Determining which Sentiment Analysis Predictive Model to use for a given Social Media Election Dataset

Lena Renshaw, Brown University, April 2020

Undergraduate Thesis, Bachelor of Science in Computer Science

Advisor: Dr. Jeff Huang
Second Reader: Dr. Ellie Pavlick

1. BACKGROUND

1.1 The Predictive Power of Social Media

As we begin the new decade, the exponential growth of the Internet, social media, and social networking sites continues to dominate our lives. Internet access is available to a wide range of citizens globally, and in step, social media usage is growing quickly. In 2005, 16% of the population had Internet access; as of 2017, this has grown to 48%. Moreover, in the developed world as of 2017, 81% of people have access to the Internet [11]. As of 2018, Facebook had 2.26 billion monthly active users (MAUs), followed closely by YouTube (1.9 billion), Instagram (1 billion), and Twitter (329.5 million) [5].

As social media usage increases, it continues to become more and more representative of populations as a whole. This has had a demonstrable effect on the world at large, with industries such as advertising, retail, and transportation being completely transformed in the past decade through the use of the web. Moreover, social media has had substantive effects on real-world politics, with users using the sites to organize protests, demonstrations, and revolts, propagandize political campaigns, and disseminate fake news [3].

In the United States and abroad, political parties and candidates have leveraged the power and reach of social media to reach out to constituents and citizens. Some have hired staffers who work exclusively on creating their image online, others choose to disseminate information seemingly at random. Using these systems smartly can dramatically change popularities and outcomes of elections; in a study done at the by Livne et al at the University of Michigan, researchers found that in the United States 2010 midterm elections, the republican party was able to more acutely and effectively leverage social media that subsequently led to their victory [9]. The speed by which politicians can access their followers on Twitter, a popular tool allowing for rapid micro-blogged tweets (status updates), is unprecedented.

The extreme growth of social media has also raised the possibility of using the Internet as a means to obtain data about different real-life phenomena, from the potential success of movies to the marketability of consumer goods to politics on a global scale [12]. The Internet is already used to explore and track different preferences of citizens, including political preferences that shape elections, democracies, and nations. It is a trove of information useful for monitoring public opinion and possibly predicting the futures of such democracies. Increasing developments in quantitative and qualitative sentiment analysis (SA) as well as natural language processing (NLP) have enabled us to exploit such information in a more reliable manner [3]. This analysis also opens up the possibilities of analyzing data beyond traditional polling metrics in more of a survey-like manner beyond best-polling candidate, such as correlation with economic and consumer confidence nationally and presidential approval rates [13]. This variety in polling questions would only be limited in scope by the topics and accuracy of the information online users choose to share [13].

Compared to traditional surveying and polling, running an analysis using social media data is compelling for a number of reasons. It can capture data about a specific event, candidate, or policy much faster than running a survey. It is also substantially cheaper: social media websites such as Twitter often provide a public API that anyone can use to scrape data on the website [3]. Social media data is also constantly updating, and in theory, if used correctly, can provide more dynamic predictions of future human behavior than traditionally static surveys administered over a period of time [3]. This data can be analyzed to observe trends and breaking points that correlate to real-world events, which can assist both scholars and analysts: researchers can benefit from the sheer amount of information available, and politicians can exploit this data to make changes to their policies or campaigns [3].

1.2 Natural Language Processing and Sentiment Analysis

Natural language processing is a subfield of computational linguistics that aims to train computers to process large amounts of human language. Sentiment analysis is a subfield of NLP in which computers are trained to analyze the sentiment behind large amounts of human language. This can be as basic as classifying whether a statement is positive or negative, or as complex as determining the specific charged emotion behind some human language tidbit.

There are many methods to implementing SA since it is a relatively new field, especially within the realm of politics. One of the most straightforward and common methods is to aggregate a portion of the data for human subjects to initially classify, and use this data as a training set to apply upon the rest of the large dataset. Using human subjects as analyzers and classifiers is often the best approach to finding out whether or not an implemented SA system is successful [4].

Most SA systems use some combination of human tagging and classifying as well as predetermined lexical features and patterns to obtain relevant data points in the first place. An example of this would be scraping the Twitter API for all tweets from a given week in time, and then grabbing from this dataset all Tweets from that time that contained “Obama” and related words. After that, human classifiers might go through this data to determine, for each Tweet, if it was a positive or negative Tweet about President Obama, or, if a sliding scale of values between positive and negative was being used for nuanced classification, they might assign it a value somewhere on this scale accordingly.

1.3 Sentiment Analysis in Relation to Political Campaigns

The current literature about the feasibility of predicting political campaigns remains relatively small. There are approaches that do not use any type of SA and approaches that do. Nevertheless, it does, on the whole, support using sentiment analysis and social media data to predict campaigns [9].

Because this field is currently still emerging, there is not a standardized method of applying SA to a given set of social media data. In past research on this topic, researchers have most commonly explored applying one method of SA to one election to predict its outcome, applying many methods of SA to one election to predict its outcome, or applying one method of SA to many different survey and polling questions to predict the response of the polls. Of course, though the overall consensus is generally that these methods are effective, there are many studies that indicate that these social media-based data collection methods are ineffective. Is this based on the method itself, the data, or both? The Jungherr study suggests that a method that might work on one dataset might not work well on others. After applying the same methodology from the Tumasjan research on the exact same dataset but not ignoring the data related to one party as the original research did, the method predicted that this party would, in fact, win the entire election, when in reality they received under 2% of the vote [7]. Other research that used logistic or other regression methods to create a key set of features that could accurately predict an election hypertuned parameters and specifically selected key features relevant to that election [8]. Additionally, there are other features such as the demographics of those voting in comparison to those online that differ between datasets and likely factor into the efficacy of each method applied to predict outcomes.

Our research uses this seemingly flawed research as a jumping point to pose a set of related questions: given a set of social media data from two weeks to an upcoming election, can we determine which SA method will align most closely with polling? Can we determine which SA method will be most effective at predicting the outcome of the election? This research will focus on American elections, specifically those for the upcoming Democratic primaries of 2020. Additional research is still needed in comparing different types of elections, both within America and internationally.

2. RELATED WORK

2.1 Overview

This section will describe key previous studies that have attempted to use social media datasets to predict political phenomena. It is important to explore this research in detail because these studies inform the methodology that we decided to use in this research. The general conclusions of this research can be found at the end of this section.

2.2 Predicting Past Elections Without the Use of Sentiment Analysis

The following are studies involving general predictive methods on large datasets that were not reliant on sentiment analysis as part of the study. They are broken down into largely successful predictive methods and mixed or unsuccessful prediction methods.

2.2.1 Successful Prediction Methods

Alexander [1]

I. *Data:* This study used a polling dataset gathered from the Huffington Post to predict the outcome of the United States general presidential election in 2016.
II. *Methodology:* To do this, they implemented maximum likelihood estimators and Gaussian conjugate prior calculation.
III. *Results:* The model worked well at predicting the winners of states and thus did well in predicting the final results of the electoral college. However, there were some pain points in this analysis: lower level candidates had imprecise final results, not enough polls were used, and performance in swing states was relatively poor, reflecting the outcome of the polls themselves. President Trump’s performance in this study was regularly underestimated, which was reflective of the polling results.

2.2.2 Mixed or Unsuccessful Prediction Methods

Lui [10]

I. *Data:* This study made use of the volume of Google searches as a predictor for the 2008 and 2010 United States Congressional elections. These races were divided into groups based on the degree they were contested. The names of candidates and the results of the elections were downloaded from a New York Times dataset.

II. *Methodology:* The primary tool of obtaining Google search data was Google Trends provided by Google Analytics. Because raw search volume data is sensitive information that Google does not publish openly, G-trends does not provide concrete numbers. However, they do provide the Search Volume Index of a keyword, which is a processed version of the data: it has been scaled based on the average search traffic of a keyword and normalized to accommodate for densely populated regions in the world.

III. *Results:* There were only a couple of races that were predicted with a better probability than chance using Google Trends. The disparities between the groups of races were large: the best group's predictions were good in 2008 (81%), this same group had an extremely poor rate two years later in 2010 (34%). This method did worse than using incumbency as a baseline for all elections and predicting based on that. These predictions were far worse than that of professional pollsters: data from the New York Times polls was almost always twice as accurate as the data gathered from this study.

2.3 Predicting Past Elections With the Use of Sentiment Analysis

The following are studies involving general predictive methods on large datasets that were reliant on sentiment analysis as a part or all of the study. They are broken down into largely successful predictive methods and mixed or unsuccessful prediction methods.

2.3.1 Successful Prediction Methods

Kim and Hovy [8]

I. *Data:* Data was collected from messages posted on an election prediction project webpage, www.electionprediction.org. The website contains various election prediction projects (e.g., provincial election, federal election, and general election) of different countries (e.g., Canada and United Kingdom) from 1999 to 2006. For this data set, they utilized Canadian federal election prediction data for 2004 and 2006.

II. *Methodology:* There were five approaches in this research. The first approach, called NGR, consisted of three steps: feature generalization, classification using support vector machines (SVMs), and SVM result integration. The program first created generalized versions of sentences from the initial input. Then it classified each sentence using generalized lexical features in order to determine the valence of and parity in a sentence. Finally, it combined results of sentences to determine valence and parity of the full message. The other methods were FRQ, measuring overall frequency, MJR, measuring messages with the most dominant party, INC, predicting just the incumbent, and JDG, using judgement words.

III. *Results:* Of these methods, MJR performed the worst (36.48%). Both FRQ and INC performed around 50% (54.82% and 53.29% respectively). NGR achieved its best score (62.02%) when using unigram, bigram, and trigram features together (uni+bi+tri). The JDG system, which uses positive and negative sentiment word features, had 66.23% accuracy. Higher accuracies were found with more finely tuned parameters, and the study claimed to achieve, as a final result, an accuracy of predicting elections of 81.68%.

O'Connor [13]

I. *Data:* This study used 1 billion Twitter messages posted over the years 2008 and 2009, collected by querying the Twitter API as well as archiving the "Gardenhose" real-time stream. The subjectivity lexicon used to determine opinion was taken from OpinionFinder.

II. *Methodology:* This study tried to predict the following three categories: consumer and economic confidence polls from Gallup, presidential approval ratings, and election results. The data was analyzed through message retrieval (identifying messages related to the topic) and opinion estimation (determining whether messages expressed positive or negative emotions about the topic, essentially sentiment analysis).

III. *Results:* Overall, a relatively simple sentiment detector based on Twitter data can, at a high level, replicate consumer confidence and presidential job approval polls. It is encouraging that expensive and 128 time-intensive polling can be supplemented or supplanted with the simple-to-gather text data that is generated from online social networking. The results suggest that more advanced NLP techniques to improve opinion estimation might end up being able to replace traditional polls, however, they have not yet reached that point. In several cases, the correlations are as high as 80%. Additionally, the textual analysis did not account for many edge cases of the Twitter data, and as such could be substantially improved.

Tumasjan [14]

I. *Data:* This study uses the context of the 2009 German federal election to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment.

II. *Methodology:* Used Linguistic Inquiry and Word Count (LIWC) software for text analysis to conduct a content-analysis of approximately 100,00 messages containing a reference to either a political party or a politician in the month before the election occurred. LIWC is a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. 12 dimensions (sentiments) were provided to the software as a means for analysis.

III. *Results:* One of the primary results for this investigation was that the number of Tweets related to each party is a good predictor of its vote share, coming close to the results of traditional election polls. Here, error was determined by comparing mean square error to actual results of the election. Additionally, the co-occurrence of political party mentions seemed to reflect close political positions and plausible coalitions accurately.

Livne [9]

I. *Data:* Used the Tweets sent by electoral candidates themselves, not the general public. Crawled Twitter to obtain this data, executing a query on Google using their name and the keyword "Twitter" and retrieved the top three results, manually filtering out bad entries.
II. *Methodology:* The content produced and associated with each candidate was considered to be the body of the tweet and the content of any URLs linked to in the tweet, and it was assumed that this content represented the opinion of the candidate. A sentiment analysis method was also used on top of this in combination with logistic regression models with the dependent variable being the binary result of the race.
III. *Results:* Reported success in "building a model that predicts whether a candidate will win or lose with accuracy of 88.0%." While this concluding statement seems strong, a closer look in the claims reveals that they found their model to be less successful, as they admit that "applying this technique, we correctly predict 49 out of 63 (77.7%) of the races."

Ceron [3]

I. *Data:* Italian political leaders throughout 2011 as well as the French presidential ballot and subsequent legislative election in 2012.
II. *Methodology:* The Italian dataset was downloaded and analyzed using the ForSight platform provided by Crimson Hexagon, while for the French cases the data was collected through Voices from the Blogs and analyses were run in R. Used the method proposed by Hopkins and King: a supervised sentiment analysis (SSA). This is in contrast to a traditional sentiment analysis, wherein results are categorized based on the use of preset dictionaries, where words appearing in various classified buckets of sentiment have some pre-set sentiment associated with them. Though traditional sentiment analysis can be completely automated, it opens up the dataset to errors stemming from edge cases, such as sarcasm, irony, and ambiguity. The HK method involves human coders creating a training set of data and determining the sentiment of the dataset they have curated. The expected error of the estimate is approximately 3%.
III. *Results:* The predictive ability of social media analysis strengthens as the number of citizens expressing their opinion online increases, provided that the citizens act consistently on these opinions. There were results for each of the three elections considered:
    A. Italian leaders' popularity ratings in 2011: the mean absolute error between comparing sentiment to actual popularity ratings varied dramatically between different leaders, and the average mass survey ratings always appeared to be higher than the social media ratings. The correlation between mass surveys and social media ratings is positive, albeit not strong.
    B. French presidential ballot in 2012: the number of preferences expressed in the week before the election enabled them to correctly forecast the outcome of the polls, yielding predictions that were very close to those made by traditional surveys.
    C. French legislative elections in 2012: past a certain threshold of voter turnout, sentiment analysis appears more and more effective.

Analysis shows a consistent correlation between social media results and results that could have been obtained from traditional mass surveys, as well as a remarkable ability for social media to forecast electoral results on average (which produces results more accurate than chance). There are also fine-grained conclusions that can be obtained through this data. Finally, some of the potential bias that arises from social media analysis may be softened in the long run as social network use increases: as we have shown, when a growing number of citizens express their opinions and/or voting choices online, the accuracy of social media analysis increases, provided that Internet users act in a way that is consistent with their statements online. Overall, if we take data from in-person polls to be the gold standard, then so too is this data.

2.3.2 Mixed or Unsuccessful Prediction Methods

<u>Bermingham and Smeaton [2]</u>

I. *Data:* This study was conducted on data from the Irish General Election, 2011. To obtain data, a "Twitter Tracker" developed by a third party was used. Data was segmented into groups based on date to conduct different analyses on the data over time.

II. *Methodology:* This study aimed to model political sentiment effectively enough to capture the voting intentions of a nation during an election campaign. They took an approach of combining sentiment analysis using supervised learning and volume-based measures. Their volume-based and sentiment analysis measures were both predictive. For the volume-based method, a calculation of proportional share of party mentions in a set of tweets for a given time period was used. For the sentiment-analysis method, the team used a prior version of the dataset to come to the conclusion that supervised learning was a more effective method of predictive polling than unsupervised learning. They trained on classifiers with data related to the election, and obtained this data from their dataset by looking for a sample of keywords relevant to the election. They then used annotaters (actual people) to label these tweets internally. Challenges included a surprisingly low level of positive sentiment, finding minority classes in an initial dataset of few examples, and using different metrics for inter-party sentiment vs intra-party sentiment. To do the supervised learning method on this data, they fit a regression to inter- and intra-party measures trained on poll data.

III. *Results:* The lowest amount of error came when using the full dataset as opposed to training on only certain portions of the dataset. In the same vein, as more data was added over time, the error reduced. Their analysis was simply: "Overall, the best non-parametric method for predicting the result of the first preference votes in the election is the share of volume of tweets that a given party received in total over the time period we study. This is followed closely by the share of positive volume for the same time period, which, despite considering only a fraction of the documents by share of volume, approaches the same error. Either overall share of volume of share of positive volume performs best for each dataset. As expected, negative share of voice consistently performs worst, though in some cases rivals the other measures. This is likely due to a correlation with the overall share of volume." In conclusion, though sentiment was able to predict with some accuracy (though lacked this in specific cases for specific Irish parties), it was a usable metric for predicting election results, with the single biggest predictive variable being the volume of tweets.

Gayo-Avello [6]

I. *Data:* A set of tweets obtained leading up to the United States 2008 Presidential elections.

II. *Methodology:* Sentiment analysis methods of O'Connor and Tumasjan are applied to tweets, assigning a voting intention to every individual user in the dataset, along with the user's geographical location. with some slight changes, in order to account for differences in the nature of electoral races (e.g., German elections had 5 major parties all vying for votes, American elections are based on "winner takes all" so a tweet cannot be counted as a vote for both opposing candidates).

III. *Results:* Electoral predictions that were computed for different states instead of simply the whole of the US found that every method examined would have largely overestimated Obama's victory, predicting (incorrectly) that Obama would have won even in Texas. In this sense, it points out that demographic bias in the user base of Twitter and other social media services is an important electoral factor and, therefore, bias in data should be corrected according to user demographic profiles. This research has revealed that data from Twitter did no better than chance in predicting results in the last US congressional elections. We argue that this should be expected: So far, knowledge of the exact demographics of the people discussing elections in social media is scant, while according to the state-of-the-art polling techniques, correct predictions require the ability of sampling likely voters randomly and without bias.

Metaxas [12]

I. *Data:* One dataset belonged to the 2010 United States Senate special election in Massachusetts ("MAsen10"), a highly contested race between Martha Coakley (D) and Scott Brown (R). The other dataset contained all the tweets provided by the Twitter "gardenhose" in the week from October 26 to November 1, the day before the general US Congressional elections in November 2, 2010 ("USsen10"). These two datasets are different. The MAsen10 is an almost complete set of tweets, while USsen10 provides a random sample, but because of its randomness, it should accurately represent the volume and nature of tweets during that pre-election week.

II. *Methodology:* Twitter chatter volume (same method as Tamasjan) and sentiment analysis (extended method from O'Connor, in that they had a subjectivity lexicon that classified words, but each tweet was labeled positive, negative, or neutral, not some combination of all of them).

III. *Results:* Electoral predictions using the published research methods on Twitter data are not better than chance. Given the unsuccessful predictions we report, one might counter that "you would have done better if you did a different kind of analysis". However, recall that they did not try to invent new techniques of analysis: they simply tried to repeat the (reportedly successful) methods that others have used in the past and found that the results were not repeatable.

2.4 Generalizations of Previous Work

There are a few key takeaways from this and a few common threads that make a prediction analysis method successful, based on this literature review. Firstly, generally, the larger the dataset, the more successful the prediction method. This makes sense generally, since larger data can often showcase predominant trends, minimize the importance of flukes, and often are more reflective of the population at large. However, while more data is usually helpful, it can also often amplify the voices of groups who are overly represented in whichever dataset is being used. This caused another common trend among these papers: when the election outcome was "surprising" - i.e., generally not predicted by news outlets covering the election, these prediction methods often reflected the predominant views set forth by news outlets rather than the actual outcomes of the election. My hypothesis for this, which is reflected in several sources, is that those with more controversial or unpopular opinions often do not want to share them publicly on Twitter or other social media websites [16]. Finally, though there seem to be more studies done using sentiment-analysis strategies in predicting election outcomes than not doing so, both sentiment analysis and not-sentiment analysis methodologies have similar success rates across the board. More sentiment analysis studies are being performed because, in theory, they are newer and will thus become more and more accurate, but at this current time that is not the case.

3. METHODOLOGY

In order to answer the question: can we determine which sentiment analysis is best for predicting election outcomes? We must compare so-called successful methods of sentiment analysis on new datasets and see which ones work for various datasets. In order to keep some semblance of coherency through this project (and not pick several different elections from around the world), we specifically looked at the United States democratic primaries during the year 2020 and apply four different sentiment analysis methods to social media data stemming from each of these locations in the weeks leading up to each election.

Social media data were obtained as Tweets from specific locations, namely, the states in which a democratic primary is occuring. These dates were gathered from two weeks prior to the election all the way up until the day before election. This is the dataset that each of these methods will be applied to. Having these various datasets is important because they will each have different features that make them unique, and thus, it is my hypothesis that one method of sentiment analysis will not be the most effective across all of the given datasets. The primaries begin on February 3rd, 2020, and were supposed to continue through June 6th, 2020. However, given the scope of the semester and the broader impacts on the elections due to the Coronavirus, these methods were only tested on states whose primaries went through March 17th, 2020.

This section describes in detail the methodologies used in each paper to predict the elections, and how we interpreted these methodologies to be used to predict the democratic primaries. Small changes had to be made for some of the studies due to differences in elections, and these are noted here. Otherwise, we stuck as close as possible to the actual paper methodology as we were able.

3.1 O'Connor

Paper Methodology

The first paper we attempted to recreate was the O'Connor, Balasubramanyan, Routledge, and Smith paper called *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* This paper attempted to explain correlations between political survey data, specifically, election polls, and Twitter sentiment. They used 1 billion Twitter messages between 2008 and 2009, removing non-english messages, and evaluated two sub-problems within this dataset: first, finding messages related to the topic at hand, and second, deriving an opinion estimation from said messages. To retrieve relevant messages, they simply looked for keywords. Multiple correlation analyses were performed, but we focused on their work pertaining to the 2008 general US presidential election.

In terms of the 2008 elections, the two keywords they used to identify relevant tweets were *obama* and *mccain*. Then, to categorize the tweets into positive and negative, they used a subjectivity lexicon from OpinionFinder (https://mpqa.cs.pitt.edu/opinionfinder/), removing the distinctions between levels of positive (the lexicon's distinctions between weak and strong words were not used). Then, they defined a sentiment score for that day for each candidate based on the tweets.

Finally, these results were compared to the presidential election polls in 2008, among others. For the purposes of this research, we focused on these results, even though other data and polls were used, such as presidential job approval rating, consumer confidence, and economic confidence. By comparing the twitter sentiment results to the presidential election polls, O'Connor et al could produce a correlation coefficient. Interestingly, for this specific poll, the correlation coefficient was the lowest among any of the polls used, at an abysmal -8%. This was in part explained by adding a lag parameter, where sentiment became a predictor of future polling results. When a lag parameter was added and fine tuned, the correlation coefficient became 44%.

Experimental Methodology

Because this paper used polls as the gold standard for accuracy of sentiment analysis results, but this paper focuses on actual election results as the barometer of truth, this experiment was divided into two sections, roughly identical. In the first, we focused on comparing the sentiment analysis results to polling data, and in the second we focused on comparing the sentiment analysis results to the actual election results.

Comparison to Polls

This first section stayed as close to the actual experiment as possible. However, because we were looking at the results over a series of elections and not one general election, instead of taking data from months before the election as the paper did, we took polling data from a week leading up to the election. This data was collected from Politico via RealClearPolitics, the same way that it was collected in the paper. This created an aggregated metric of various polls for each region. One caveat with this is that some regions, especially southern and/or more conservative states, did not have extensive, comprehensive polling because there was not seen to be a competitive race. For example, there was only one poll in Arkansas about the democratic presidential primary, and there were no polls at all in North Dakota. In these states, Biden was the overwhelming favorite among the more conservative democrats who represented the constituent voting base.

Tweets were scraped in the areas of each of the states for the week leading up to the election, and subsequently categorized by sentiment. For each day, a score was totaled using the same metric as in the paper, where the sentiment score x_t on day *t* represented the ratio of positive versus negative messages on the topic, counting from that day's messages:

$$x_t = \frac{count_t(pos.\ word \wedge topic\ word)}{count_t(neg.\ word \wedge topic\ word)} = \frac{p(pos.\ word \mid topic\ word,\ t)}{p(neg.\ word \mid topic\ word,\ t)}$$

Both polls and tweet sentiment values were normalized across the candidates to create metrics that were comparable to one another.

Comparison to Election Results

In this second section, we took Tweets from two weeks leading up to each election and applied the OpinionFinder subjectivity lexicon on each of the tweets. From this, for each candidate we developed a score, which is the same sentiment score used above, representing the ratio of positive versus negative messages on the topic. Because we took data from a longer period of time (instead of day by day), there was a lot more data to work with, and so this data was more representative of the general public opinion of the candidate leading up to the election. We then compared these scores for each candidate with the actual outcome of the election, taking into consideration the percentage of delegates up for grabs won in each state and comparing that to the share of the total sentiment each candidate had.

3.2 Ceron

Paper Methodology

The paper used the Hopkins and King method of supervised analysis (SA), which essentially assigned each text to an opinion category based on the presence of words within the text. The formal methodology for employing this was detailed in another paper, but there exists both software and code that is public and used in both papers. Essentially, the software can produce a probability distribution, P(S), where S represents the word profiles used and D indicates the opinions expressed.

$$P(S) = P(S|D)P(D)$$

Then, it can produce an estimate of P(D), which is the opinion distributed over the posting population. This can be calculated as follows:

$$P(D) = P(S|D)^{-1}P(S) = P_T(S|D)^{-1}P(S)$$

In the paper, each of these probability distributions was created by hand-labeling tweets in order to predict the vote shares of the first round of the 2012 French legislative elections and the French presidential ballot. In this paper, the social media sentiment results closely mirrored the actual results, giving them a very low mean absolute error of 2.38%. Additionally, further models also took into consideration the number of tweets and abstention rates of each area to make these predictions. This resulted in a "remarkable ability for social media to forecast electoral results on average (a careful prediction that could not be simply due to chance)" [17].

Experimental Methodology

For this experiment, we were able to pretty effectively replicate the original experiment. To hand label the data, we used Amazon Mechanical Turk to classify a dataset of 400 tweets each from different states at around 2-3 weeks prior to the election. To ensure that this data was coherent and consistent, we had three different workers label each dataset and also hand classified some of the data ourselves to make sure it matched what the workers were doing. We used this data as the training set for the experiments for each of the states. We were able to use the same software and code that the paper originally used, an R plugin called isaX. The model was trained twice, once using the tweets themselves as actual data to be assigned a sentiment score, and once extracting each word from the tweet and trying to assign a sentiment score to each word. Each word or tweet that we fed into isaX we stemmed and we used the labels from the mechanical turking data as labels for the sentiment data. Each could be classified into strong positive, positive, neutral, negative, strong negative, or N/A (if the tweet was in a foreign language or otherwise incomprehensible to a worker). Then, we created a probability distribution with the results of the sentiment analysis, running two more experiments, one in which we considered the number of tweets, and one in which we did not. We did not consider rates of abstention as there was not a good way to apply that data over the state (unless we did it by neighborhood, and our data was not granular enough to do so). We then compared the results of these experiments to the distribution of delegates in the primary elections in each state.

3.3 Tumasjan

Paper Methodology

This paper attempted to use political tweets to predict the outcome of the German National Election in 2009. There were six parties represented in the German parliament and a group of politicians associated with these parties. In the month during which the tweets were collected, approximately 100,000 relevant tweets were amassed.

To extract the sentiment of each of these tweets, the software Linguistic Inquiry and Word Count (LIWC) 2007 was used. When the paper was published, LIWC was the premier Natural Language Processing software. They identified the following dimensions to profile political sentiment: future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money. The paper found that the degrees of each of these were closer between candidates that had similar political stances. Additionally, in using just the sheer volume of tweets to predict the election result, the mean absolute error of this prediction was 1.65%.

Experimental Methodology

To replicate this experiment, we employed the use of the LIWC software as used in the paper, but we also used a tool called Vader to replicate the experiment as well. Vader is the current technological equivalent of LIWC, which is now seen to be outdated. To quote Hutto and Gilbert in their paper *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,* “VADER retains (and even improves on) the benefits of traditional sentiment lexicons like LIWC: it is bigger, yet just as simply inspected, understood, quickly applied (without a need for extensive learning/training) and easily extended. Like LIWC (but unlike some other lexicons or machine learning models), the VADER sentiment lexicon is gold-standard quality and has been validated by humans” [17]. In addition, we did not have a good metric for measuring how similar candidates were to one another to verify whether or not the results of a similarity test would be accurate. Instead, we took the dimensions that the paper identified for use in LIWC and aggregated them into a sum representing the amount of positive sentiment for a candidate. Certain positive features were added to a score, while the other negative features were subtracted. This gave us a sentiment score for each candidate based on the features identified in LIWC.

3.4 Livne

Paper Methodology

In contrast to the three previous experiments, the Livne paper makes use of twitter data from actual candidates themselves instead of users on the platform. This paper attempted to analyze the 2010 midterm elections, and in doing so used twitter data from almost 700 different users involved in the political arena. This data was collected going back three years from the date of the election.

This paper made heavy use of graphs and graph algorithms to create networks of candidates with edges between them representing how close they were. Candidates had “closer” or better edges if they were more highly connected with one another. For example, if one candidate was following another, that edge would be closer between the two candidates. These graphs were used to determine scores for candidates in different categories, including in/out degree (number of edges in/out of a node), closeness in/out/all (how close candidates were based on in/out/all edges), HITS’ Authority score, and PageRank, measuring the relative importance of nodes in the graph.

Additionally, a Language Model (LM) was built for every single candidate. This involved the following set of calculations:

Functions to note: $tf(t,\ D)$ represents the frequency of a term in each document, $udf(t,\ D_u) = df(t,\ D_u)\ /\ D_u$ represents the document frequency of a term in a set of users’ documents, and $idf(t)\ =\ log(1 + |D|/df(t))$ represents the inverse document frequency.

The initial weight of a term in the user LM is

$$w(t,\ u) = tf_{AVG}(t, D_u)udf(t, D_u)idf(t, D)$$

where

$$tf_{AVG}(t, D) = \sum_{d \in D_u} tf(t, d)/D_u$$

Then, the marginal probability of each term in the full corpus vocabulary is

$$P(t|D) = tf_{AVG}(t, D)udf(t, D)$$

The terms were normalized as follows:

$$P^N(t|D) = \frac{P(t|D)}{(\sum_{t \in V} P(t|D))}\ ;\ W^N(t, u) = \frac{w(t,u)}{(\sum_{t \in V} w(t,u))}$$

The weights were smoothed using a value of $\lambda = 0.001$ as follows:

$$P(t|u) = (1 - \lambda)w^N(t, u) + \lambda P^N(t|D)$$

Finally, the values were normalized.

$$P^N(t|u) = \frac{P(t|u)}{(\sum_{t\in V} P(t|u))}$$

Then, a metric called KL (Kullback-Leibler) divergence was also calculated between the LM for a certain candidate and the LM of their party (KL-party), as well as the LM for a certain candidate and the LM of the entire corpus (KL-corpus). This was calculated as:

$$D^S_{KL}(P_1\|P_2) = \sum_{t\in V} P_1(t)\frac{logP_1(t)}{logP_2(t)} + P_2(t)\frac{logP_1(t)}{logP_2(t)}$$

Finally, variables were also obtained representing the party a candidate was in, if they were an incumbent, if they were in the same party as the seat was held currently, and the total number of tweets, hashtags, replies, and retweets for a candidate.

All of these values were then centralized and a logistic regression model was created, using each of these values as a variable for one piece of training data for the model. The model was trained on some of the data and used these features to predict the outcomes for the other part of the data, the testing data. The outcome was whether or not a candidate won. The most important factors in determining this were the same party, incumbent, indegree, and closeness all, which all had over a 70% accuracy in determining the winner. Using all of these features, the paper was able to obtain a logistic regression model with 88% accuracy on predicting the election outcomes.

Experimental Methodology

To recreate this experiment, we scraped twitter data for each candidate from the two weeks leading up to each election. We did not use location on twitter to weed out tweets, as many candidates could not visit every state before the election, and doing so would have severely limited our pool of data. Instead, we considered the corpus of tweets relevant to each candidate for each state to be the tweets they made two weeks leading up to each election. For most of the features required for the logistic regression model were extracted or calculated directly from the dataset. These included any features related to graphs as well as counts of hashtags or tweets. However, some features we had to make some judgements about, such as the difference between the KL divergence between the corpus and KL divergence from the party. Since all candidates were democrats, they were technically in the same party. To make these different values, we classified each candidate into the further subcategories of “moderate” and “liberal”. Then, we used these values as the party values. This was also helpful in determining whether or not candidates were in the same party as the previous winner -- we compared these to the democratic primaries in 2016 to assess whether the “moderate” or “liberal” candidate won each state that year. Finally, we asserted that Biden was the “incumbent” candidate, as he had previously held office from 2008-2016. Though we had to make these decisions with our dataset in mind, they seemed reasonable and helped us recreate the experiment as closely as we possibly could.

The logistic regression model used was sklearn.linear_model.LogisticRegression from Python. We also used train_test_split to split up the data.

4. RESULTS AND CONCLUSIONS

In order to determine the success of the predictions that each analysis will make, we have to establish the baseline 100% accuracy. Depending on the study used, baseline markers will consist of either or both:

1. Actual election result data.
2. Actual polling data from RealClearPolitics, which aggregates all of the predominant polls together to create one clear benchmark.

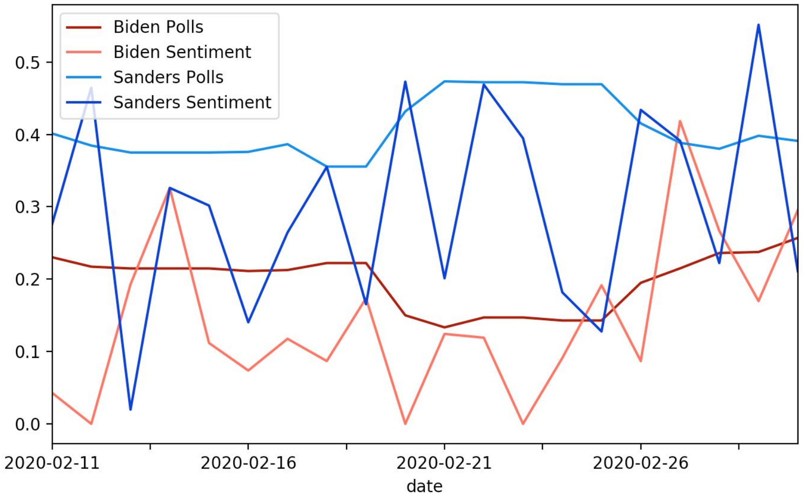
Using both of these will provide a clearer window into whether or not this data is reflective of polling data, actual election results, both, or neither. The following is the analysis of each political prediction method tested, named for the first author on the paper which we recreated.

4.1 O'Connor

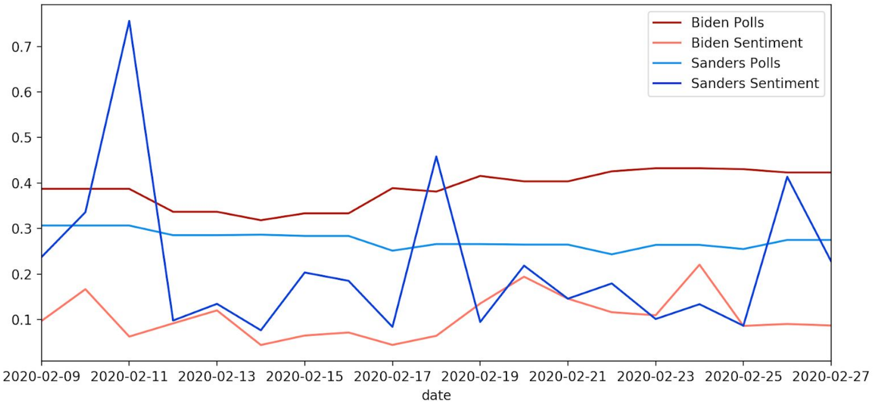
Comparison to Polls

The overwhelming conclusion with regards to polling data was that day by day, this was an ineffective way to estimate polling data. To evaluate these probability distributions of spread between candidates in terms of sentiment and polls, we used two methods. First, we calculated a Pearson's correlation coefficient, as used in the paper. Over all states tested, this value averaged out to be 80%. At first glance, this seems relatively strong, but some of this has to do with the fact that candidates were left in models who had already dropped out of the race or who were still running but barely being considered by voters. When we account for that by removing such candidates, we receive a correlation coefficient value of 63%. Still, this is not terrible considering the fact that at this point in the paper doing the same calculation, the tests had a negative correlation coefficient. However, we argue that this is not the best metric to calculate this data. This can be shown visually. The following are a few graphs in various states, comparing polling data to the collected sentiment data. The same-colored lines represent the same candidates, and should be correlated if we are to assume that sentiment and polling data are positively correlated.

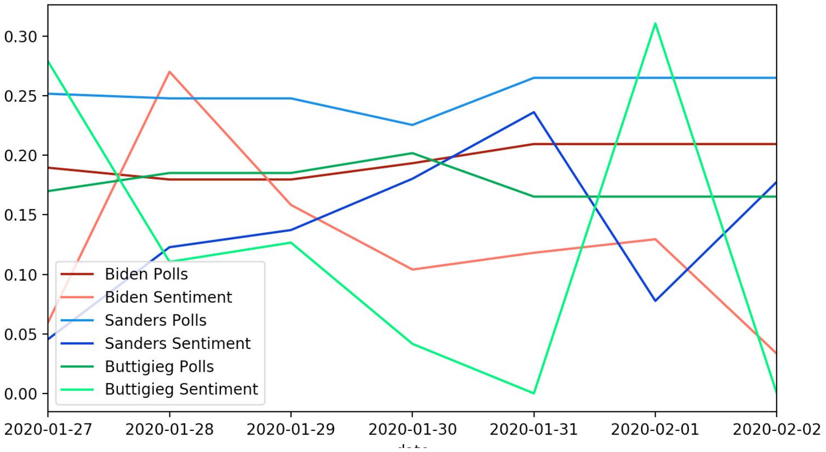
California, Sanders won



South Carolina, Biden won



Iowa, Buttigieg won



As we can see, Pearson's correlation coefficient might explain some of these small trends, but visually it does not appear that these probability distributions are actually meaningful. To further verify this, we used a chi-squared goodness of fit test, which explicitly measures whether probability distributions are correlated.

State by state results:

Iowa;	X^2(9, N = 1266) = 0.06, p < .05;	Model rejected
New Hampshire;	X^2(9, N = 9489) = 0.09, p < .05;	Model rejected
Nevada;	X^2(6, N = 7567) = 0.07, p < .05;	Model rejected
South Carolina;	X^2(6, N = 9027) = 0.12, p < .05;	Model rejected
Alabama;	X^2(4, N = 6839) = 0.12, p < .05;	Model rejected
Arkansas;	X^2(4, N = 186) = 0.03, p < .05;	Model rejected
California;	X^2(4, N = 6857) = 0.04, p < .05;	Model rejected
Colorado;	X^2(4, N = 198) = 0.08, p < .05;	Model rejected
Maine;	X^2(4, N = 6868) = 0.04, p < .05;	Model rejected
Massachusetts;	X^2(4, N = 6832) = 0.01, p < .05;	Model rejected
Minnesota;	X^2(4, N = 207) = 0.09, p < .05;	Model rejected
North Carolina;	X^2(4, N = 284) = 0.09, p < .05;	Model rejected
Oklahoma;	X^2(3, N = 233) = 0.01, p < .05;	Model rejected
Tennessee;	X^2(4, N = 221) = 0.03, p < .05;	Model rejected
Texas;	X^2(4, N = 244) = 0.04, p < .05;	Model rejected
Utah;	X^2(4, N = 6854) = 0.03, p < .05;	Model rejected
Vermont;	X^2(4, N = 221) = 0.21, p < .05;	Model rejected
Virginia;	X^2(4, N = 197) = 0.04, p < .05;	Model rejected
Idaho;	X^2(2, N = 11708) = 0.05, p < .05;	Model rejected
Michigan;	X^2(2, N = 207) = 0.01, p < .05;	Model rejected
Mississippi;	X^2(2, N = 187) = 0.04, p < .05;	Model rejected
Missouri;	X^2(2, N = 223) = 0.0, p < .05;	Model rejected
Washington;	X^2(1, N = 182) = 0.01, p < .05;	Model rejected
Arizona;	X^2(2, N = 195) = 0.0, p < .05;	Model rejected
Florida;	X^2(2, N = 6727) = 0.19, p < .05;	Model rejected
Illinois;	X^2(2, N = 197) = 0.01, p < .05;	Model rejected

Here, our null hypothesis was that the polling proportions in each category were consistent with the sentiment analysis data. Across all states tested, for every state, every single null hypothesis was rejected. The difference here is that while the correlation coefficient just measures the linear correlation between two variables, the goodness of fit calculation compares the expected result to the measured result, attempting to determine if one can reasonably explain the other.

In this experiment, we arguably reported better outcomes than the original experiment, as our correlation coefficient was higher than both the regular and the adjusted correlation coefficient reported in the paper. However, we have also shown that these trends are not explicative of any meaningful correlation between sentiment and polls. In fact, we argue that the paper did not go far enough in assessing its results, and potentially could have drawn some incorrect conclusions based on the data provided. Some critical facts to note are as follows:

1. The Twitter dataset used was not nearly as comprehensive as the one in the paper. We only scraped about a month's worth of data, whereas the paper scraped data going back a year. The reason we took a smaller dataset for these elections (and continued to do so for the other methodology implementations as well) is because for a majority of the states, there was not heavy Twitter traffic for those primaries until about a month before the election. Though there was substantial election discourse on Twitter, the Tweets were often not geotagged or simply not coming from the primary states we were interested in, so they were not possible to use in our dataset because it was unclear in what election (if any) that Twitter user might be voting in. It is possible that some of the correlation here could simply be a fluke of polls going up and down. Additionally, we only used tweets that had a recognizable geolocation as being in the area of the election, which seriously dampened our pool of tweets.
2. As the paper noted as well, the lexicon used to train the data only had about 1600 words in it, and often mis-tagged certain parts of speech with labels that did not necessarily correspond to the meaning of the data. A more advanced lexicon such as the one used in Veilkovitch et al. would have significantly improved these results.
3. In the paper, text sentiment was shown to be a partial leading indicator of polls, whereas for our data we did not adjust any of the timeframes to take into account this leading factor. The main reason for this is that we used a dramatically smaller twitter dataset and timeframe, so adjusting for a few days or even a week would have significantly impaired our results. Additionally, the point of this experiment was to see if sentiment would be an indicator of poll results, and to find a meaningful and consistent method of predicting election outcomes. In the paper, different lengths of time were used to see how much of a time-lag was needed for sentiment to correlate with actual polling data. We did not want to arbitrarily adjust this length of time to get the best possible output, because it seemed arbitrary for each election.

Comparison to Election Results

With our chi-square goodness of fit assessment, this data was moderately more accurate as compared to the polling data. To analyze the results, we did the same as above, doing both a chi-square goodness of fit measurement and a correlation coefficient measurement.

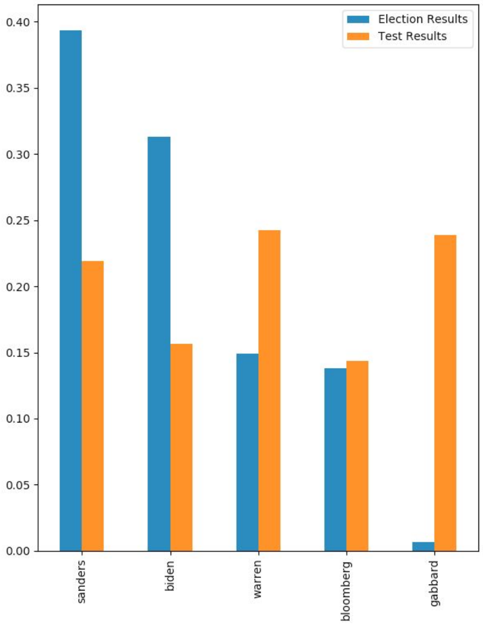
State by state results:

Iowa;	X^2(10, N = 1266) = 162.43, p < .05;	Model rejected
New Hampshire;	X^2(9, N = 9489) = 65.82, p < .05;	Model rejected
Nevada;	X^2(9, N = 7567) = 227.92, p < .05;	Model rejected
South Carolina;	X^2(9, N = 9027) = 425.91, p < .05;	Model rejected
Alabama;	X^2(9, N = 6839) = 374.47, p < .05;	Model rejected
Arkansas;	X^2(9, N = 186) = 76.84, p < .05;	Model rejected
California;	X^2(10, N = 6857) = 185.82, p < .05;	Model rejected
Colorado;	X^2(9, N = 198) = 59.26, p < .05;	Model rejected
Maine;	X^2(9, N = 6868) = 213.29, p < .05;	Model rejected
Massachusetts;	X^2(10, N = 6832) = 191.89, p < .05;	Model rejected
Minnesota;	X^2(10, N = 207) = 67.62, p < .05;	Model rejected
North Carolina;	X^2(10, N = 284) = 213.18, p < .05;	Model rejected
Oklahoma;	X^2(10, N = 233) = 41.99, p < .05;	Model accepted

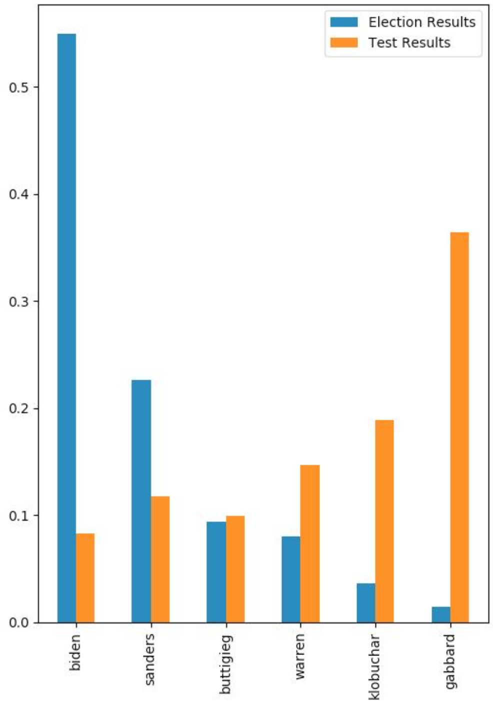
Tennessee;	$X^2(10, N = 221) = 124.93, p < .05;$	Model rejected
Texas;	$X^2(10, N = 244) = 40.04, p < .05;$	Model accepted
Utah;	$X^2(10, N = 6854) = 148.4, p < .05;$	Model rejected
Vermont;	$X^2(9, N = 221) = 68.3, p < .05;$	Model accepted
Virginia;	$X^2(10, N = 197) = 143.49, p < .05;$	Model rejected
Idaho;	$X^2(10, N = 11708) = 396.83, p < .05;$	Model rejected
Michigan;	$X^2(9, N = 207) = 69.16, p < .05;$	Model rejected
Mississippi;	$X^2(9, N = 187) = 40.73, p < .05;$	Model accepted
Missouri;	$X^2(10, N = 223) = 131.22, p < .05;$	Model rejected
North Dakota;	$X^2(10, N = 274) = 195.9, p < .05;$	Model rejected
Washington;	$X^2(10, N = 182) = 41.56, p < .05;$	Model accepted
Arizona;	$X^2(10, N = 195) = 117.46, p < .05;$	Model rejected
Florida;	$X^2(10, N = 6727) = 540.83, p < .05;$	Model rejected
Illinois;	$X^2(10, N = 197) = 25.08, p < .05;$	Model accepted

The average correlation was 33%. The chi-square goodness of fit calculation yielded 26% of winners to be chosen correctly and 22% of probability distribution models to be accepted.

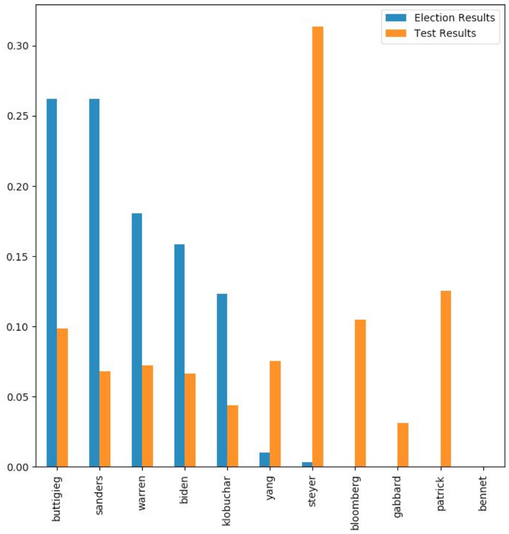
California, Sanders won



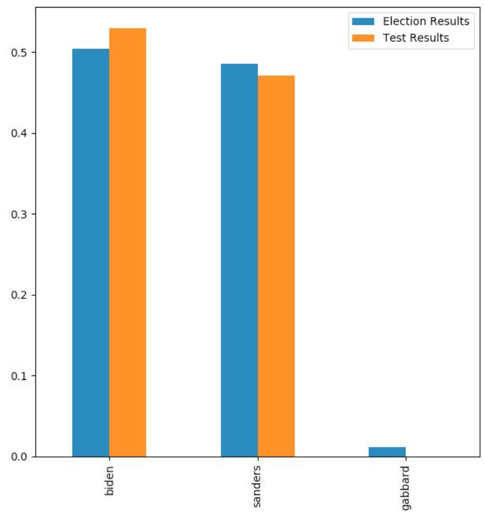
South Carolina, Biden won



Iowa, Buttigieg won



Washington, Sanders won



This experiment arguably was a more thoughtful analysis of sentiment data as compared to the polling methodology above. First of all, a fraction of the models calculated were accepted as valid probability distributions for this dataset, meaning that the sentiment analysis predictions were actually meaningful. Because many more tweets were able to be used, it is far less likely that one or two tweets significantly impacted the dataset and caused inaccurate results. However, this 22% is a far cry from the 44% toted by the paper itself and even the 73.1% correlation after smoothing the data. One major difference between this election and the one in the paper was that there were multiple candidates to consider, so increased twitter chatter about one specific candidate might not have necessarily been meaningful, it could have just been a very active supporter. This theory is supported by the fact that the only time the model correctly predicted the winner is when Biden won the election, which often was easy to predict in areas where there wasn't any chatter about any of the other candidates.

Additionally, some of the less mainstream candidates never had anything negative tweeted about them, likely because they were generally out of the election discourse generally. We attempted to cap these values so that they wouldn't be extreme outliers in the data, eventually settling on giving them a score that was equal to the number of positive tweets they'd gotten. Still, this reflects that the actual number of tweets was sometimes a non-factor in this ranking system. For example, a non-mainstream candidate with 15 positive tweets and no negative tweets would be more favorable than a mainstream candidate with 600 positive tweets and 150 negative tweets, whose score would be 4.

Finally, we can see from the sampling above that many of the elections and the sentiment data correlated well with one another, but many of them did not (especially the more challenging to predict elections with no clear front runner). Thus, we conclude that the methodology set forth by the O'Connor paper is not an acceptable way to assess either the outcomes of polling data or the elections at hand, especially when the Twitter dataset is only thousands of tweets and when there are more than two candidates.

4.2 Ceron

The main takeaway here was that the tremendous amount of hand-labeling of data required to create a comprehensive, large dataset was difficult to achieve. Because our dataset was so small and many tweets were categorized as neutral (effectively not adding anything to the sentiment data), when the isaX algorithm tried to classify tweets, it often was unable to produce results that weren't completely neutral. Thus, using the data we had hand-labeled, we were unable to get a sentiment analysis evaluation for several states. This model was trained twice as explained before, once with the tweets themselves being classified, and once with the words within each tweet being classified. While the method in which we broke down each word was able to give us more results (and was able to classify every state), these classifications were not more accurate and the overall results between the two methods remained roughly the same.

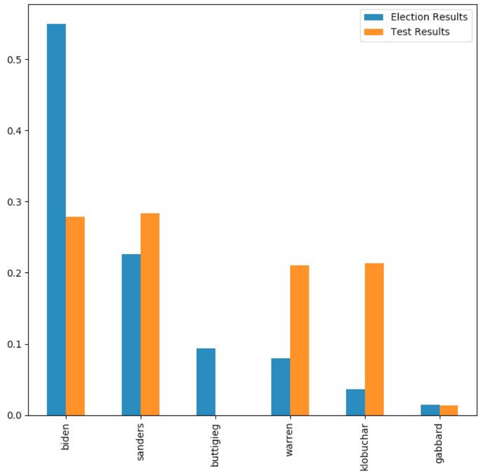
Moreover, when hand-labeling the data, there was a skew toward labeling tweets as positive, and thus with fewer negative data points, many of the training data tweets were not labeled as negative.

State by state results:

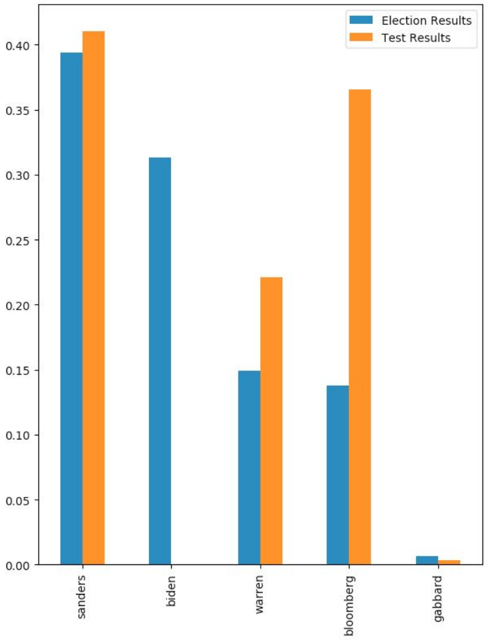
Iowa;	X^2(10, N = 3895) = 304.73, p < .05;	Model rejected
New Hampshire;	X^2(8, N = 23487) = 30.58, p < .05;	Model accepted
Nevada;	X^2(5, N = 12501) = 44.32, p < .05;	Model accepted
South Carolina;	X^2(5, N = 24068) = 100.6, p < .05;	Model rejected
Alabama;	X^2(4, N = 11256) = 785.43, p < .05;	Model rejected
California;	X^2(4, N = 11319) = 804.35, p < .05;	Model rejected
Maine;	X^2(4, N = 11318) = 1186.13, p < .05;	Model rejected
Massachusetts;	X^2(4, N = 11230) = 115.8, p < .05;	Model rejected
North Carolina;	X^2(4, N = 823) = 892.16, p < .05;	Model rejected
Utah;	X^2(4, N = 11173) = 357.85, p < .05;	Model rejected
Idaho;	X^2(2, N = 22017) = 5.73, p < .05;	Model rejected
Florida;	X^2(2, N = 11487) = 24.8, p < .05;	Model accepted

Taking this into consideration, after performing a chi-squared goodness of fit test with each model, we were able to create models to predict the results of each election, 25% of which were accepted. If we take into consideration the states where no results could be predicted, this number drops to 11%. Similarly, of these accepted models, the winner was correctly predicted each time. Interestingly, when we took into consideration the number of tweets in the dataset, these numbers remained the same, but the likelihood of predicting the correct winner of each election shot up significantly, to 70% (as opposed to 25% earlier). This shows us that the mere number of tweets is a significant indicator of political success, as the Ceron paper explained as well.

South Carolina, Biden won



California, Sanders won



Thus, we can see that the methodology set forth by the Ceron paper might be an acceptable way to assess the elections, provided that the sheer volume of training data was increased tenfold. Even so, our preliminary findings indicate that a dataset increased in size might not necessarily give more accurate predictive results. To test this, we used the same language classifier from the O'Connor paper and created a test dataset of Tweets scraped in the weeks prior to the Tweets that were scraped for each election. This is the same dataset that we used for the mechanical turk classification, but for this dataset, we did not have to pare it down, as a machine was doing the classification. This was not necessarily the method set forth by the paper, but it was an additional test done just to verify that increasing the size of the dataset would lead to better results, provided that the mechanical turk data was similar to the O'Connor classification for

Tweets. We found that with this modified dataset, we produced models that were accepted 33% of the time and correctly predicted the winner of each election 44% of the time, a substantial step up from the results with the severely reduced dataset.

So far from this dataset, we have to dispute the conclusions in the Ceron paper, as so few of our models were deemed acceptable when compared to polling data. Because a tremendous amount of technology exists wherein classifying words and sentiment is heavily trained and tested, we assert that this is a better option to use than hand-labeling tweets. In fact, initially before we had hand-labeled tweets, we used OpinionFinder to classify each of the tweets for the training set. These results were much more comprehensive, as we were able to label thousands of tweets pertaining to each election. These distributions gave us an average chi-squared goodness of fit value of 37%, not outstanding, but better than labeling tweets by hand, which cost time and money and arguably added very little value to these results.

4.3 Tumasjan

Of the three experiments conducted here: using Vader, LIWC, and number of tweets to identify election results, Vader performed the worst. This is pretty surprising given that Vader is considered the industry standard for language classification, but looking at this a little closer, we can come to see why. 37% of the probability distribution models created by Vader were accepted using a chi squared goodness of fit test, and it was only able to pick the correct winner of the election 7.4% of the time. The LIWC model performed better, with 52% of the probability distribution models accepted and the winner predicted correctly 33% of the time.

A good example of this is from the tweet dataset used to predict the outcomes of the South Carolina primaries. In these primaries, Biden won, and the (un-normalized) predictive scores were as follows:

	LIWC	**Vader**
Biden	463650.33	1.30
Sanders	319139.7	2.86

Normalized, this becomes:

	LIWC	**Vader**
Biden	.59	.31
Sanders	.41	.69

Why, then, if Vader is considered industry standard, was it not predicting the outcomes of these elections correctly? This is likely because Vader provides a holistic sentiment based on many specific, nuanced traits. In contrast, the LIWC analysis selected for dimensions specifically relevant to political aspirations and prowess. In this specific example, both Vader and LIWC (which was broken down into specific traits) agreed that Sanders had more tweets with positive sentiment and thus in terms of positive emotion alone, should have been the front-runner. This is reflected in the Vader scoring. However, LIWC takes many more traits into account, including clout, authenticity, tone, focus, and money-focus, many of which (in this specific example) swung in favor of Biden. So, the Tumasjan paper refined the results of LIWC with more accuracy than the standard Vader analysis by pre-selecting for these subtle traits.

This is obviously subject to scrutiny, as there is no set of traits that are deemed “political” over other traits. However, generally, the dimensions selected here fit into this category. Additionally, in calculating the LIWC value, the number of tweets was a factor that played into the sentiment score. This was not true of the Vader results. To try to correct for this, we did an additional test where we weighted the Vader scores by the number of tweets about each candidate. Still, when using the number of tweets as a factor for sentiment by multiplying the score obtained by number of tweets for the Vader test, only 48% of models were accepted and only 15% could predict the winner correctly.

The best model of the three was the one that just used the number of tweets about a candidate as a measure of political preference.

State by state results using LWIC and number of tweets:

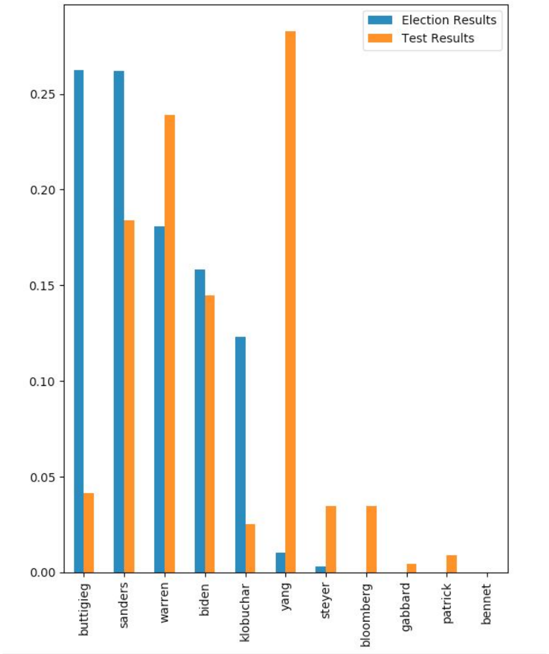
Iowa;	X^2(10, N = 145) = 127.62, p < .05;	Model rejected
New Hampshire;	X^2(8, N = 2528) = inf, p < .05;	Model rejected
Nevada;	X^2(5, N = 470) = 12.78, p < .05;	Model accepted
South Carolina;	X^2(5, N = 2376) = 29.4, p < .05;	Model accepted
Alabama;	X^2(4, N = 504) = 328.42, p < .05;	Model rejected
Arkansas;	X^2(4, N = 12) = 102.85, p < .05;	Model rejected
California;	X^2(4, N = 506) = 47.2, p < .05;	Model accepted
Colorado;	X^2(4, N = 13) = 37.96, p < .05;	Model accepted
Maine;	X^2(4, N = 506) = 73.21, p < .05;	Model rejected
Massachusetts;	X^2(4, N = 508) = 72.45, p < .05;	Model rejected
Minnesota;	X^2(4, N = 14) = 164.26, p < .05;	Model rejected
North Carolina;	X^2(4, N = 83) = 62.6, p < .05;	Model accepted
Oklahoma;	X^2(4, N = 18) = inf, p < .05;	Model rejected
Tennessee;	X^2(4, N = 13) = inf, p < .05;	Model rejected
Texas;	X^2(4, N = 15) = inf, p < .05;	Model rejected
Utah;	X^2(4, N = 506) = 14.72, p < .05;	Model accepted
Vermont;	X^2(4, N = 12) = inf, p < .05;	Model rejected
Virginia;	X^2(4, N = 12) = inf, p < .05;	Model rejected
Idaho;	X^2(2, N = 1414) = 1.72, p < .05;	Model rejected
Michigan;	X^2(2, N = 20) = 14.76, p < .05;	Model accepted
Mississippi;	X^2(2, N = 17) = 124.14, p < .05;	Model rejected
Missouri;	X^2(2, N = 21) = 44.14, p < .05;	Model accepted
North Dakota;	X^2(2, N = 35) = 48.15, p < .05;	Model accepted
Washington;	X^2(2, N = 14) = 27.41, p < .05;	Model accepted
Arizona;	X^2(2, N = 28) = 6.64, p < .05;	Model accepted

Florida;	$X^2(2, N = 816) = 22.07, p < .05$;	Model accepted
Illinois;	$X^2(2, N = 25) = 25.11, p < .05$;	Model accepted

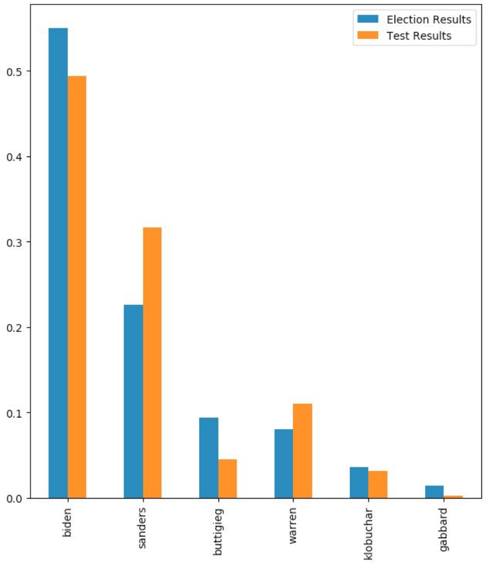
This model produced a probability distribution that was accepted by a chi square goodness of fit test 56% of the time, and correctly predicted the winner of each election 51% of the time. This is a far cry from the mean absolute error of 1.65% cited in the paper, so these results certainly do not match the success of the paper in this specific technique. Still, up to this point, this model was the most successful of those tested, and it did not rely on one shred of word classification.

The following graphs are from the quantity of tweets model.

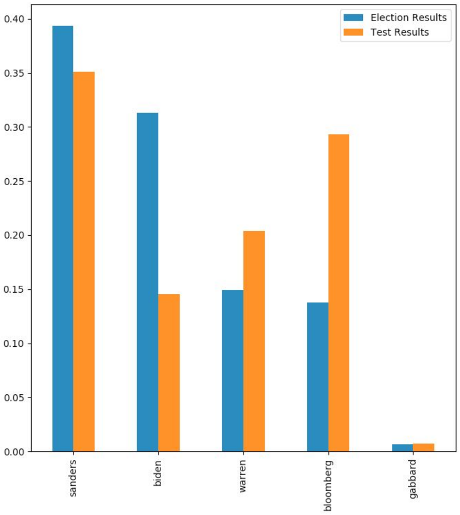
Iowa, Buttigieg won



South Carolina, Biden won



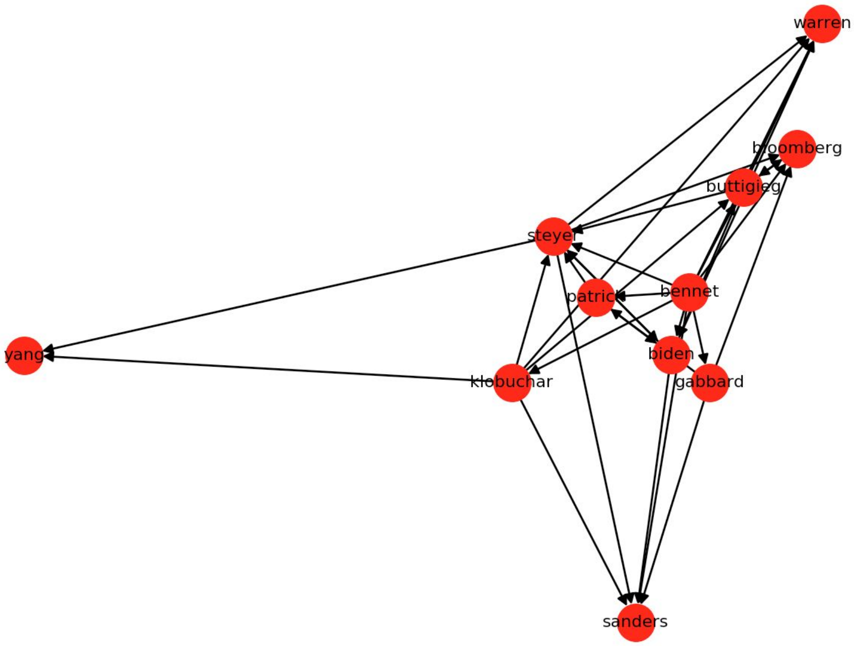
California, Sanders won



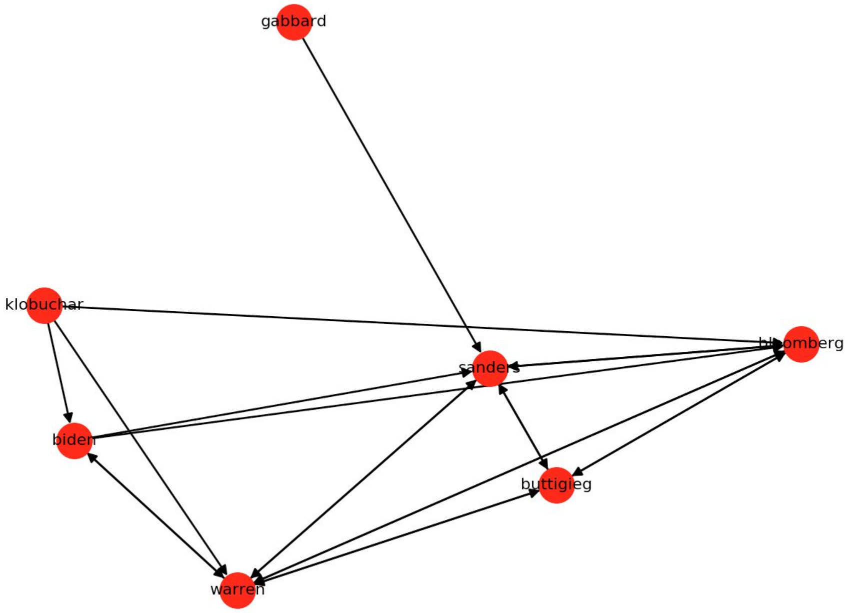
4.4 Livne

Here are some of the graphs generated between candidates, where the closer the candidates are to one another, the more similar they are. More graphs in/out of the candidate mean that they are more connected to the rest of the candidates.

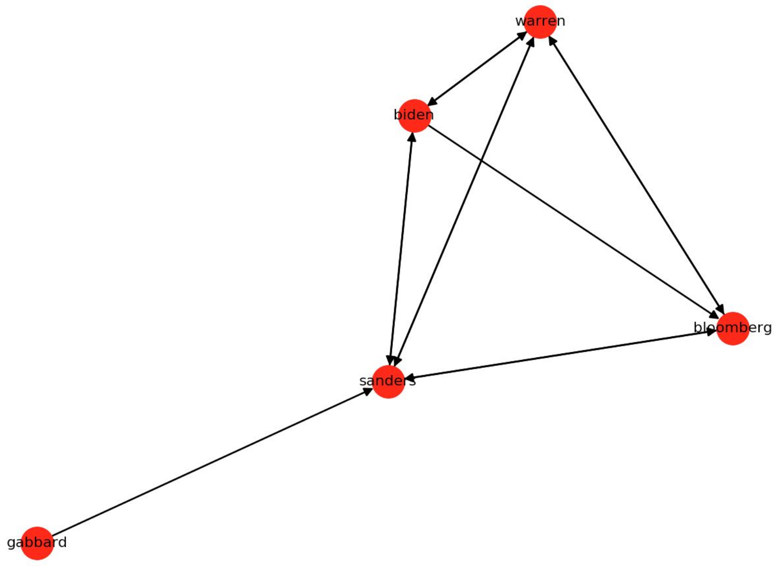
Iowa, Buttigieg won



South Carolina, Biden won



California, Sanders won



This bag of words model was incredibly effective in analyzing the data. In further exploration, the paper wanted to assign sentiment to each of the values as yet another metric, but from what we've seen with the previous projects, we suggest that the bag of words model might be more effective without the sentiment analysis. In addition, using just the candidates themselves eliminated much of the unpredictability with users' tweets, which potentially could have thrown off the data. The major downside to this model is that the outcomes of each election were known beforehand in order to accurately train the data. Because "state" was a variable within the logistic regression model as well, but we used all data from the two week interval it is unclear how this model would perform on data in future states without having the chance to train on that kind of data accurately.

```
Linear Regression Results
=======================================================================================
Dep. Variable:                    won   R-squared (uncentered):                   0.726
Model:                            OLS   Adj. R-squared (uncentered):              0.659
Method:                 Least Squares   F-statistic:                              10.79
Date:                Mon, 06 Apr 2020   Prob (F-statistic):                    4.66e-12
Time:                        10:53:20   Log-Likelihood:                          5.6813
No. Observations:                  76   AIC:                                      18.64
Df Residuals:                      61   BIC:                                      53.60
Df Model:                          15
Covariance Type:            nonrobust
====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
closeness_in          0.0281      0.042      0.666      0.508      -0.056       0.112
closeness_out         0.0617      0.036      1.735      0.088      -0.009       0.133
closeness_all        -0.0954      0.052     -1.838      0.071      -0.199       0.008
pagerank              1.1536      0.679      1.699      0.094      -0.204       2.511
hits                 -0.0127      0.303     -0.042      0.967      -0.619       0.594
in_degree            -0.0133      0.034     -0.393      0.696      -0.081       0.054
out_degree            0.0188      0.032      0.591      0.557      -0.045       0.082
```

```
kl_divergence       -0.0004      0.000     -0.894      0.375     -0.001       0.000
incumbent            0.4415      0.189      2.337      0.023      0.064       0.819
same_party           0.1308      0.060      2.179      0.033      0.011       0.251
tweets               0.0032      0.004      0.816      0.417     -0.005       0.011
replies_total     1.369e-05   5.13e-06      2.668      0.010   3.43e-06    2.39e-05
replies_average     -0.0020      0.001     -2.993      0.004     -0.003      -0.001
retweets_total   -5.476e-06   1.89e-06     -2.905      0.005  -9.25e-06   -1.71e-06
retweets_average     0.0008      0.000      3.543      0.001      0.000       0.001
```

In a divergence from the paper, the number replies and retweets and the averages of these values seemed to be the variables that most explained whether or not a candidate would win a race. Second to those were incumbent, same party, followed by closeness all and page rank. The other variables had p-values that were much too high to rely on for predicting the outcomes of the election. The R squared value for this regression, or the percentage of variation that is predicted for by our independent variables, is 72.6%. Finally, our logistic regression classifier accuracy was 85%, only three percentage points lower than the accuracy results in the paper. Thus, this methodology has substantial merit, this paper valid, and this model be further explored in data collection.

5. CONCLUSION

Of the four models tested, there was only one that came close to accurately and consistently predicting the outcome of the election: the logistic regression model proposed in the Livne paper. This model predicted the outcomes of elections with 85% accuracy across the board, in a field with varying numbers of candidates in each election and various political preferences in each region. Of course, it must be noted that this model required the results of the election for earlier political contests to be able to predict outcomes for each election. This is a tremendous advantage, and this core reason was likely the model was able to perform so well. However, this advantage is not the sole reason this test outperforms other models. Although it does use this extra data, we found that the most compelling factors that influence election outcomes for this data were number of likes and number of retweets on candidates' tweets. These were pure, quantitative measures that had nothing to do with the state itself, but seemed to yield tremendous impact on predicting the outcomes of the elections. Even the next highest predictive factors were still indicators regardless of state: incumbency and same party as previous victor. This was the methodology set forth in the paper, so it was what we followed. Thus, the "predictive power" so to speak of this model remains to be tested, and suggestions for this can be found in the next section, further investigation.

However, there are two additional takeaways from these results. Firstly, sentiment analysis of tweets produced by citizens, in this study, are not the best indicator of political preference, no matter how you analyze them. This is likely due to a number of factors, including the same group of people comprising most of the user-base of twitter, as well as critical, negative emotions often being commonplace in politics and not necessarily an indicator of worse performance. An easy example of this is the constant mainstream media's disparaging of President Trump, even though his approval rating has been relatively consistent for his past three years in office.[15] Secondly, political candidates themselves on Twitter are a good indicator of later political performance, and quantitative metrics such as number of likes and retweets on posts or bag of words language models can provide meaningful election predictions. Based on these results, sentiment analysis, by any of the measures used in these models, is not a good indicator of political performance. In fact, sentiment about a candidate is likely overall a poor measure of political performance, as candidates who are mentioned more often across the board, positively or negatively, tend to fare better in polls and elections. This was a statement that most of the papers we reviewed concurred with. Thus, to predict an election, we suggest a model that does not use sentiment as a guiding factor.

Finally, this is a complete breakdown of all four models and whether or not they predicted the election in each state for side by side comparison:

State	**O'Connor**	**Ceron**	**Tumasjan**	**Livne**
Iowa	Reject	Reject	Reject	Accept
New Hampshire	Reject	Accept	Reject	Accept
Nevada	Reject	Accept	Accept	Reject
South Carolina	Reject	Reject	Accept	Accept
Alabama	Reject	Reject	Reject	Accept
Arkansas	Reject	Reject	Reject	Accept
California	Reject	Reject	Accept	Accept
Colorado	Reject	Reject	Accept	Accept
Maine	Reject	Reject	Reject	Accept
Massachusetts	Reject	Reject	Reject	Accept
Minnesota	Reject	Reject	Reject	Reject
North Carolina	Reject	Reject	Accept	Accept
Oklahoma	Accept	Reject	Reject	Accept
Tennessee	Reject	Reject	Reject	Accept
Texas	Accept	Reject	Reject	Accept
Utah	Reject	Reject	Accept	Accept
Vermont	Accept	Reject	Reject	Reject
Virginia	Reject	Reject	Reject	Accept
Idaho	Reject	Reject	Reject	Accept

Michigan	Reject	Reject	Accept	Accept
Mississippi	Accept	Reject	Reject	Accept
Missouri	Reject	Reject	Accept	Accept
North Dakota	Reject	Reject	Accept	Accept
Washington	Accept	Reject	Accept	Accept
Arizona	Reject	Reject	Accept	Accept
Florida	Reject	Accept	Accept	Accept
Illinois	Accept	Reject	Accept	Accept

This side-by-side breakdown shows that there is not necessarily a "right" way or a "right" method to use on a particular dataset, as theorized at the beginning of this paper. It is clear that one of these methods (Livne) consistently trumps the others in terms of accuracy of predictions. However, there are a few notable takeaways. Firstly, the O'Connor method (SA classification with OpinionFinder) performed best in southern states. Additionally, later elections with fewer candidates were easier to predict across the board, but this was especially true with the Tumasjan model (LIWC scoring). Finally, the states in which the Livne model (Logistic Regression) did not accurately predict the outcome all did not have Biden as a winner; in fact, a majority of the time Livne was predicting Biden to be the winner, and Biden was the candidate who ultimately prevailed and became the Democratic nominee in 2020.

6. FURTHER INVESTIGATION

To enhance the Livne model to its utmost potential, it needs to be tested on data that it has not been trained on. Namely, it needs to be tested and validated on data before an election occurs. The main flaw of the Livne model is that it requires training data from the election once it has already taken place. To expand on this study, the Livne model needs to be finely tuned and tweaked so that only the most important dimensions of data are included, and it also needs to be validated on future datasets. There needs to be a way for each election to train on a previous dataset, perhaps the data from the last time or couple of times that election occurred, and generate a consistent and accurate prediction for data for upcoming elections. Otherwise, this model is no more accurate than any of the other sentiment analysis models.

Finally, all code and documentation for this project can be found here:
https://github.com/lenarenshaw/thesis

7. REFERENCES

[1] B. Alexander. "A Bayesian Model for the Prediction of United States Presidential Elections." 2019. doi: 10.1137/17s016166, URL:

https://www.semanticscholar.org/paper/A-Bayesian-Model-for-the-Prediction-of-United-Alexander/98e1c1db4e7c523f5f8b175af552e6b02cb033df

[2] A. Bermingham and A. Smeaton, "On Using Twitter to Monitor Political Sentiment and Predict Election Results," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP*, 2011, pp. 2-10.

[3] A. Ceron, L. Curini, S. M. Iacus, and G. Porro. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," *New Media & Society*, vol. 16, no. 2, 2013, pp. 340-358.

[4] A. Clark, C. Fox, and S. Lappin. *The Handbook of Computational Linguistics and Natural Language Processing*. Hoboken, NJ: Wiley-Blackwell, 2013, pp. 271-295.

[5] J. Clement. "Number of Social Media Users Worldwide 2010-2021," *Statista*, 2019 [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed: Dec 10, 2019].

[6] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj. "Limits of Electoral Predictions Using Twitter." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 490-493.

[7] A. Jungherr, P. Jürgens, and H. Schoen. "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to 'predicting elections with twitter: What 140 characters reveal about political sentiment,'" *Social Science Computer Review*, 2011, pp. 230-234.

[8] S. Kim and E. Hovy. "Crystal: Analyzing Predictive Opinions on the Web," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 1056-1064.

[9] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic. "The Party is Over Here: Structure and Content in the 2010 Election," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 201-208.

[10] C. Lui, P.T. Metaxas and E. Mustafaraj. "On the Predictability of the U.S. Elections through Search Volume Activity," *Proceedings of the IADIS International Conference on e-Society*, 2011.

[11] *Measuring Digital Development: Facts and Figures 2019*. [Online]. Available: http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx. [Accessed: 10-Dec-2019].

[12] P. T. Metaxas, E. Mustafaraj and D. Gayo-Avello. "How (Not) to Predict Elections," *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 165-171.

[13] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. “From tweets to polls: Linking text sentiment to public opinion time series.” *In Proc. of 4th ICWSM*, 2010, pp. 122-129.

[14] A. Tumasjan, T. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment.” *In Proc. of 4th ICWSM*, 2010, pp. 178-185.

[15] NateSilver538. “How Popular Is Donald Trump?” FiveThirtyEight, April 15, 2020. https://projects.fivethirtyeight.com/trump-approval-ratings/.

[16] Mcdonald, R. I., Salerno, J. M., Greenaway, K. H., & Slepian, M. L. (2020). Motivated secrecy: Politics, relationships, and regrets. Motivation Science, 6(1), 61–78. doi: 10.1037/mot0000139

[17] Hutto, C. J., & Gilbert, E. (2014). http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf. Georgia Tech. Retrieved from http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf