UNITED STATES
MILITARY ACADEMY

WEST POINT, NEW YORK

HONORS THESIS

PREDICTING POLLING USING TWITTER SENTIMENT IN PRESIDENTIAL PRIMARIES

by

John Lee

May 2020

Thesis Advisor:	Lieutenant Colonel Kevin Cummiskey
Co-Advisor:	Lieutenant Colonel Christopher Weld
Second Reader:	Major Mario Andriulli

PREDICTING POLLING USING TWITTER SENTIMENT IN PRESIDENTIAL PRIMARIES

John Lee
Cadet, U.S Army
B.S., United States Military Academy, 2020

Submitted in partial fulfillment of the
requirements for the degree of
BACHELOR OF SCIENCE
in **MATHEMATICAL SCIENCES**
with Honors
from the
UNITED STATES MILITARY ACADEMY
May 2020

Author: John Lee

Advisory Team: Lieutenant Colonel Kevin Cummiskey
Thesis Advisor

Lieutenant Colonel Christopher Weld
Co-Advisor

Major Mario Andriulli
Second Reader

Colonel Tina Hartley
Chair, Department of Mathematical Sciences

TABLE OF CONTENTS

Abstract

Can sentiment analysis methods supplement polling as measures of public opinion in presidential primary elections? We investigate using a dataset of tweets and 2020 Democratic primary polls over four months. Using text lexicons, we extract measures of sentiment from these tweets associated with four major Democratic candidates. Granger Causality and Correlation analysis reveal weak associations between each of the candidates' sentiment/poll pairings at time lags under five days. We fit VAR and ARIMA models to find insufficient evidence that sentiment adds useful forecasting information to polling at short lags. Sensitivity analysis reveals that with optimized parameters we can improve sentiment-based models. Case studies on the Iowa Caucuses and South Carolina Primary suggest that sentiment may be useful in uncovering underlying opinion trends when they exist. These findings highlight the potential of inexpensive sentiment collection as an alternative to polling, but also emphasize the need for refined methods of sentiment analysis for better performance.

Introduction

After more than a century of continuous use, polling remains the conventional method for measuring the standing of candidates in upcoming elections. While polls typically do a sufficient job of assessing public opinion ahead of elections, they remain difficult and expensive to administer. A standard telephone poll of 1000 respondents can cost tens of thousands of dollars to run (O'Conner et al., 2010). The cost-related shortcomings of traditional polling techniques may amplify in volatile, primary-style elections where candidates enter and exit the field regularly. This volatility may compound through the relative sparsity and questionable scientific rigor of primary polling in individual states (Rakich, 2019). For example, it is difficult to determine whether the difference between December 2019 and March 2020 polls from a well-respected Iowa polling firm, David Binder Research, was reflective of actual opinion-changes versus a result of the change in candidates offered in each poll. Political operatives and voters alike could stand to benefit from additional ways of measuring public opinion in these contexts.

The onset of widespread text-based social media has given rise to a widely available source of millions of opinions every day. Particularly in election-seasons, these opinions are often increasingly associated with some political sentiment (Tumasjan et al., 2010). Can this information be harnessed to augment or replace political polling in primary elections? Similar studies have already shown the potential of Twitter textual data for augmenting polling in German Federal elections (Tumasjan et al., 2010) and in various national economic measures, approval ratings, and electoral elections (O'Conner et al., 2010). Primary-style elections present a new challenge; the volatility of candidate support and frequent entry and exit of candidates makes the process difficult to model. Nevertheless, the success of similar models in other contexts suggests we attempt to assuage the difficult undertaking of primary polling with sentiment analysis methods.

In this paper, we analyze the potential of Twitter-based sentiment analysis as a supplement or a substitute for polling in primary-style elections. We examine the 2020

Democratic Presidential Primary through the lens of four of its top contenders: Vice President Joseph Biden, Senator Bernard Sanders, Senator Elizabeth Warren, and Mayor Peter Buttigieg. Tweets and polls are gathered daily (or in the case of polls, as often as they are released) over a period from early November to late February, two days before the South Carolina Primary. We gather data on a national level for the United States and on a state level for Iowa and South Carolina, two crucial early voting states in the nomination process. We extract sentiment from tweets through processing key words in each tweet, which we then score through external lexicons. We then create time series for sentiment and polling of each candidate.

With our data formatted into eight separate time series, one sentiment and one polling for each candidate, we return to the motivating question for our study: Is Twitter sentiment useful in forecasting polling in primary-style elections? We propose polling as a response to sentiment by the reasoning that sentiment is the more instantaneous opinion estimator. Polls take days to conduct before they are released, while sentiment gathers on demand. We hypothesize then that sentiment *leads* polling or polling *lags* sentiment. Of course, it is entirely possible that polling can also lead sentiment given the benefits of being a front-runner such as name-recognition and momentum. We are interested only in investigating the former possibility, however, given our desire to provide a suitable alternative or supplement for traditional polling methods, and it would not make sense to provide a ‘slower’ alternative.

We first investigate basic relationships between polling and sentiment time series. Correlation analysis and Granger Causality reveals weak relationships for each candidate between sentiment and polling at time lags under five days. Next, we model the relationship between polling and sentiment using auto-regressive and moving average techniques, ultimately finding insignificant forecasting improvement of polling from using Twitter sentiment. We modify parameters of sentiment weighting, lexicon choice, and forecast windows to demonstrate that sentiment analysis methods can be significantly improved. Case studies on Iowa and South Carolina reveal sentiment may be useful in improving forecasts through detecting underlying trends. These findings reinforce the need for improved sentiment analysis methods, but also provide hope for the potential of sentiment analysis to bolster traditional polling in primary election settings.

Literature Review

Early forecasting models utilized large-scale factors to make sweeping election predictions. The advent of widely available data through social media and more advanced collection methods has allowed researchers to reach new levels of depth in political analysis. In this literature review, we explore the transition from traditional forecasting models to sentiment-based models and scope how this study fits in the latter category.

The seminal papers in political forecasting predominantly revolved around assessing ‘structural’ factors as explanatory variables for election outcomes. In Lewis-Beck et al., (1984) the study used broad factors traditionally assumed by political analysts to be causal in U.S presidential elections. These variables consisted of national economic performance, candidate political experience, and incumbent approval rating. The study then aggregated these factors into a multivariate model with the popular vote result as the dependent variable. Within the multivariate model, economic performance and presidential approval variables were both statistically significant in predicting election results. The model had an absolute error only slightly larger than that of the last Gallup poll before the election in the study, implying the potential for substituting analysis for traditional polling. This study established that political and economic factors can statistically predict election results.

Other studies explored election prediction on a smaller scale. Linzer et al. (2013) relied exclusively on presidential election polling but used an aggregate technique at the state level to continually update the model. Similarly, Jerome et al. (2012) used political and economic factors with an added effect from the long-term electoral trends in each state to make state level election forecasts. From there, the study aggregated the state predictions into a national forecast. These studies incorporated more of the variation that goes into a national level election based on individual state results. Breaking election forecasting down to lower levels of analysis presents an attractive continued method of research.

Political forecasting studies using more traditional methods are an essential base for all work in the field. However, modern data collection techniques and increased data availability correspond with an increased interest in studies of social media analysis. Since its genesis in the early 2000s, social media has rapidly expanded in popularity. For political forecasting purposes, Twitter has proven to be a useful social media platform. Twitter offers a wealth of accessible and compact data in uniform sizes. Despite its relative novelty, researchers have conducted numerous studies on political analysis through Twitter.

Studies in Twitter-based political forecasting often begin with more general goals than predicting election outcomes. They instead focus primarily on whether there is an association in the first place between Twitter data and political sentiments. In Tumasjan et al. (2010), the study used German federal elections in 2009 to analyze whether Twitter is used as a political forum, whether Twitter data reflects overall political sentiment, and whether Twitter data can predict the popularity of political parties. The study set the standard for several pioneering techniques and conclusions in Twitter data analysis. First, the study established effective methods of analyzing sentiment in Twitter data. The volume of party mentions alone correlated to party success substantially. The frequency of words appearing with these party names were shown to give a complete picture of political sentiments. Next, this study came to the significant conclusion that Twitter can predict electoral success nearly as well as polls. Specifically, when using the volume of tweets as a predictor of election results, the mean absolute error of the model was comparable to that of many mainstream polls. Overall, this study set the stage for future political analysis using Twitter by finding significant relationships between Twitter sentiment and political discourse, as well as Twitter forecasts and election results.

Further studies continued to build on the prospect of election forecasting using Twitter. Bermigham et al. (2011) delved into the insight sentiment analyses can provide for election forecasting, this time using the Irish general election of 2011. This study assigned groups of words for specific Irish political parties in order to classify Twitter mentions as either positive or negative. The study then investigated the performance of two sentiment categories: positive and negative, and compared them to a volume-based approach. Interestingly, the volume-based approach performed the best in predicting election results, followed by the positive and then negative categories, respectively. The study also revealed some limitations to sentiment analysis-based research. First, sentiment can be notoriously fickle; which the study demonstrated through huge sentiment swings in the days leading up to the election. Another limitation is that sentiments usually have far smaller sample sizes than volume-based approaches. Overall, this study dug further into the merits and drawbacks of sentiment analysis in Twitter, demonstrating its potential for further research.

Other studies investigated the potential of Twitter sentiment measures as an alternative or supplement to polling measures, which will be the focus of this paper. O'Connor et al. (2010) used more straightforward measures to assess U.S political sentiment between 2008 and 2009, focusing primarily on the correlation between sentiment and rolling poll data over this period. During the period it investigated, the study examined the similarity between Twitter sentiments and consumer confidence indices, presidential approval rating, and presidential election polling between President Barack Obama and Senator John McCain. The Twitter text sentiments were classified by using a method like that of Tumasjan et al. (2010). Tweets mentioning the economy, presidential approval, or either presidential candidate were classified as positive or negative. The study then aggregated Twitter sentiment into rolling averages to coincide with that of national level polling, approval ratings, or confidence indices. Ultimately, the study found that a substantial correlation exists between Twitter sentiment as a leading indicator of polling on a national scale.

It is fair to wonder at this point why researchers would bother substituting Twitter data for polling in the first place, even if it can offer similar predictability. O'Connor et al. (2010) noted that Twitter data offers the advantage of substantially lower costs in comparison with polling. In political forecasting, reducing the expensive process of data collection can allow for more detailed analysis. Additionally, this study brought up an area of research that is also suggested in Bermigham et al. (2011): the potential for weighting 'likes' or 'retweets' in sentiment analysis. Weighting would allow for a greater depth of inferences from political data that can be difficult to achieve with only polling. The benefits of gathering Twitter data are evident from these studies.

This paper will investigate the potential of Twitter sentiment to support or substitute for traditional polling in primary-style elections. Tumasjan et al. (2010) investigated the predictive strength of Twitter data versus that of various polls in multi-party federal elections. Meanwhile, O'Conner et al. (2010) established the utility of Twitter sentiment as a leading indicator of polling and other national level metrics on a broad scale. One area the field of research has lacked in thus far is the use of Twitter to examine primary elections where candidates and their bases of support are particularly volatile. This is to say, how effectively can Twitter sentiment augment or even replace the problematic prospect of polling primary elections? We answer this question and related sub-queries on the accuracy of sentiment-augmented forecasting versus polling forecasting in the critical early-voting states Iowa and South Carolina.

Data

There are two central bodies of data associated with this study: Twitter textual data or 'tweets', and polls from several different organizations. Since time shapes the context of election opinion estimation, it is necessary to first establish the periods over which we collected data. Unlike traditional elections, primary-style elections have rolling results over multiple state election days. Early-voting states often play a crucial role in deciding primary elections because they can establish momentum (Abramowitz, 1989). Two important states in U.S presidential primaries are Iowa and South Carolina. The 2020 Iowa Democratic Caucuses took place on February 3, 2020. The 2020 South Carolina Democratic Primary took place on February 29, 2020. We began gathering data on November 6, 2020 and concluded on February 27, 2019, two days before the South Carolina primary and 3 days before Mayor Pete Buttigieg dropped out of the race.

Twitter

Twitter is a social media platform which revolves around the concept of 'micro-blogging,' meaning each tweet on Twitter is limited to just 140 characters. Tweets can also include pictures, videos, emojis, and hyperlinks to other tweets or websites. Users can convey approval of tweets through 'favoriting' or 'retweeting,' the latter of which re-posts the tweet on the feed of the new user. Due to the terse nature of tweets, the widespread popularity of Twitter, and the relative ease of collecting data using the Twitter API, Twitter is an ideal candidate for political sentiment analysis.

In this study, we collect approximately three million tweets and retweets between November 6, 2019 and February 27, 2020, mentioning Vice President Biden, Senator Warren, Senator Sanders, or Mayor Buttigieg. Tweets are queried from three locations: The United States as a whole, Iowa, and South Carolina. Each day, we gather approximately 26,000 tweets and retweets in total. The variation is due to an improved collection process implemented on December 21st, 2019 which allowed us to gather tweets three times a day rather than twice. Fifty percent of these are gathered at the national level, twenty five percent are gathered from Iowa, and twenty-five percent are gathered from South Carolina. The tweets from each of these locations are split evenly between mentioning each of the four candidates, though it should be noted that many of these tweets overlap among candidates, since political tweets often mention multiple candidates in the same tweet. We handle these tweets through counting them for each candidate they mention. Tweets are queried for through keyword association with the candidates. For instance, Senator Biden's tweets are identified through searching for tweets that contain: 'Biden' or 'Joe Biden 'or '@JoeBiden' or '#JoeBiden.' This is a relatively simple method of querying, and certainly comes with some limitations. For example, our queries would classify a tweet mentioning a person named 'Steve Warren' as a tweet associated with Senator Warren. Despite these limitations, applying loose keyword association criteria assures the capture of as many legitimate mentions of the candidates as possible.

Polling

The traditional method of opinion estimation relies on polling. For this study, we collect polls over the same 114-day time period as Twitter sentiment in order to conduct future analysis using the two methods. There are countless polling organizations in the United States with varying levels of reputability. Relying on a few, established polls in this study, like Gallup or Pew Research Center, becomes problematic due to the relative infrequency with which these organizations conduct polls. A single polling organization may only release polls of the same location monthly or less. We circumvent these issues using polling aggregation sources. Political analysis websites like 'fivethirtyeight.com' and 'realclearpolitics.com' collect and categorize polls from a broad swath of organizations as they are released. We select the former (fivethirtyeight.com) due to their regularly updated and easily accessible database of election polling.

In order to make an adequate comparison with daily Twitter sentiment, we assume that the more polls gathered for this study, the better. For this reason, though fivethirtyeight.com grades pollster reputability on an 'A' through 'F' scale, we opt to include every poll that meets our contextual criteria. These criteria specify that we gather polling over the same geographical and time parameters as we collect tweets. We gather national level Democratic primary polls for Vice President Biden, Senator Warren, Senator Sanders, and Mayor Buttigieg from November 6, 2019 to February 27, 2020. We do the same for state level polls in Iowa and South Carolina. We gather all polls on their final day.

There are several factors we must deal with to consolidate polling data such that we have one value for each candidate on each day in our 114-day period. First, we eliminate 'head-to-head' polls from our data set, which pit pairs of candidates against one another. These polls artificially spike the polling numbers of candidates since it eliminates the usual broad field of options that other polls consider. Another issue in our dataset is missing values. Even considering every poll we can, there are some days where no organizations release a poll within our geographic criteria. Approximately 35% of days in our set were not polled across our three geographic locations. We assume polling remains relatively stable over periods of less than a week for each candidate, so we fill missing values with the last known value. The greatest polling gap we encounter in our datasets is six days. We also encounter the opposite problem, where multiple organizations release polls meeting our criteria on the same days. We solve this problem through simply averaging the poll values for each candidate on days where multiple polls are conducted.

Automated Data Collection

We automated the collection of Twitter data to avoid some of the difficulties associated with manual collection processes. The Twitter API limits the range we can collect past tweets to only the past 7-9 days. We circumvent this issue through gathering tweets in real-time and storing them in a database. The Twitter API also limits the quantity of tweets that can be gathered in 15-minute periods. We initially created a script that gathers tweets according to our desired parameters (four candidates, three locations) and ran the script manually at two staggered intervals throughout the day. Though we maintained this system from November 6th to December 29th, two issues subsequently arose: running the script manually is tedious and local storage of the data files was limited. We solved both issues through implementing an automated system on an external data-collection server called the DSP (Data-Science Playground). We set up a CRON job using the Linux operating system to automatically run our script at three staggered time periods each day. Files were then automatically stored on the server and could be retrieved locally whenever needed. Automating this process saved time and resources in collecting Twitter data but did result in one data-loss incident. The server went down on December 30th and was not restored until January 5th, resulting in 7-days of lost Twitter data. We fill these missing days using the last known sentiment values on December 29th, and further mitigate this issue through smoothing methods described in the Methodology.

Methodology

We aim to utilize the following established measures of time series analysis to determine whether sentiment is useful in forecasting polling:

- Sentiment Analysis
- Data Visualization
- Temporal Smoothing
- Correlation Analysis
- Granger Causality
- ARIMA and VARIMA Modeling

Sentiment Analysis

Once our corpus of Twitter data is scraped, we employ numerous data cleaning techniques to attain sentiment values for each candidate on each day of our 114-day period. We exclude retweets since we want our data to represent as broad a sample of opinions as possible, and our data set would be dominated by relatively few tweets with many retweets otherwise. We also strip words which give little indication of sentiment from our tweets called ‘stop words.’ These words include articles like ‘a,’ ‘the,’ ‘and’ as well as non-descript adjectives like ‘this’ or ‘that.’

We use pre-made sentiment lexicons to match words in each tweet with associated sentiments. In this study, we employ two premade lexicons: *bing* from Bing Liu et al., and *AFINN* from Finn Arup Nielsen. The *bing* lexicon categorizes 6786 words in a binary fashion as ‘positive’ or ‘negative.’ The *AFINN* lexicon categorizes 2477 words on a scale of -5 to 5, with -5 being the most negative and 5 being the most positive. In the case of the *bing* lexicon, we assign a sentiment value to each tweet through summing the ‘positive’ (+1) and ‘negative’ (-1) words in each tweet. We do the same with the *AFINN* lexicon, with the only difference being a greater range of potential values to sum. With both lexicons, we then ‘standardize’ the tweets as either -1 (negative) if their value is less than 0, 0 (neutral) if their value is 0, or 1 (positive) if their value is greater than 0. The reason we scale values down is that we do not want individual tweets to have disproportionate effects on candidate sentiment, which could occur if a tweet happens to contain several words in either of these lexicons.

One evident weakness in these sentiment analysis methods is that they are remarkably inflexible. For instance, consider the imaginary tweet: “Joe Biden’s speech was not good.” The *bing* lexicon recognizes only the word ‘good’ as a positive word from this sentence, so our simple algorithm would classify this word as a 1 (positive) overall. Clearly, this is a false classification, which occurred for the simple reason that our algorithm ignores the negating adverb ‘not,’ which effectively reverses the meaning of ‘good.’ We mitigate these misclassifications through adjusting our algorithm using bi-grams (pairs of consecutive words). We separate each tweet into bi-grams and reverse the value of successive lexicon words if the first word in each pair is a negating word: ‘not,’ ‘no,’ ‘never,’ or ‘without.’ Finally, in the case of both lexicons, we sum the sentiment values of tweets for each candidate on each day.

Data Visualization

Data exploration methods in time series analysis enable a general understanding of the data before further analysis. Three plots in time series analysis are particularly useful: the data itself, the Auto-Correlation Function (ACF), and the Partial Auto-Correlation Function (PACF). Plotting the data itself is used to identify the general trends of the data. Plotting the data itself can be enhanced through employing built-in smoothers from programming software like Locally Estimated Scatterplot Smoothing (LOESS). LOESS is a weighted least squares regression of locally estimated subsets of data determined by a k-nearest neighbor algorithm. The ACF gives the lagged correlations between a time series and itself at specified lags. This can be useful in interpreting the structure of a time series, and subsequently, if it is suitable for a variety of models. The PACF gives the lagged correlations between a time series and itself at specified lags, conditionally on the intermediary lags. More specifically, for a correlation between time series a_t and itself at a_{t-3}, the PACF takes into account the correlations between a_t and a_{t-2}, a_{t-1} as well. The PACF is also useful for determining model orders.

Temporal Smoothing

In their current state, our sentiment time series and to a lesser extent, our polling time series have extremely high day to day variance. O’Conner et al. emphasize the importance of temporal smoothing techniques to mitigate this variation before modeling with time series. Ordinarily, a simple moving average algorithm of our observations is sufficient to smooth out ‘white noise:’

$$MA_t = \frac{x_{t-k+1} + \cdots + x_t}{k}$$

where MA_t is the moving average at day t, x is an observation (either sentiment or polling), and k is the smoothing ‘window’ over which values are averaged (O’Conner et al., 2010). There is no statistical test to determine the optimal value of k for our purposes. We seek to select a parameter that balances between smoothing out noise that could negatively impact our models, but not smoothing out valuable information. We choose to smooth on a weekly basis since this is an interpretable unit of time that smooths over the largest gaps of our missing data (7 days) while not approaching the broad trends observed on a monthly scale.

Correlation Analysis

The sample cross-correlation (CCF) is a basic test for determining if there is an association between two time-series. The CCF of two time series a_t and b_t calculates the set of sample correlations at h lags of a_t, denoted as a_{t+h} (Penn State Eberly College of Science, 2020). For negative values of h, we are indicating the correlation between time series a at a time before t and time series b at time t. In this case we say a_{t+h} lags b_t. For negative values of h, we are indicating the correlation between time series a at a time after t and time series b at time t. In this case we say a_{t-h} leads b_t (Penn State Eberly College of Science, 2020). If we let Twitter sentiment for candidate i be a_{ti} and polling for candidate i be b_{ti}, we can use the CCF to examine significant correlations of different lags of sentiment and polling. Significant leads of sentiment against polling reveals potential predictors of polling, a necessary first step before we can build more complex models down the line.

Granger Causality

Granger causality is a statistical test that determines if one time series is likely to influence change in another. Like CCF, this presents another potential key indicator for predicting polling using leads (past lags) of sentiment. Consider once again two time series: a_t and b_t. If we are testing to see if a_t granger-causes b_t then we create two models M_1 and M_2 both of which predict b_t. M_1 uses only past lags of b_t as explanatory variables, while M_2 uses past lags of b_t and a_t as explanatory variables (Granger, 1969):

$$M_1: b_t = \beta_0 + \beta_1 b_{t-1} + \cdots + \beta_k b_{t-k} + \varepsilon$$

$$M_2: b_t = \beta_0 + \beta_1 b_{t-1} + \cdots + \beta_k b_{t-k} + \alpha_1 a_{t-1} + \cdots + \alpha_k a_{t-k} + \varepsilon$$

We seek to determine if M_1 adequately explains b_t or if M_2 is superior. A Wald test is used to simulate and test the following hypotheses:

$$H_0: a_i = 0\ for\ all\ i\ in\ [1, k]$$

$$H_a: a_i \neq 0\ for\ some\ i\ in\ [1, k]$$

If we obtain a sufficiently 'low' p value from the Wald test (typically less than 0.05), we reject the null hypothesis and find that a_t granger-causes b_t, meaning that statistically a_t and past values of b_t are more useful in predicting future values of b_t than past values of b_t alone. While this is not a strong enough test to establish true causation, it is adequate to establish one time series is useful in forecasting future change in another. If sentiment time series granger cause polling time series, we gain stronger evidence that we can model this relationship.

ARIMA and VARIMA

Correlation and Granger-Causality tests allow us to establish a relationship between time series but fail to provide adequate methods for forecasting future behaviors. We turn to Auto-Regressive Integrated Moving Average (ARIMA) and its multivariate extension, Vector Auto-Regressive Integrated Moving Average (VARIMA) models to meet these objectives.

ARIMA

In an ARIMA model, an Auto-Regressive term is a past lag of the time series a_t multiplied by a coefficient. Thus, we say a first-order AR model is denoted by the equation:

$$a_t = \alpha_0 + \alpha_1 a_{t-1}$$

where α_0 and α_1 are coefficients. A second-order AR model would include a term with lag $t - 2$, with each successive order continuing in this way (Penn State Eberly College of Science, 2020).

Integrative refers to the 'differencing' applied to the time series in the ARIMA model. One of the assumptions underpinning ARIMA models is trend stationarity. Put simply, this refers to the general trend of the time series being constant over time (neither increasing nor decreasing). If a trend is present in the data, differencing can eliminate it. Sometimes, a time series variable a_t requires successive differences to achieve trend stationarity. Differencing order matches the number of required differences to achieve trend stationarity. We denote a first-order differencing, z_t, for a univariate time series a_t (Penn State Eberly College of Science, 2020):

$$z_t = a_t - a_{t-1}$$

The last components of ARIMA models are Moving Average terms. These terms are like Auto-Regression terms, but instead of lagged time series multiplied by coefficients, they are past 'white noise' or random-shock errors multiplied by coefficients. Let w_t, be a random-shock error that is identically, independently distributed with a normal distribution about a mean of 0 for a time series a_t. Then we designate a Moving Average first-order model (Penn State Eberly College of Science, 2020):

$$a_t = \mu + w_t + \theta_1 w_{t-1}$$

where μ is the mean value of the time series, and θ are coefficients.

We designate ARIMA models, therefore, by indicating the order of each of the three components. For example, an ARIMA(1,1,1) model is differenced once to achieve stationarity, and has AR(1) and MA(1) components. We identify the 'right' ARIMA model for a given univariate time series through observing the time series data itself, the Auto-Correlation Function (ACF) of the data, and the Partial Auto-Correlation Function (PACF) of the data. The first identifies the need for differencing order. The latter two identify the order for AR and MA terms through significance tests. Maximum Likelihood Estimation methods are used to find the coefficients in the model.

VARIMA

VARIMA models extend ARIMA models into multivariate time series scenarios. The only difference is rather than fitting lagged AR and MA terms of one time series to predict itself, we consider lagged terms from itself and other time series. This can get rather complicated as we increase VARIMA model order, since we consider a large amount of terms. Complex multivariate time series models can lead to problems with both interpretability and overfitting. For this reason, multivariate time series analyses often restrict models to AR terms, leaving us with VAR models. This change is reflected in prominent time series analyses packages in R, such as Bernard Pfaff's *vars*.

We still need to account for trend-stationarity in VAR models. This can be achieved by either appropriately differencing each time series in the model or fitting the trend as its own variable. If we choose the latter process, then a VAR(1) model considering two time series variables a_t, and b_t is as follows (Penn State Eberly College of Science, 2020):

$$a_t = \alpha_0 + \alpha_1 a_{t-1} + \beta_1 b_{t-1}$$

$$b_t = \beta_0 + \alpha_1 a_{t-1} + \beta_1 b_{t-1}$$

The same processes are used to fit coefficients in this model as in ARIMA modeling.

Evaluation Metrics

There are numerous metrics to evaluate ARIMA and VAR model performance. For this study, we select Mean Absolute Error (MAE) for model evaluation. We select MAE because it is a commonly used time series analysis metric which preserves the units of the response variable. Preserving the units of the response variable makes MAE easily interpretable to our study question. The formula for MAE is as follows (Hyndman et al., 2018):

$$MAE = \frac{\sum_{t=1}^{n} |y_t - x_t|}{n}$$

where *n* is the number of observations in our data, y_t are ARIMA/VAR predictions, and x_t are actual values.

Later we compare model performance in projecting results for the Iowa Caucuses and South Carolina Primary. We select the Sum of Squared Errors (SSE) for this task since it is the simplest way to compare aggregated prediction errors for a fixed election day. The formula for SSE is as follows (Archdeacon, 1994):

$$SSE = \sum_{t=1}^{n} (x_t - y_t)^2$$

where x_t are actual election results for each candidate and y_t are the model's predicted values on election day.

Results and Analysis

Data Visualization

Our data consists of eight times series -- a sentiment and polling time series for each of our four candidates. We begin by running a LOESS smoother through each time series to get an idea of broad trends in the data. The gray-bars on each side of the smoother represent 2% confidence intervals for the weighted least squares regression of locally estimated subsets of data determined by a k-nearest neighbor algorithm. We spare further details of LOESS smoothing methods because this is not the smoothing method we will use in our time series analysis (we seek a more-stringent smoothing technique to capture more variance in our data) but it is useful in examining the overall trends in our data. We first examine sentiments based on the *bing* lexicon, which will be used for the initial model of our study due to its greater sample size of words compared with the *AFINN* lexicon.

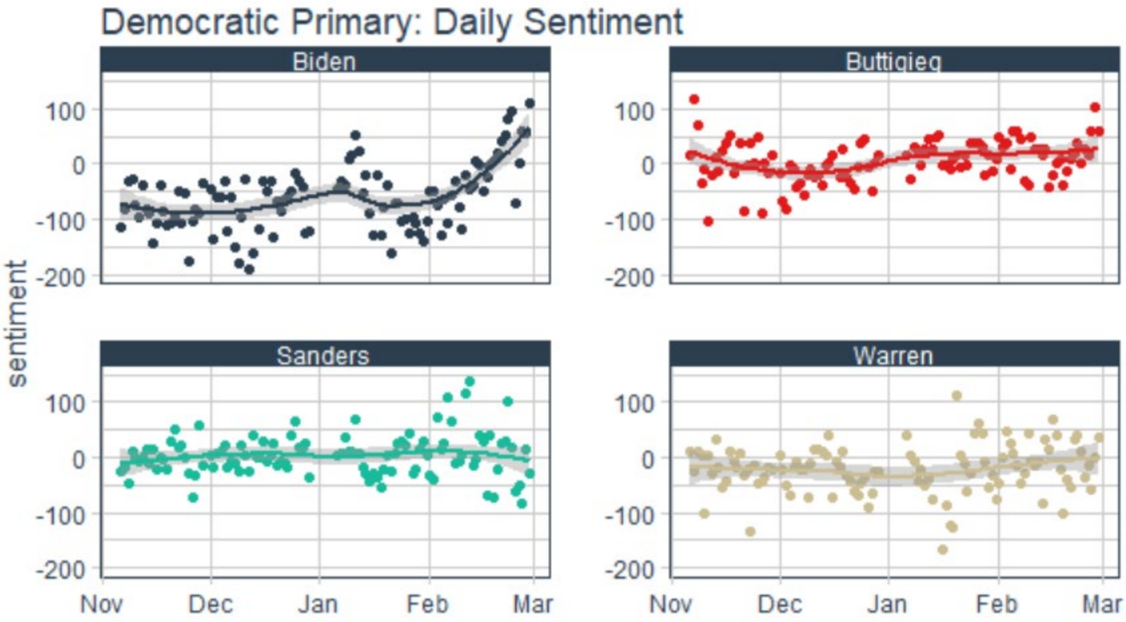


Figure 1

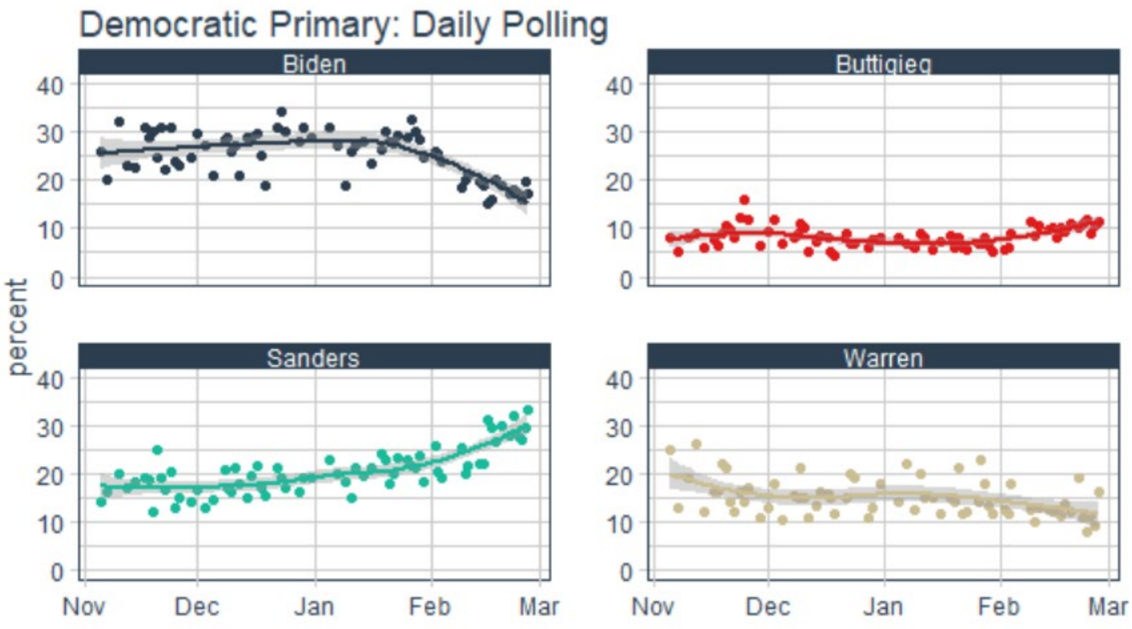


Figure 2

There appears to be broad negative correlation on a monthly scale in some instances, particularly in January and February with Vice President Biden. This implies sentiment may not be powerful enough to overcome broad structural factors in primary elections. However, we are still interested on comparing sentiment and polling on a smaller scale.

Besides comparing the candidate time series visually, these plots are also useful for identifying the trend stationarity of our data. While the trends of some of the time series appear to generally remain stationary over time (Sanders, Warren sentiment and Buttigieg polling), the majority seem to have at least some increasing or decreasing trend over time. Biden's sentiment and Sanders's polling series have particularly sharp increasing trends. Down the line, we will need to deal with this lack of stationarity in our modeling.

It can also be helpful at this point to observe the ACF and PACF plots for each time series. ACF plots display the degree to which a time series is correlated with itself, considering all effects between lags. PACF plots display the degree to which a time series is correlated with itself, separating the effects of each lag. Both are useful for determining AR or MA order in later modeling. Displayed below are Biden's respective plots.

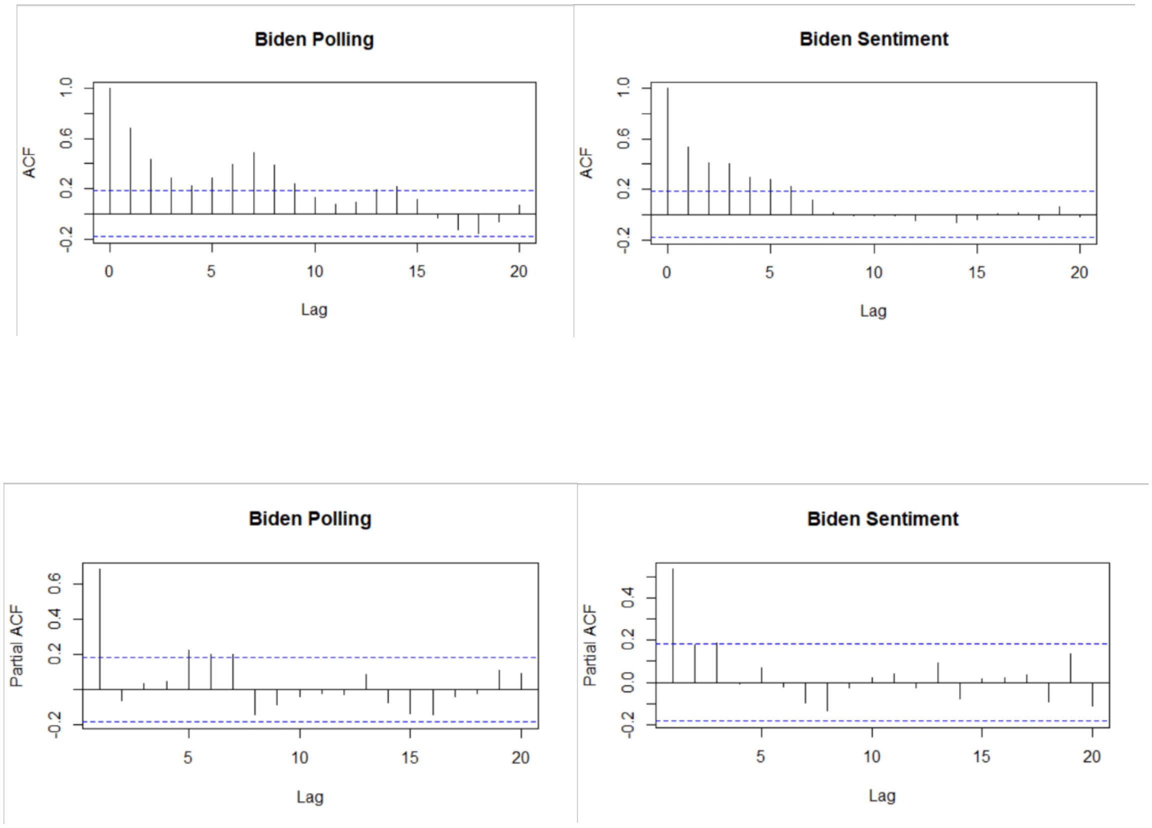


Figure 3: Biden ACF and PACF

The plots show a pattern seen in all four candidates: ACF correlation in polling across many lags, but only limited sentiment correlation in the first few lags. PACF correlation sporadic across the board, with few significant lags if any. Our polling series, in general, tend to be more auto-correlated than our sentiment series.

Temporal Smoothing and Differencing

Temporal smoothing is a critical issue for time series with high variance. We use practical knowledge to determine the amount of smoothing necessary for us to filter out white noise but preserve critical information. Observing the raw data, we see variance in short multi-day bursts, but general trends over week-long periods in both polling and sentiment series. Additionally, week-long increments are longer than our largest spans of missing data, so it can encapsulate or 'smooth over' this information. The following shows Biden's raw data and smoothed data with seven day rolling averages.

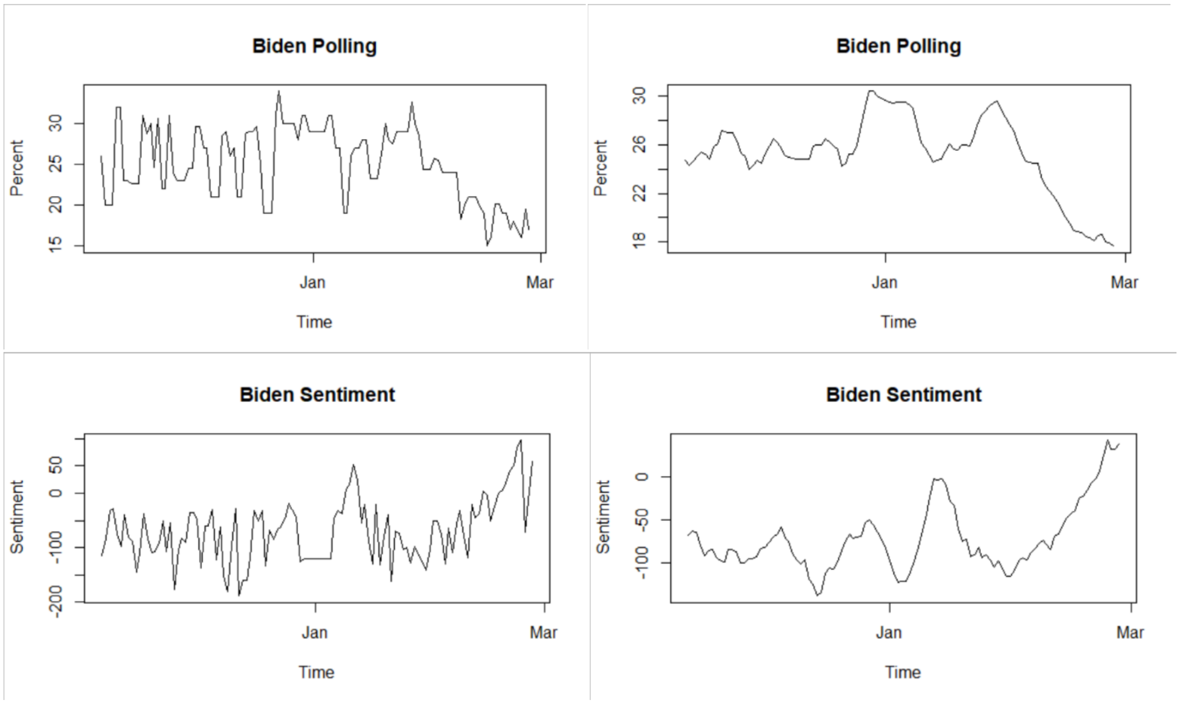


Figure 4: Biden Unsmoothed and 7-day Smoothed Data

The resulting plots demonstrate that a large amount of day to day variations can be filtered out while preserving general week to week trends. We observe in *Figure 4,* for instance, that Biden's polling and sentiment both oscillate until around February, where polling rapidly drops off and sentiment quickly increases.

Biden's plots both show potential for another anticipated problem: trends. Polling shows a negative trend over the time frame, while sentiment shows a positive time trend. Many of the other candidate time series share this issue. Removing these trends for stationary time series is necessary to fulfill assumptions for future modeling. Built-in tools in programming software can check for the differencing order each plot needs by calculating the unit root of the data. Hyndman et al.'s *forecast* package in R carries the *ndiffs* function which does just that. We find that each smoothed series relies on a first difference to achieve stationarity, except for Warren's sentiment and Buttigieg's polling. Because we maintain stationarity when we difference these two exceptions, we apply a first difference to them as well for the sake of consistency. Below is the transition from Biden's smoothed series to Biden's differenced series.

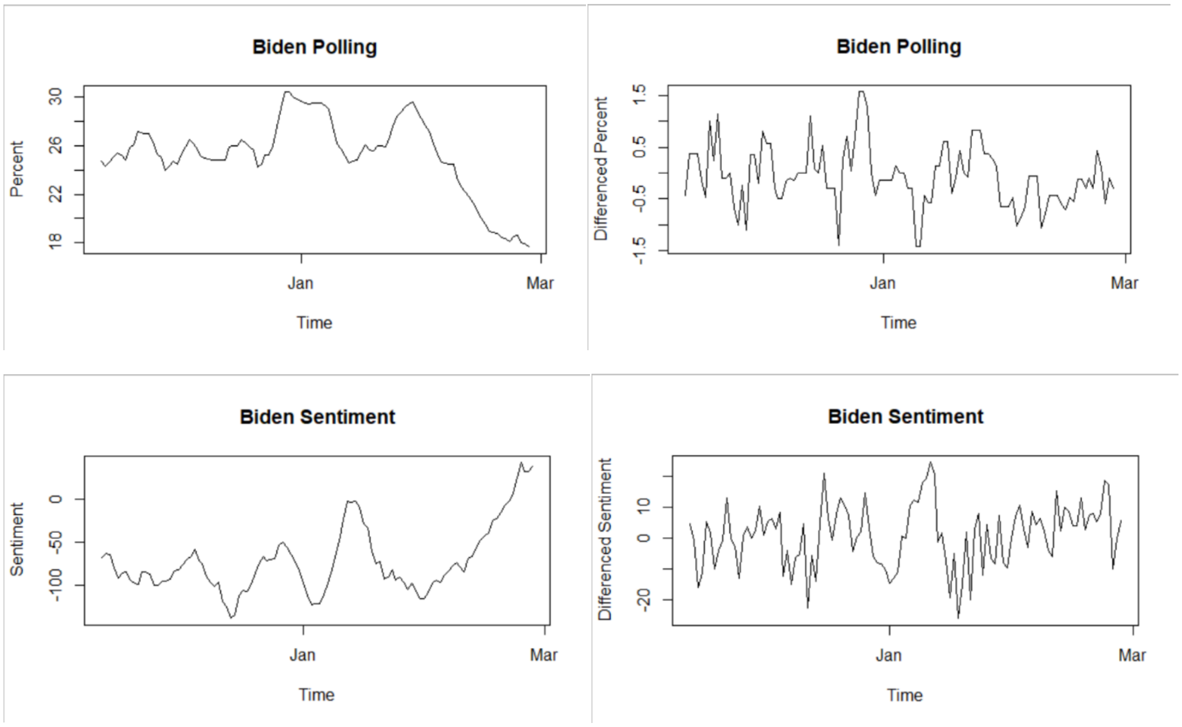
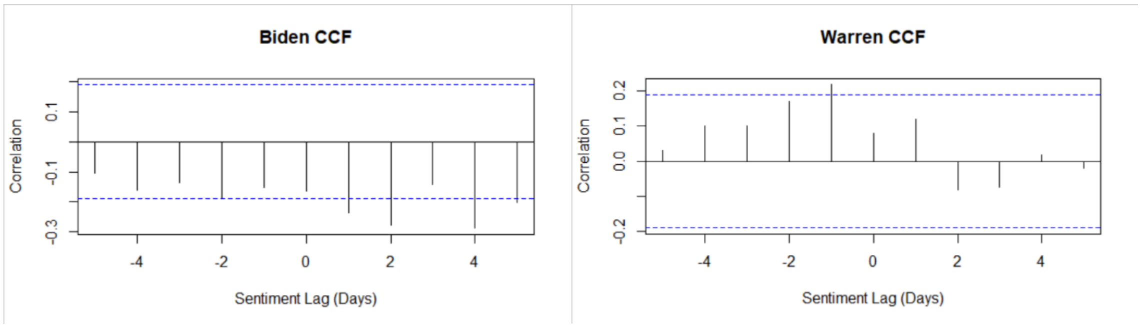


Figure 5: Biden Smoothed and Smoothed/Differenced Data

The results appear to show that we get back similar series to what we had before smoothing, only with the trend removed. This is not the case however, as differencing the unsmoothed series results in substantially greater variation and white noise than what we end up with. Now that our time series are pre-processed, we are ready to conduct comparative analyses.

Correlation Analysis

We now turn to investigating relationships between sentiment and polling time series. We examine the cross-correlations between each candidate pairing at different lags. Since we are examining whether sentiment is useful in forecasting polling, we hold each polling series constant and calculate the cross correlations at different lags of the sentiment series for each candidate.



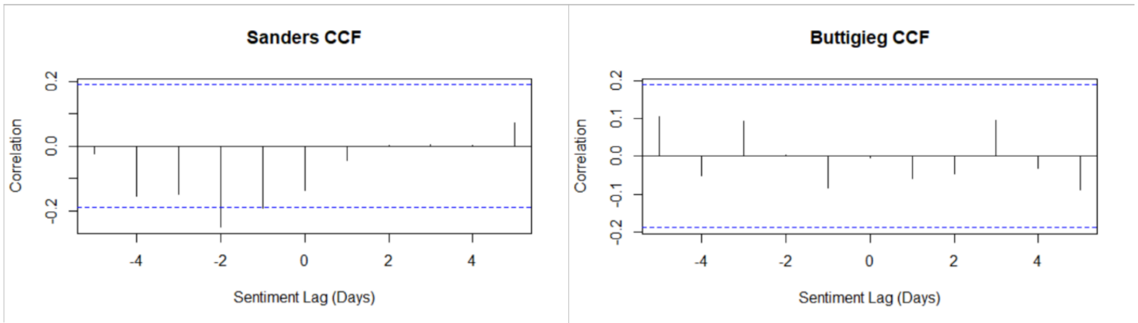


Figure 6: Candidate Cross-Correlations between Polling(fixed) and Sentiment(lagged)

Positive values of sentiment lag are correlations with polling at day *t* and sentiment at day *t+h* (where *h* is the sentiment lag value on the x-axis). Negative values of sentiment lags are correlations with polling at day *t* and sentiment at day *t-h*. Thus, strong correlations at negative lags in these plots indicate sentiment leads polling. We consider lags up to 5 days in keeping with our assumption that sentiment could plausibly only register a few days earlier than polling in public opinion. The blue line indicates statistical significance (a correlation of at least 20%), so lags that reach this line indicate possible useful forecasting. Biden, Warren, and Sanders have significant correlations in our desired time frame of 0 to -5 days, but Buttigieg does not. We note that the strongest correlations at negative sentiment lags appear between 0 and -2 days.

Granger Causality

Granger Causality is another test for whether sentiment is useful in forecasting polling. Granger Causality requires stationarity of its involved time series. Since we are now dealing with first-differenced series, we are now investigating whether changes in sentiment are useful in forecasting changes in polling. We conduct pair-wise tests for each candidate series. Once again, we limit lag order consideration to 5 since we assume it is not feasible sentiment could lead polling by more than 5 days.

P-Values	Sentiment Leads Polling	Polling Leads Sentiment
Biden	0.56	0.03
Warren	0.27	0.16
Sanders	0.19	0.98
Buttigieg	0.21	0.45

Table 1: Granger Causality Test P-Values for Differenced Candidate Series

The P-values above show we fail to reject the hypothesis that changes in sentiment adds useful information to forecasting changes polling. If we assume a traditional significance level of 0.05, only Biden's test for polling leads sentiment meets the cut-off. We note that overall Sanders and Buttigieg have more significant results for sentiment leading polling, while Biden and Warren have more significant results for polling leading sentiment. However, we do attain significance levels less than 0.3 for all candidates but Biden for models where sentiment leads polling. These significance levels give us reason to press on with short-run modeling techniques where sentiment predicts polling.

ARIMA and VAR Modeling

We are interested in showing whether sentiment is useful in predicting polling at short term lags of five days or less. Since we had to apply a first difference to our time series to achieve stationarity, we modified our question to whether changes in sentiment can predict changes in polling. However, since we have shown that each of our time series can be made stationary with a first difference, we can simply leave the time series undifferenced and include the trend directly in the model as a variable. Including the trend directly allows for us to interpret results as polling rather than changes in polling. ARIMA and VAR models are common modeling techniques for forecasting using time series. We propose four models to test our research question:

- Model 1: Separate Polling ARIMAs
- Model 2: Joint Polling VAR
- Model 3: Separate Polling/Sentiment VARs
- Model 4: Joint Polling/Sentiment VAR

Model Fitting

Model 1 fits ARIMA models to each candidate polling series *individually*. This model uses only information from each polling series alone, without accounting for any covariance between them in forecasting. Hyndman et al.'s *forecast* package in R provides the function *auto.arima* which automatically chooses AR and MA order for a time series based on its ACF and PACF and subsequently fits coefficients:

$$M1$$

$$PBi_t = 0.53PBi_{t-1}$$

$$PW_t = -1.29PW_{t-1} - 0.37PW_{t-2} + 1.79Pw_{t-1} + 0.92Pw_{t-2}$$

$$PS_t = -0.41PS_{t-1} + 0.47PS_{t-2} + 0.81Ps_{t-1} + 0.13\mu$$

$$PBu_t = 0.51PBu_{t-1}$$

where capital variables represent candidate polling at day *t*, lower-case variables represent white-noise errors at day *t*, and μ represents the mean of the time series.

Model 2 fits a single VAR with all four candidate polling time series. This model uses information from past lags of each polling time series as well as past lags from other candidate polling series for forecasting. Thus, it does account for covariance between each candidate polling series. Pfaff's *vars* package in R gives the *VARselect* function which yields an AR order of 1 for this data. This model includes four equations, each considering two lags from itself, two lags from all other polling series, and a trend variable. We only show Biden's equation for simplicity:

$$M2$$

$$PBi_t = 1.38PBi_{t-1} - 0.44PBi_{t-2} + 0.08PW_{t-1} - 0.1PW_{t-2} + 0.15PS_{t-1} - 0.17PS_{t-2} - 0.17PBu_{t-1} + 0.05PBu_{t-2} - 0.004t + 3.59$$

where *t* is the overall trend of each time series and the other variables are as above.

Model 3 consists of separate pair-wise VARs for polling and sentiment series of each candidate. This model isolates the effects of each candidate's sentiment on its polling, not considering covariation with other candidates. We again select model order and fit the trend directly into each model. Biden's pairwise VAR is as follows:

$$M3$$

$$PBi_t = -0.004SBi_{t-1} + 1.5PBi_{t-1} + 0.004SBi_{t-2} - 0.52PBi_{t-2} - 0.004t + 0.9$$

were *SBi* are sentiment lags and all other variables are as above.

Model 4 combines all eight polling and sentiment series into a single VAR. This accounts for the effects of sentiment on polling, and all covariance between the time series. We display Biden's total VAR below:

$$M4$$

$$PBi_t = 0.001SBi_{t-1} - 0.003SW_{t-1} - 0.001SS_{t-1} - 0.004SBu_{t-1} + 0.97PBi_{t-1} - 0.04PW_{t-1} + 0.01PS_{t-1} - 0.13PBu_{t-1} - 0.006t + 2.53$$

Model Evaluation

We make predictions using ARIMA and VAR models by forecasting over a specified period. Since we project that sentiment estimates political opinion only a few days ahead of polling, we assume that our models will be used to forecast 3 days ahead for now. We will return to this issue in a sensitivity analysis. The simplest way to evaluate the models would be to divide all our data into a training set and test set, fit the models again over the training set, and then forecast using the test set. We would then assess model performance with a performance metric like MAE. However, this leaves us vulnerable to a biased test set where we could 'get lucky' with favorable data, or vice versa. To mitigate this issue, we cross validate each model. The accepted method of cross validation with time series analysis is to 'cascade' across the data, gradually increasing the training set in size (Hyndman et al., 2018). This cross validation technique preserves the temporal order of the time series data while varying the test set. A visual representation of this method of cross validation is as follows:

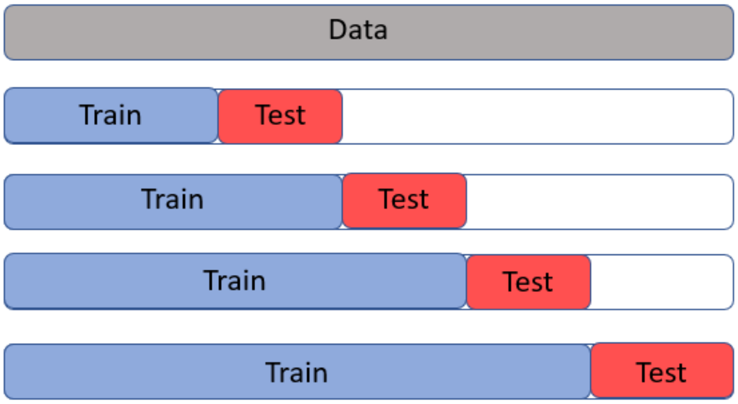


Figure 7: Time Series Cross Validation

We select a period of a month from the beginning of our smoothed data on November 12, 2019 to December 12, 2019 as the first training set. The first test set is the consecutive three-day period December 13-15. We then evaluate the MAE of each model using these training and test sets, before moving the windows forward three days and evaluating again. We continue this process until the test window reaches the final three days of the data. Note that we use smoothed data for our training windows but unsmoothed for test windows to simulate real-world usage of the models. We then average the test set MAEs of each fold in each model to attain the total model MAE. This yields the following results.

Model	**MAE**
Model 1: Separate Poll ARIMAs	2.300078
Model 2: Total Poll VAR	2.364915
Model 3: Separate Sentiment-Poll VARs	2.319908
Model 4: Total Sentiment-Poll VAR	2.434311

Table 2: 3-Day Forecasts with Bing Sentiment and Unweighted Tweets

The results show that the less information there is from sources other than the polling series itself, the better the three-day forecasting performance. We do see a strong performance by the pair-wise Sentiment-Polling VAR model, indicating some promise from the inclusion of sentiment in forecasting polling. However, this is overshadowed by the stronger performance of individually fitted Separate Poll ARIMAs model. Our initial takeaway is that sentiment may yield a small benefit in forecasting polling, but our sentiment analyses measures likely could use improvement to better estimate public opinion. We attempt to confirm this finding by using weighted sentiment measures and substituting the AFINN lexicon in our sensitivity analysis. Regardless, there seems to be a more significant impact from reducing the variable interactions in the models than including sentiment time series.

Sensitivity Analysis

We made assumptions to select numerous parameters in this study. We selected the *bing* lexicon, chose not to account for retweets/favorites in our sentiment analysis, and used 3-day prediction intervals in cross validation. We modify these parameters to determine effects on results in the following comparisons:

- unweighted versus weighted retweets/favorites
- *bing* versus *AFINN* sentiment lexicons
- 2-day, 3-day, 7-day prediction intervals

Tweet Weighting

Tweets can be retweeted by other users on Twitter, usually as a means of expressing approval. We incorporate retweets into our sentiment analysis by weighting tweets based on their retweet count (simply multiplying the raw sentiment score of each tweet by its retweet count suffices). Using identical parameters as before besides this change yields the following results.

Model	**MAE**
Model 1: Separate Poll ARIMAs	2.300078
Model 2: Total Poll VAR	2.364915
Model 3: Separate Sentiment-Poll VARs	2.299402
Model 4: Total Sentiment-Poll VAR	2.381964

Table 3: 3-day Forecasts with Bing Sentiment and Weighted (Retweets) Tweets

We see a small but marked improvement in both our sentiment-inclusive models. The Separate Sentiment-Polling VARs model surpasses the Separate Polling ARIMAs as the best performing model, and the Total Sentiment-Polling VAR closes in on the Total Polling VAR. Including favorite count in our weighting (raw sentiment multiplied by the sum of retweet and favorite count) generates the following results.

Model	**MAE**
Model 1: Separate Poll ARIMAs	2.300078
Model 2: Total Poll VAR	2.364915
Model 3: Separate Sentiment-Poll VARs	2.299106
Model 4: Total Sentiment-Poll VAR	2.381562

Table 4: 3-day Forecasts with Bing Sentiment and Weighted (Retweets and Favorites) Tweets

These results show an incremental change (in the thousandths of a percent), but nonetheless improve our sentiment-based models (Models 3 and 4) further. We conclude that refined sentiment weighting does improve model performance and strengthens the case that sentiment can assist in forecasting polling.

Lexicon Choice

We modify the sentiment lexicon used in our analysis from *bing* to *AFINN* to observe sensitivity to word choice. Using initial parameters and not adjusting for retweet/favorite count, we find the following results.

Model	MAE
Model 1: Separate Poll ARIMAs	2.300078
Model 2: Total Poll VAR	2.364915
Model 3: Separate Sentiment-Poll VARs	2.328047
Model 4: Total Sentiment-Poll VAR	2.386943

Table 5: 3-day Forecasts with AFINN Sentiment and Unweighted Tweets

Compared to the *bing* lexicon, this improves both sentiment-based models (Models 3 and 4). Testing with the *AFINN* lexicon and weighting by retweet/favorite count yields the following table.

Model	MAE
Model 1: Separate Poll ARIMAs	2.300078
Model 2: Total Poll VAR	2.364915
Model 3: Separate Sentiment-Poll VARs	2.308731
Model 4: Total Sentiment-Poll VAR	2.327982

Table 6: 3-day Forecasts with AFINN Sentiment and Weighted (Retweets and Favorites) Tweets

We observe a substantial improvement from the Total Sentiment-Poll VAR model, surpassing the Total Poll VAR model easily. The Separate Sentiment-Polling VARs model also improves, but not by quite enough to surpass the same model's performance in *Table 4*.

Forecasting Interval

Our original models forecasted three days ahead of time. The 3-day forecasting window was based on our assumption that sentiment effects should be felt a few days ahead of polling. We modify this interval to test 2-day and 7-day forecasting periods. We maintain the 'best' parameters from our previous analysis: weighted tweets and the AFINN lexicon. 2-day forecasting yields the following results.

Model	MAE
Model 1: Separate Poll ARIMAs	2.158614
Model 2: Total Poll VAR	2.219302
Model 3: Separate Sentiment-Poll VARs	2.190178
Model 4: Total Sentiment-Poll VAR	2.216096

Table 7: 2-day Forecasts with AFINN Sentiment and Weighted (Retweets and Favorites) Tweets

As expected with shorter prediction intervals, we see lower errors across the board for our models. The same order of performance remains from *Table 6*. Next we check 7-day prediction intervals.

Model	MAE
Model 1: Separate Poll ARIMAs	2.351701
Model 2: Total Poll VAR	2.387332
Model 3: Separate Sentiment-Poll VARs	2.390327
Model 4: Total Sentiment-Poll VAR	2. 389508

Table 8: 7-day Forecasts with AFINN Sentiment and Weighted (Retweets and Favorites) Tweets

The longer prediction intervals yield the highest errors yet. Polling-based models (Models 1 and 2) perform the best in longer term predictions, in keeping with our assumption that sentiment is a relatively instantaneous measure.

Case Studies: The Iowa Caucuses and South Carolina Primary

Though we were not definitively able to show that sentiment can be useful in forecasting polling in primary elections, we remain interested in how our models would perform in the real world. We select our 'best' model parameters and conduct forecasts on state level data from Iowa and South Carolina. Our model parameters are weighted tweets, AFINN sentiment lexicon, and a 2-day forecasting window. We calculate the Sum of Squared Errors (SSE) between the 2-day forecasts for each candidate, and the actual percentage vote for each candidate in the Iowa Caucuses on February 2nd.

Model	SSE
Model 1: Separate Poll ARIMAs	166.8070
Model 2: Total Poll VAR	137.7671
Model 3: Separate Sentiment-Poll VARs	190.2634
Model 4: Total Sentiment-Poll VAR	134.2666

Table 9: Iowa 2-day Forecasts with AFINN Sentiment and Weighted (Retweets and Favorites) Tweets

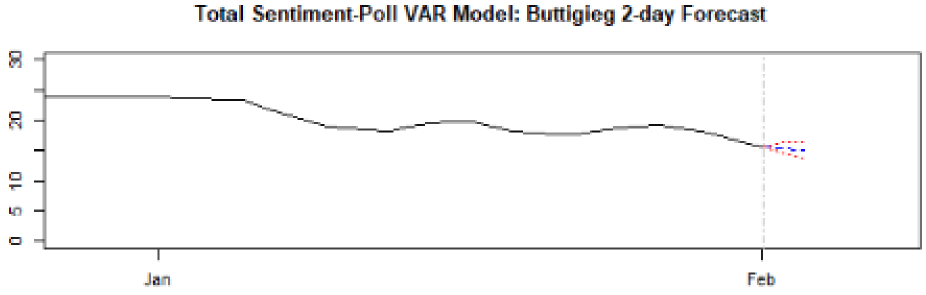

Total Sentiment-Poll VAR Model: Buttigieg 2-day Forecast

Figure 8

We observe that the models with more information and covariance performed better in Iowa. The Total Sentiment-Poll VAR model performs the best, implying some benefit from including sentiment in our forecasting. The Iowa Caucuses had the somewhat surprising result of Mayor Buttigieg and Senator Sanders virtually tying for the victory, while Vice President Biden and

Senator Warren had unexpectedly weak performances. Sentiment may have been useful in picking up underlying trends that polls in Iowa failed to capture.

We perform an identical analysis with the South Carolina Primary, which took place on February 29th. Again, we use 2-day forecasting from the end of our dataset (February 27th), weighted tweets, and the AFINN sentiment lexicon.

Model	**SSE**
Model 1: Separate Poll ARIMAs	187.7195
Model 2: Total Poll VAR	244.5692
Model 3: Separate Sentiment-Poll VARs	263.1933
Model 4: Total Sentiment-Poll VAR	319.4662

Table 10: S.C. 2-day Forecasts with AFINN Sentiment and Weighted (Retweets and Favorites) Tweets

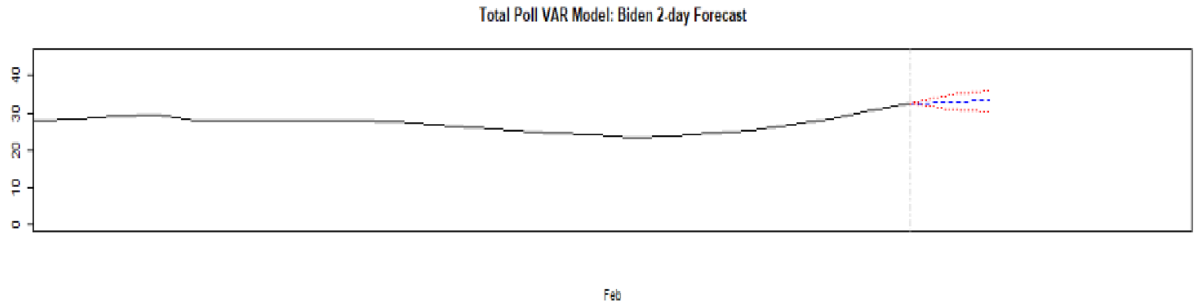


Figure 9

We obtain very different results from Iowa. In this case study, both the polling models (Models 1 and 2) perform substantially better than the models including sentiment (Models 3 and 4). Less information and covariance between candidates lead to better results. The South Carolina Primary was a blowout victory for Vice President Biden, who won every single county. Polls consistently showed Biden with a large cushion on his rivals, and the results bore this out. In South Carolina, there seemed to be little room for sentiment to explain an underlying trend because there was no such trend. Polls showed Biden with broad support, and the result coincided.

Conclusion

We set out in this study to determine whether Twitter-based sentiment could be useful in forecasting polling in primary-style elections. The benefit of an affirmative answer would be supplementing or replacing expensive polling methods with cheaper sentiment analysis. Similar studies on national elections both in the U.S and abroad were largely successful in establishing this relationship. We anticipated greater challenges with primaries due to their volatile candidate field and staggered elections. The results bore these challenges out. Weak statistical tests between candidate sentiment and polling corresponded with sentiment-based models failing to outperform polling-based counterparts consistently. Models predicting future polling typically did just as well when based solely on past polling than when based on past polling and sentiment. Despite failing to answer our study question positively, sensitivity analysis and case studies point to promising areas of future study in primary election forecasting. Logical improvements to sentiment analysis methods through weighting sentiment by retweets and favorites produced substantive improvements in models. Differences in predictability between the Iowa Caucuses and South Carolina Primary paralleled differences in model performances at predicting their results. Sentiment-based models outperformed polling-based models in the more surprising Iowa Caucuses, where Senator Sanders virtually tied upstart Mayor Buttigieg. Conversely, polling-based models performed strongly on the widely expected result in the South Carolina Primary, where Vice President Biden cruised to victory. These findings point us to two areas of future interest: improved sentiment analysis algorithms and sentiment as a predictor of primary election polling shocks. We anticipate the former methods like basing sentiment on entire tweets rather than bi-grams within each tweet could greatly reduce classification error and lead to better performance of sentiment-based models. The latter could leverage the ability of sentiment to pick up on underlying or surprising trends to improve primary election forecasting.

Acknowledgements

This study was supported by the Department of Mathematical Sciences at the United States Military Academy. We would like to give a special thanks to the following individuals: Lieutenant Colonel Kevin Cummiskey and Lieutenant Colonel Christopher Weld for advising, Major Bryan Jonas for enabling tweet file scraping automation, and Captain Mario Andriulli for review.

References

Abramowitz, Alan I. "Viability, Electability, and Candidate Choice in a Presidential Primary Election: A Test of Competing Models." *The Journal of Politics 51, no. 4* (1989): 977-992.

Archdeacon, Thomas J. *Correlation and Regression Analysis: a Historians Guide*. University of Wisconsin Press, 1994.

Baseballot. "How The Giant Democratic Primary Field Messes With Polls." *FiveThirtyEight*, FiveThirtyEight, 4 Apr. 2019, fivethirtyeight.com/features/how-the-giant-democratic-primary-field-messes-with-polls/.

Bermingham, Adam, and Alan Smeaton. "On using Twitter to monitor political sentiment and predict election results." *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*. 2011.

DataDhrumil. "National President: Democratic Primary Polls." *FiveThirtyEight*, 6 May 2020, projects.fivethirtyeight.com/polls/president-primary-d/national/.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, vol. 37, no. 3, 1969, p. 424., doi:10.2307/1912791.

Hanck, Christopher, et al. *Introduction to Econometrics with R*. University of Duisburg-Essen, 2019.

Hyndman, Rob J., and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.

Jerôme, Bruno, and Véronique Jerôme-Speziari. "Forecasting the 2012 US presidential election: Lessons from a state-by-state political economy model." *PS: Political Science & Politics* 45.4 (2012): 663-668.

Lewis-Beck, Michael S., and Tom W. Rice. "Forecasting presidential elections: A comparison of naive models." *Political Behavior* 6.1 (1984): 9-21.

Linzer, Drew A. "Dynamic Bayesian forecasting of presidential elections in the states." *Journal of the American Statistical Association* 108.501 (2013): 124-134.

Liu, Bing. *Opinion Mining, Sentiment Analysis, Opinion Extraction*, www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.

Nielsen, Finn Arup. *AFINN*, www2.imm.dtu.dk/pubdb/pubs/6010-full.html.

O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth International AAAI Conference on Weblogs and Social Media*. 2010.

Silge, Julia, and David Robinson. *Text Mining with R: a Tidy Approach*. OReilly Media, 2017.

STAT 510. *PennState: Statistics Online Courses*, Penn State Eberly College of Science, online.stat.psu.edu/stat510/.

Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Fourth international AAAI conference on weblogs and social media*. 2010.