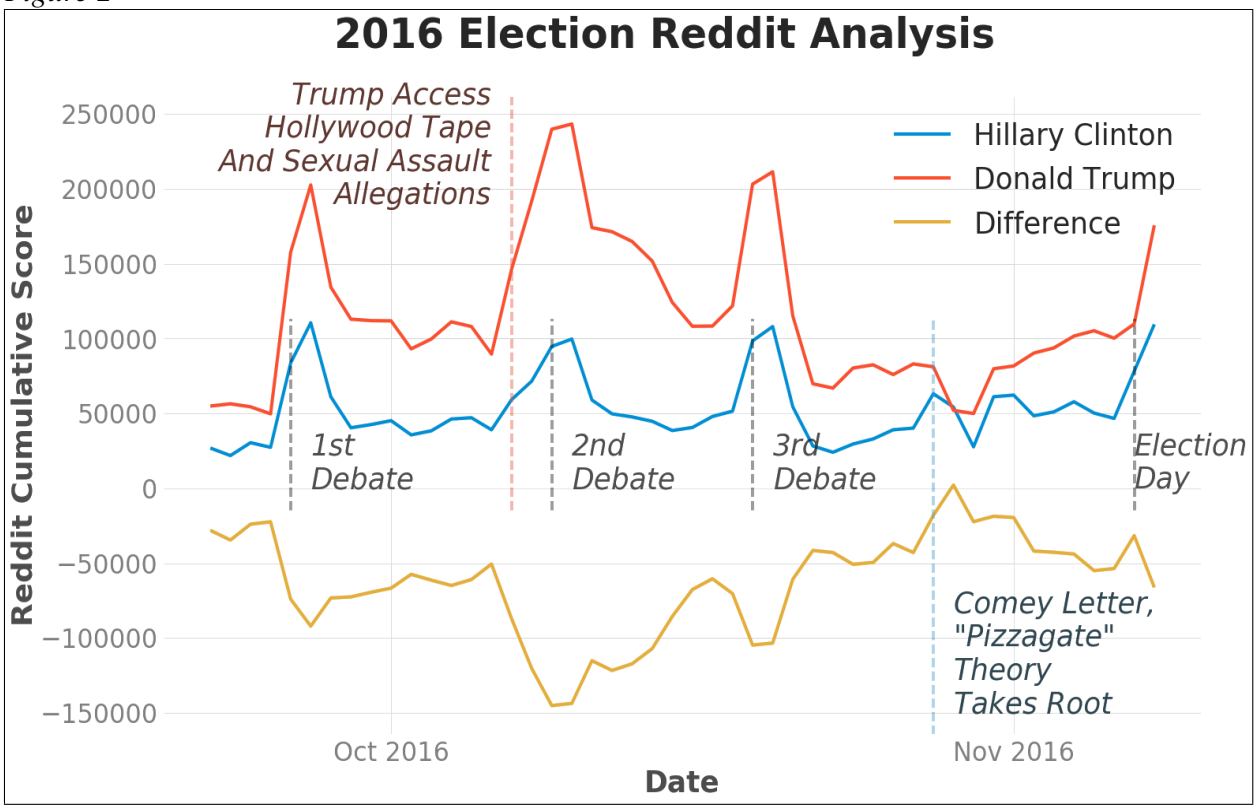
3.1.3 Preliminary Data Analysis

Even before putting this data through a model, the connection between Reddit discussion data and real-world news events is immediately apparent. The marked news events in *Figure 2* line up with

large spikes in online discussion about various news events during the 2016 election cycle. Even with a slight moving average ($k = 2$), the data is extremely responsive to news events. It also shows the relative impact of different news events that is useful when correlating to public polling. For instance, the impact of the Access Hollywood tapes and sexual assault allegations against Donald Trump had a larger affect on online discussion than the first or second debate.

The yellow "Difference" line on the graph shows the potential for this data to correlate with polling. Inverting the data to account for the constant negativity of online discussion about candidates, the yellow line closely follows the contours of the polling average throughout the election if interpreted as Clinton's polling lead over Trump. Trump's polling was *weakest* during mid-October in the aftermath of the Access Hollywood tape, as shown by the dip in the line at the same time. Trump's polling peaked in the weeks leading up to the election, with a slight dip around election day. This pattern is also shown in the later part of the graph.

Figure 2



3.2 Generating a Model

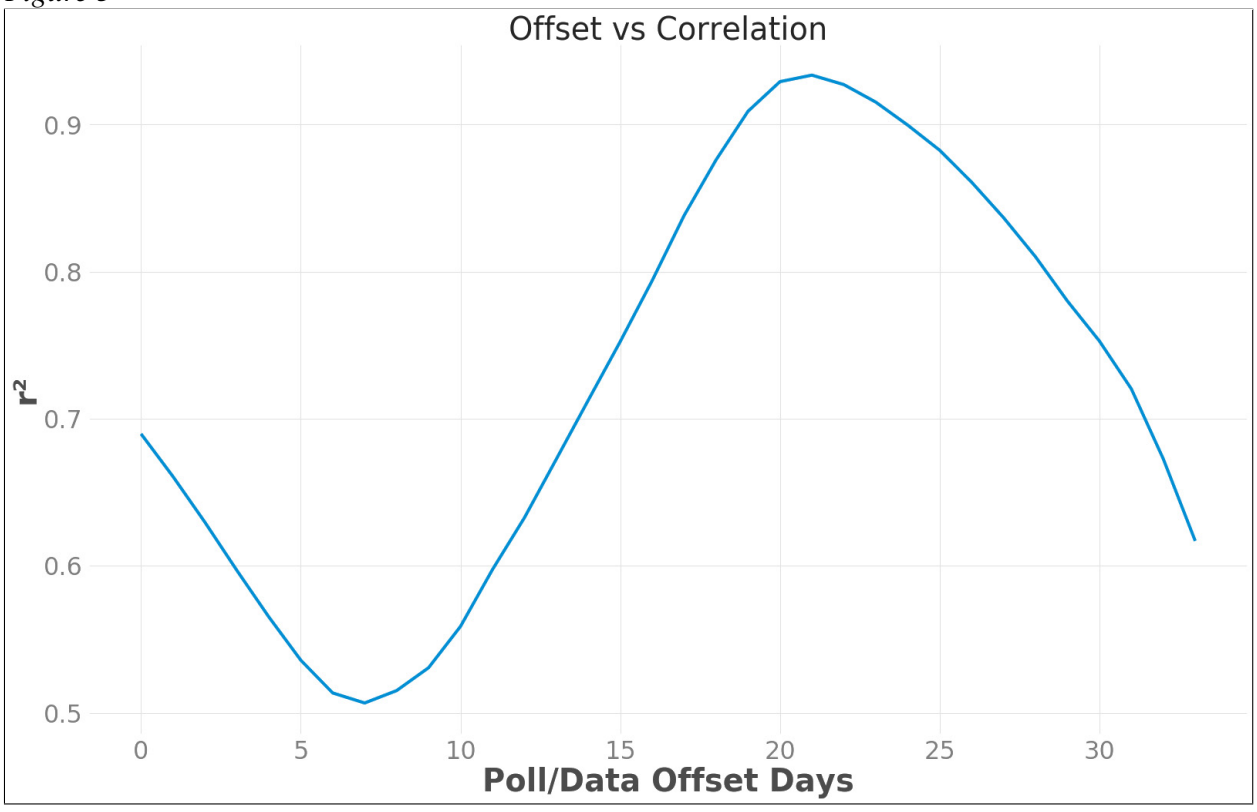
After the data preparation steps, the last steps are choosing a timeframe of data, and offset between polling data and Reddit data.

3.2.1 Choosing the Offset

One important difference between our model and traditional polling methods is the time delay. Social media data can be collected within 24 hours of an event happening, while polls take several days to conduct, and then several more for the polling agency to verify, analyze, and publish the final results. This means that in order to properly compare the Reddit data to the polling data, we will have to find an optimal offset and apply it before the linear regression is run.

An offset of o days is applied in the data by shifting all the poll data o days forward. To find the optimal offset value, we run our model several times with different values and choose the one with the highest correlation coefficient. *Figure 3* shows a graph showing correlation coefficients for different days of offset, using a moving average of $k = 14$. The results show that o is optimal at 21 days of offset. Accounting for the 14-day moving average, this means that polls lag the Reddit data by around 6 or 7 days. This is what we would expect, since polls usually take about a week after polling starts to publish the results.

Figure 3



3.2.2 Defining the Model

Since our research compares two datasets over the same time series that measure the same ground truth, a standard linear regression model is the most appropriate model. We have multiple independent variables for the output representing the discussion levels of each candidate, so the multiple

linear regression variant will be used. A multiple linear regression allows us to include the data for both candidates when correlating the output poll data.

In addition, we create a multiple linear regression model for each of the outputs of the polling API: Trump's popular vote percentage (PVP), Clinton's popular vote percentage, and Clinton's percentage lead over Trump. Clinton's lead was used for the third variable in order to keep the values above zero. Note that we do not use the clinton_lead ("Difference") independent variable shown in Figure 2 because it is a simple linear combination of clinton_score and trump_score. This means that the linear regression receives no new information by including the data. We verified this by creating regressions with clinton_lead and without it, and received the same output model each time.

Our final model consists of three multiple linear regression models, as described in Table 4.

Table 4 (Training Data)

Model	**Independent Variables**	**Dependent Variable**
A	trump_score, clinton_score	Clinton PVP
B	trump_score, clinton_score	Trump PVP
C	trump_score, clinton_score	Clinton Lead PVP

These models can be described mathematically as follows, where the multiple linear regression solver learns α, β_1, and β_2 (with error term ϵ) from the training data for each model.

$$
\begin{aligned}
t &= \text{date};\ o = \text{offset} \\
A_{t+o} &= \alpha + \beta_1 \text{trump\_score}_t + \beta_2 \text{clinton\_score}_t + \epsilon \\
B_{t+o} &= \alpha + \beta_1 \text{trump\_score}_t + \beta_2 \text{clinton\_score}_t + \epsilon \\
C_{t+o} &= \alpha + \beta_1 \text{trump\_score}_t + \beta_2 \text{clinton\_score}_t + \epsilon
\end{aligned}
$$

3.2.3 Interpreting the Model

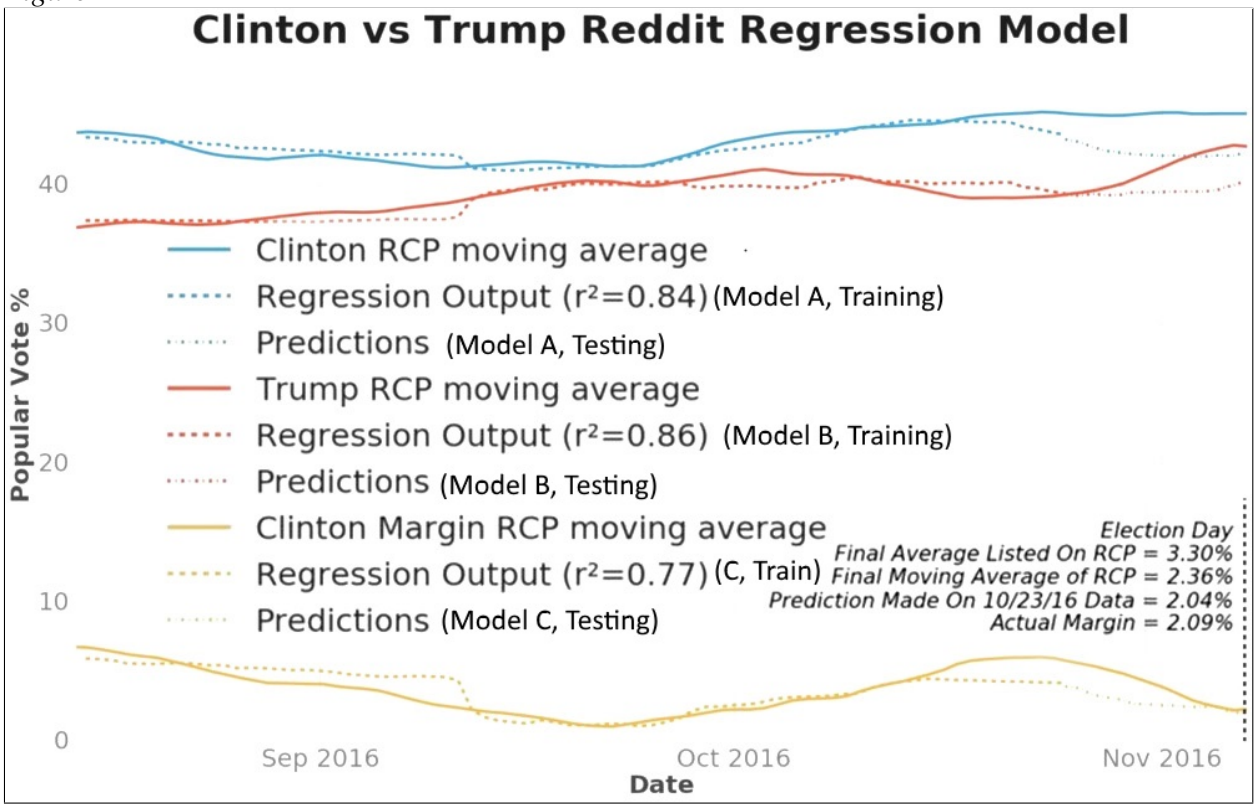
Before applying the model to see the relationship between Reddit data and polling averages, it is important to understand the nature of the relationship and why it exists. As noted earlier, media coverage of both candidates in the 2016 election was near unanimously negative in tone and content. Similarly, public perceptions and online discussion of the candidates tended to be very negative as well. Going into the election season in May, both Trump and Clinton had record-low favorability ratings for major party nominees for President [7]. These polls reflected widespread disappointment among the public of the quality of candidates available to them. This distaste was reflected by much stronger than average support for third parties in polls and the general election as well.

One consequence of the uniquely negative nature of the 2016 election is that the amount of online discussion of a candidate *inversely* correlated with their performance in the polls. That is, the less a candidate was talked about relative to the other candidate, the better they would do in the polls. This means that our model assumes that all comments were negative in nature. Although there were a small amount of positive comments of course, those were usually crowded out by the negative comments. This assumption allowed us to avoid using more complicated models to determine the sentiment of various comments.

3.2.4 Applying the Model

With our offset and moving average values selected, we created a linear model to predict the popular vote margin of the 2016 election using data up through October 23rd, 2016 (2 weeks prior to the election). We used `sklearn`'s `LinearRegression` module to create a multiple linear regression model that used the prepared time series dataset for each candidate to predict the the moving average of the RCP polling average, as described earlier in *Table 4*. The graph showing the outcome of this model is in *Figure 4*, with labels corresponding to *Table 2*.

Figure 4



3.2.5 Verifying the Model

As shown in the graph, our data has produced an accurate model for predicting polls. We can verify the accuracy of the model through the results of the linear regression, comparison to related works, and an examination of the raw data.

- **Coefficient of Determination (r^2):** The r^2 coefficients we obtained, between 0.77 and 0.84, signify that Reddit may be a strong of a source for measuring public sentiment and even predicting polls that have not released yet. According to this metric, 84% of the variation in poll data could be explained by variation in the discussion of either candidate on Reddit. This is a very high result considering that the polls themselves have margins of error from the true sentiment of the population. Normally, comparing two imperfect measures of a ground truth would be expected to have a somewhat low r^2 coefficient, so this high value may indicate that the model based on Reddit data tracks public sentiment very accurately.

- **Testing Set** We created a testing set to verify the model could predict real election results on its own. This testing data was the predicted PVPs of the final 14 days of the race, labelled "Predictions" in *Figure 4*. For Model A and B (Clinton and Trump), these predictions were inaccurate relative to the PVP of actual result. However, the polling averages also included this error. Generally, polling accuracy is evaluated by the predicted margin rather than the raw PVPs of each candidate, since polls tend to overrepresent third party voters and also include an option for "Undecided" or "Don't know," which are not options on the presidential ballot. Since our model is based on the polls, we should also use the margin of victory to judge the accuracy of the model. The output of our model for this testing set proved to be highly accurate, predicting that Clinton would win the popular vote by 2.04% when she won by 2.09%. Our results show significant potential for the use of Reddit data as a basis for predicting not only public opinion polls but the popular vote margin of elections as well.

- **Raw Data Analysis:** Examining the data from *Figure 2* and comparing it to the polling data in *Figure 4* shows visually how this model works. The "Difference" data both responds to news events and also corresponds to changes in the RCP polling average. This visualization shows that it is highly unlikely that the correlation between our data and the polls is the result of a coincidence in a more intuitive way than the statistical tests.

- **Related Work Replication:** O'Connor[10] et al. were able to achieve a maximum r^2 of 0.86 when using Twitter to correlate with polls in the 2008 Presidential Election. O'Connor used a more sophisticated sentiment analysis scheme, which may have allowed them to achieve a higher coefficient of determination. In addition, they were operating on an election 8 years prior on a different social media platform. However, the fact that they achieved a similar result on a different social media platform on a different election cycle shows that the process of using social media to predict election outcomes is effective and replicable.

3.3 Topic Analysis

For analysis on the specific events or topics that affected each candidate's polling averages, the comments can be divided into separate political topics, such as the identity politics, Wikileaks, and the FBI investigation into Clinton's email server. These topics each contain several keywords that are used to create a time series data set in a similar way as the previous section. This method of search allows us to see which issues were most important at different times during the election cycle and see how they may have affected either candidates' polling average.

The algorithm for transforming the raw comment data into the prepared dataset is a modified version of the earlier algorithm. The main difference is that this analysis uses relative, rather than absolute scoring, of comments. This means every score was divided by the *total score* of all keyword categories in order to control for varying amounts of total discussion over time. By controlling for the variance in the total amount of discussion over time, we can more easily understand how different political topics affected the election at different points in time. Without controlling for the total amount of discussion (which generally increased over time), topics that were discussed earlier in the election would appear to be less significant than they really were when they happened.

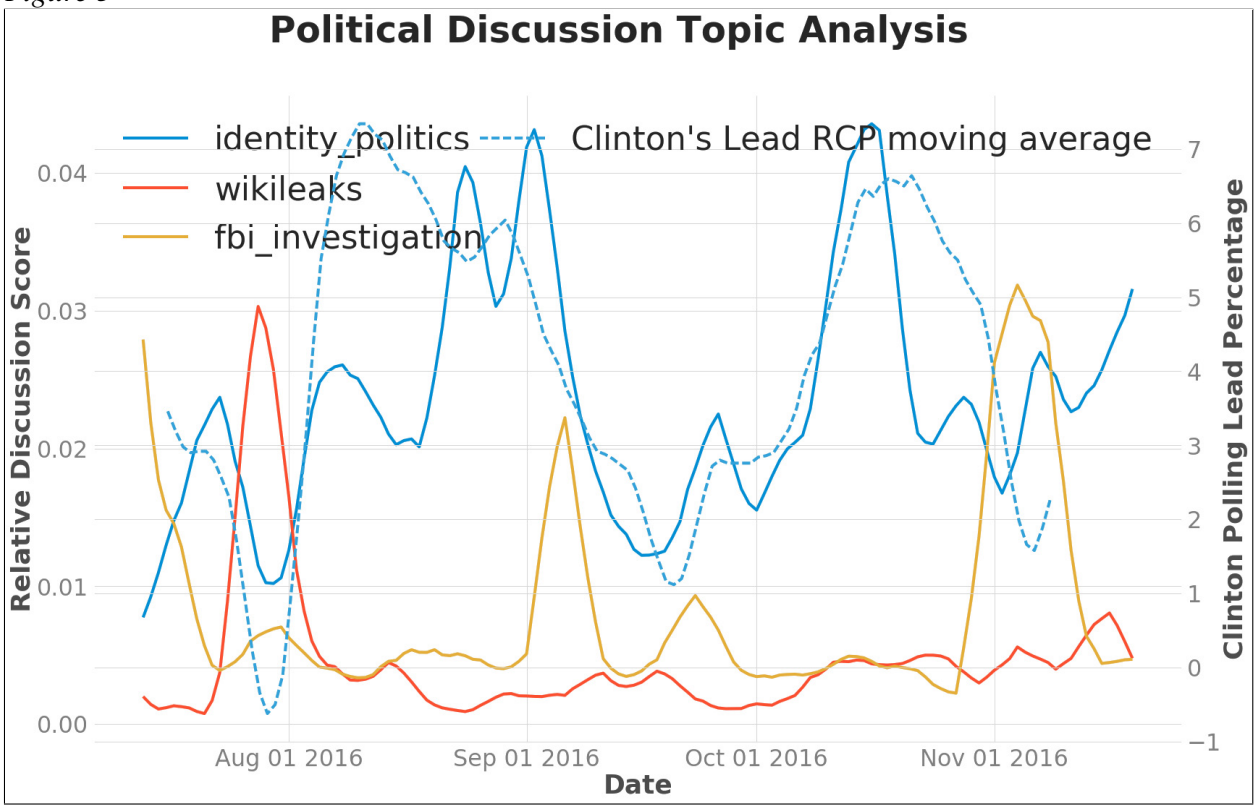
$$\begin{aligned}
\text{topic_score}_{\text{t}} &= \sum_{i \in S} \frac{i_{comment_score}}{total_score} \\
t &= \text{The given date} \\
S_d &= \text{The set of all comments posted on date } t \text{ that contain at least one keyword in } k \\
comment_score &= \text{The cumulative score of a comment} \\
total_score &= \text{The cumulative score of all comments matching any keyword} \\
k &= \text{Set of keywords for the given topic}
\end{aligned}$$

To demonstrate this technique, several topics related to the election are graphed below in *Figure 5*. The keywords used to identify each topic are included on the last page in the Appendix.

This type of data can be used for analysis that is far more advanced than what you can normally glean from public polling data. Although polling surveys often ask how voters feel about certain topics, these questions rarely show whether a certain news story actually *changes* the vote of a voter. Using the data in *Figure 5*, we can draw a rough outline of what topics and events drove both online discussion and public polling in the 2016 election.

1. Clinton's polling lead was driven largely by discussions of *identity politics,* especially in the aftermath of the Access Hollywood scandal in mid-October (where a tape was released in which Trump could be heard talking in vulgar language about women). The importance of identity politics in driving Clinton's polling numbers is clear from the data, but runs in contrary to the post-election analysis of many people who blamed the loss of Clinton on

Figure 5



identity politics. Clinton was strongest when the discussion *was* about identity politics, and weakest when discussion shifted to other topics.

2. Trump gained most when the discussion was about either the controversy surrounding Clinton's private email server (early July and early November), or around the emails leaked by Wikileaks from both the DNC email servers (released in late July) and John Podesta's campaign email (released in early October). Considering that the US intelligence community widely believes that the emails released by Wikileaks were acquired through a concerted hacking effort by Russia, this data would imply that these hacking efforts had a substantial negative effect on Clinton's polling averages at several points during her campaign.

3. Polling data was highly sensitive to the most recent news events throughout the election cycle. This would imply that the news events happening in the very last moments leading up to the election would have the largest impact on the outcome. Indeed, roughly a week before the election, James Comey revived Hillary Clinton's email controversy by notifying Congress that the FBI had reopened the investigation into Clinton's emails after discovering new material while investigating Anthony Weiner for an unrelated case. The effect of this news event is clear–discussion about the FBI investigation spiked, while Clinton's polling lead evaporated nearly instantly.

4. This data also shows that despite the common perception that /r/politics is an echo chamber for liberal beliefs, a significant amount of discussion featuring conservative discussion topics that hurt Hillary Clinton also occurred on the subreddit. The difference between the perception and reality is likely due to the fact that most users only read the top few comments of Reddit posts (see *Figure 2*), which do usually have a liberal bias. However, there are still plenty of conservative comments that are lower in the comment section that are included in our model. Since our model is a more comprehensive view of Reddit, it allows us to see that the overall bias of the subreddit is not as overwhelming as it may seem.

The conclusion that the election outcome was largely shaped by the Comey Letter is shared by Nate Silver, editor of the popular political prediction statistics website FiveThirtyEight [15]. However, his reasoning for the conclusion was merely that Clinton's polling dip correlated with the timing of Comey's letter. The Reddit analysis provides more robust evidence that there was a causal link between the Comey letter and the last-minute decline of Clinton's polling performance, since it shows that the Comey Letter was a large subject of discussion at the final stage of the election.

4 Comparison to Related Work

O'Connor et al. created a similar model as ours using Twitter data on the 2008 Presidential Election between Barack Obama and John McCain [10]. Instead of just using candidate names, they used a slightly more complicated model that used keywords to measure sentiment. Since the 2008 election was less uniformly negative than the 2016 election, this increased complexity may have been necessary. After O'Connor's study, researchers have been able to repeat the connection between public polling and Twitter sentiment data in several other studies. In 2012, researchers were able to predict use Twitter data to predict public opinion during the 2012 President Primary Elections [14]. Another study using Twitter data found that "ambient happiness of selected words correlates well with solicited public opinions", and also found a time delay between data on Twitter and the data in public polling [5]. The concept of using social media data for political prediction has been replicated to predict elections in other countries as well, with one study using social media data to create a prediction model for the 2014 Taiwanese elections [16].

A systematic literature review of studies conducted in 2017 that use social media to predict the future found that the power of social media of prediction extended to many topics besides politics, such as public health, stock markets, or movie ticket sales [12]. Our model, which uses Reddit data on the 2016 election, verifies that the measurable connection between online discussion and public polling can be replicated on social media websites other than Twitter. Until now, there have been no studies published that successfully connected Reddit data to public opinion polls. Kruitwagen attempted to make this connection, but was unable to find meaningful results [9].

In addition, our type of detailed political topic analysis was not conducted in other studies. We believe that this type of analysis using data from social media could be groundbreaking for resolving public debates about which news events or topics contributed most to the outcome of the election.

5 Future Work Discussion and Potential Limitations

Although our work and the work done in related studies have shown that social media has robust capability for predicting election polling, this method still has several drawbacks that will require additional work to address. Since our model still relies on public polling to create the baselines model, it would be subject to any systematic bias in polling for the given election. For 2016, however, *national* polls were fairly accurate at estimating the outcome of the election. The 2-week moving average of the RCP polling average estimated that Clinton would win the popular vote by 2.36%, and her actual margin was 2.09%. However, if an election included systematic problems at the national level, that would taint the Reddit-based model as well. Therefore, it would be best to create a way to estimate the results of a public opinion poll without relying on them to create the baseline in every case. This could be done by creating a *general model* that relies on polls that were known to be accurate. The general model could then be used on any arbitrary set of Reddit data to estimate how that data translates into public sentiment as a whole.

An issue specific to US elections is the model's weakness to the peculiarities of the Electoral College. The anonmyized nature of Reddit means that we can't produce anything other than national public sentiment. However, as seen in the outcomes of the 2016 and 2000 Presidential Elections, a candidate winning the national popular vote on the whole does not necessarily translate into winning the electoral votes of enough states to win the election. This could be rectified by creating individual models for each state based on that states' polling average, and then morphing these individual models into a complete prediction of how many electoral votes each candidate is likely to win. However, this approach would be susceptible to the same polling biases mentioned earlier. State polls are also less frequent and less accurate than national polls, which would increase the uncertainty of the overall model. Incorporating geographical data from Twitter or even Facebook could help alleviate this problem, but users usually do not opt to share their location publicly on those websites. In addition, collecting geographical data of millions of users would raise privacy concerns for the users, even if the data is publicly listed.

The unique nature of social media also poses certain challenges for using it as the basis for a public prediction model. The 2016 election showed how foreign actors can intentionally influence the public through the use of social media and the internet [3], and these same actors could potentially sabatoge a model based on public social media data. It would only take a few hundred dollars worth of fake accounts to directly manipulate the data that feeds into the model through fake comments or fake votes on comments. Public opinion polling does not have this issue, since polls retrieve their data directly from real voters.

Finally, although our model may prove to be robust from a statistical standpoint, it may take a considerable amount of effort and outreach to convince the public that such models are as credible as public opinion polls. Public opinion polls have been prominent in the US since the 1940's and their methods to gather and interpret data are understood by many voters [13]. Using social media to predict elections, in contrast was only discovered in 2010 and is still a developing field. In order to address this credibility gap, news organizations or bloggers who wish to use these models will need to convince their readers that the models are accurate and backed by science. Interactive tools and open data and source code could allow individuals to verify that these techniques work by themselves. Endorsements by well-known and trusted news organizations or websites would also help cement the credibility of such models. Eventually, the models could speak for their own credibility through a history of accurate predictions.

6 Conclusion

Overall, the positive results of our models show promise for using data from Reddit to predict elections. Since the same type of analysis has been conducted on Twitter for separate election years with good results, we believe that the model is highly flexible and could be used for many types of elections across the world at a very low cost. While public acceptance may be an obstacle in the short term, one could see a future where social media data from websites like Reddit are tracked in the news the same way public opinion polls are today.

We believe that this paper demonstrates how social media data analysis could be a public good. Politics on social media to an individual user is often highly confusing and lacks context. However, on a broad level that captures data across long periods of time, social media data on Reddit can be used as a tool to determine what topics swung an election with scientific precision. Providing the public with more robust and detailed information about national political trends could help voters understand each other better, and reduce the amount of confusion inherent to the current political climate.

7 Appendix

Topic Analysis Keywords

Topic	Keywords
fbi_investigation	'fbi', 'investigation', 'email server', 'confidential', 'comey', 'benghazi', 'anthony weiner'
identity_politics	'identity politics', 'sexual assault', 'rape', 'rapist', 'grope', 'sexism', 'sexist', 'blm', 'black lives matter', 'racist', 'racism', 'hate crime', 'black people', 'blacks', 'muslim', 'bigot', 'mexican', 'latino', 'hispanic', 'asian', 'gay', 'lgbt', 'lesbian', 'transgender', 'bisexual'
wikileaks	'wikileaks', 'pedo', 'pedophile', 'pizzagate' , 'podesta','spirit cooking', 'rigged primary', 'debbie', 'pay for play', 'leaked email', 'assange', 'open trade and open borders','donna brazile', 'paid speeches', 'transcripts'

References

[1] General election: Trump vs. clinton.

[2] reddit.com traffic statistics.

[3] Assessing russian activities and intentions in recent us elections, 1 2017.

[4] Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 1. reddit news users more likely to be male, young and digital in their news preferences, Feb 2016.

[5] Emily Cody, Andrew Reagan, Peter Dodds, and Christopher M. Danforth. Public opinion polling with twitter. 08 2016.

[6] Nate Cohn. A 2016 review: Why key state polls were wrong about trump, May 2017.

[7] Harry Enten. Americans' distaste for both trump and clinton is record-breaking, May 2016.

[8] John Allen Hendricks and Dan Schill. The social media election of 2016, Jan 1970.

[9] Marlijn Kruitwagen. Does reddit know better? 2 2017.

[10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. 11, 05 2010.

[11] Thomas E. Patterson. News coverage of the 2016 general election: How the press failed the voters. Dec 2016.

[12] Lawrence Phillips, Chase Dowling, Kyle Shaffer, Nathan Hodas, and Svitlana Volkova. Using social media to predict the future: A systematic literature review. 6 2017.

[13] Lily Rothman. Political polling history: How gallup got his start.

[14] Lei Shi, Neeraj Agarwal, and Rahul Garg. Predicting us primary elections with twitter. 2012.

[15] Nate Silver. The comey letter probably cost clinton the election, May 2017.

[16] M. H. Wang and C. L. Lei. Boosting election prediction accuracy by crowd wisdom on social forums. In *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 348–353, Jan 2016.