2020 Democratic Primary Twitter Sentiment Analysis

Kevin McGuire

May 4, 2020

Abstract

An exploration of the potential for using sentiment analysis on social media content to predict election results.

1 Introduction

Predicting election results is one of the most popular yet challenging questions of our time. The primary tool that has historically been used to estimate results, polling of individuals, faces a number of limitations that even the best technologic innovations cannot address.[5][4] Some of these limitations are inherent to the nature of traditional polling, while others have emerged from this particular social moment. No matter the method used to administer a poll, some level of lag will exist between it being launched in the field and the point at which results have been cleaned and prepared for analysis. Emerging forms of technology have opened doors for new methods, such as online polling, that may shorten the lag, but introduce new unknowns about response rates and representative samples. In any method of polling, response rates will never reach an ideal point.

Perhaps a better solution, or at least one that can be tested alongside more traditional methods, resides in the mining of user generated data. In a world where people willingly post endless political content on social media, can analyzing this vast, raw source of data provide new or better insights? My project was to perform a basic exploration of the potential of

using sentiment analysis on political Tweets as a predictor of election performance. The Democratic presidential primary would be my test case.

2 Literature Review

A number of studies using sentiment analysis on Twitter data have been done. For the purposes of my project, I identified three key studies that were done about a decade ago that form the basis for understanding the application of this methodology to elections.

A study conducted by the University of Technology in Munich, Germany explored many of the foundations for using Twitter as predictive of both election results and as a medium for political communication. While the study focused largely on concepts less relevant to American politics, such as ties between parties, it also confirmed the validity of many factors that were valuable for me know in advance. [1]

A study conducted by Dublin City University focused on the application of Twitter sentiment analysis to Irish elections, building off of existing predictive models using social media. This study was important as I deter-

mined what data to collect, specifically on the importance of including raw volume statistics in addition to just sentiment. [2]

The final study I used in detail for preparing this project, was conducted by the University of Southern California on the 2012 presidential election. While this project was the most similar in concept to my own, it was also primarily a demonstration of how to collect this sort of data, rather than making any broad predictive claims. While the methods for collecting data in it are now largely outdated as Twitter has improved their API and new packages have been developed to better use these tools, the USC project examined the sorts of events that cause spikes in political Twitter volume, which was crucial for understanding some of my results. [3]

3 Data

Due to the fact I was limited by the constraints of the free Twitter API, data collection required decisions to be made well in advance of when I'd be able to start even preliminary analysis. In order to collect as large a data set as possible, I made the decision to download all Tweets

that tagged a major candidate[1]. In cases where a candidate had more than one Twitter handle (e.g. @BernieSanders and @SenSanders), whichever was primarily associated with their campaign activities was used. Retweets of a candidate without additional comment were not included in the data. I ran my collection script daily from February 4th to March 10th. This captured the majority of the competitive primary season. Ultimately, I trimmed the analysis range to end on March 4th due to the race winding down, and not wanting to include coronavirus Tweets. The final data set resulted in over 1.1 million Tweets for analysis.

These Tweets were cleaned, run through a sentiment analysis package in R, and then summarized into new variables. Based on previous studies and preliminary exploration of my data, I ended up with seven variables that could be used to summarize the sentiment data at the candidate and date level. In many cases, I converted the data into multiple similar variables, with slight differences to test the performance of each version. In some cases one was clearly superior, while in others each version pointed to unique underlying information.

[1]Major candidates were determined by polling average and debate invitations, including: Biden, Bloomberg, Buttigieg, Klobuchar, Sanders, Steyer, Warren, and Yang

3.1 Count Tweets

$$\sum(candidate\,tweets)$$

This variable served as the raw measurement of how much engagement a candidate was receiving on a daily basis. This was included based off of the findings of the Dublin and Munich studies on the importance of including Twitter volume with sentiment analytics.

3.2 Percent Tweets

$$\frac{\sum(candidate\,tweets)}{\sum(total\,tweets)}$$

This variable serves a similar function to Count Tweets, but normalizes the volume by considering how much of the total daily candidate Tweets is represented by a given candidate. This protects from days with debates or similar events that boost the volume of all candidates having an outsize impact.

3.3 Total Sentiment

$$\sum(candidate\ sentiment)$$

This variable shows the sum of all the sentiment scores for every Tweet tagging a given candidate. This measurement is designed to show which candidates are well-liked at the given time by offsetting positive scores with the negative ones.

3.4 Total Absolute Sentiment

$$\sum(\ |candidate\ sentiment|\ )$$

This variable shows the sum of the absolute values for every Tweet tagging a given candidate. This measure is designed to capture which candidates are most polarizing by considering positive and negative scores equally.

3.5 Percent Sentiment

$$\frac{\sum(\ |candidate\ sentiment|\ )}{\sum(\ |total\ sentiment|\ )}$$

This variable is similar to Total Absolute Sentiment, but normalizes the values over time by showing a candidate's sentiment relative to the total polarization on a given day. This works to balance against the effects of major events that boost volume such as debates.

3.6 Average Sentiment

$$\frac{\sum(\,candidate\ sentiment\,)}{\sum(\,candidate\ tweets\,)}$$

This variable measures the average sentiment in each Tweet directed at a candidate. It serves primarily as an indicator for low ranking candidates whether they have room for growth as they gain exposure. Possibly limited in utility by constraining the study to major candidates relatively late in the process.

3.7 Average Absolute Sentiment

$$\frac{\sum(\,|candidate\ sentiment|\,)}{\sum(\,candidate\ tweets\,)}$$

This variable works almost identically to Average Sentiment, but

measures total polarization. Its purpose is essentially the same, just measured slightly differently.

4 Methods

I had a multistage process for collecting the Tweets and transforming them into useful data. The Tweets would be downloaded using Twitter's API and the rtweet package. Each downloaded Tweet contained 90 columns. All but the date and text would be removed, and the content run through the sentimentr package to determine a sentiment score for each Tweet. Files would be combined and aggregated by candidate and day.

The resulting sentiment data would be compared against both traditional polling data (national polling average as calculated by FiveThirtyEight) and results from the primaries and caucuses that occurred during the measurement window. As a result of the limited number of data points provided where all could be measured against each other, as well as the exploratory nature of this project, evaluations were limited to correlations.

5 Findings

Many of my variables performed with similar correlations to the election results as did the traditional polling average. One of the variables, Percent Tweets, even outperformed the polling average by a small margin. This being a measurement of Twitter volume, is consistent with prior studies. The fact that Percent Tweets is a more accurate predictor than Count Tweets in this experiment is also consistent with my expectation that adjusting for total volume would be effective. The full correlation results can be seen on Table 1 in the appendix.

The overall tracking of Percent Tweets followed the results relatively closely, both in ranking the candidates and in their relative movements. The charts of actual results and the Percent Tweets variable can be seen in Chart 1 and Chart 2. Polling average can be seen in Chart 3.

While not one of the best predictors, Average Sentiment appears to have some value in other areas, as it tracks closely with events that are heavily reported in the political media as they happen. A prime example of this is Bloomberg reaching the lowest Average Sentiment Score of the entire

study during his disastrous debate debut. This can be seen in Chart 4. There may be value in a study of Twitter sentiment verses political media coverage to determine if certain types of events are over or under covered relative to immediate reactions as a future branch of research.

6 Conclusion

Overall, promising results were observed through the use of Twitter sentiment as a predictor of election performance. Further study is certainly needed to determine whether it is a viable alternative to traditional polling in a larger context, or if it will remain only a tool for finding interesting points for analysis. These additional studies will need to explore a larger sample of elections across a variety of time frames and contexts, preferably not while public discourse is occupied by a global pandemic. Ideally, premium Twitter location data would also be used to better match Tweets occurring in the locations voting at specific times, instead of using the complete national data set.

References

[1] P. G. S. I. M. W. Andranik Tumasjan, Timm O. Sprenger, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *International AAAI Conference on Web and Social Media*, 2010.

[2] A. Bermingham and A. F. Smeaton, "On using twitter to monitor political sentiment and predict election results," *International Joint Conference on Natural Language Processing*, pp. 2–10, 2011.

[3] A. K. F. B. S. N. Hao Wang, Dogan Can, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," *Annual Meeting for the Association for Computational Linguistics*, 2012.

[4] N. Silver. Is the polling industry in stasis or in crisis? [Online]. Available: https://fivethirtyeight.com/features/is-the-polling-industry-in-stasis-or-in-crisis/

[5] ——. The state of the polls 2019. [Online]. Available: https://fivethirtyeight.com/features/the-state-of-the-polls-2019/

	Vote	Polling	Total_Absolute_Sentiment	Percent_Sentiment	Total_Sentiment	Count_Tweets	Percent_Tweets	Average_Sentiment	Average_Absolute_Sentiment
Vote	1.0000000	0.58365432	0.4904613	0.5673366	0.3711342	0.5210140	0.5935236	0.28693591	0.3097615
Polling	0.5836543	1.00000000	0.5873825	0.7769470	0.3226651	0.6296876	0.8128638	0.02348151	0.4591911
Total_Absolute_Sentiment	0.4904613	0.58738254	1.0000000	0.8450909	0.8983913	0.9966833	0.8362123	0.31501476	0.3492089
Percent_Sentiment	0.5673366	0.77694699	0.8450909	1.0000000	0.6586708	0.8557691	0.9960999	0.25110323	0.4708854
Total_Sentiment	0.3711342	0.32266511	0.8983913	0.6586708	1.0000000	0.8795963	0.6395810	0.57225391	0.2471578
Count_Tweets	0.5210140	0.62968757	0.9966833	0.8557691	0.8795963	1.0000000	0.8532276	0.30177992	0.3437784
Percent_Tweets	0.5935236	0.81286385	0.8362123	0.9960999	0.6395810	0.8532276	1.0000000	0.23415641	0.4549303
Average_Sentiment	0.2869359	0.02348151	0.3150148	0.2511032	0.5722539	0.3017799	0.2341564	1.00000000	0.4633467
Average_Absolute_Sentiment	0.3097615	0.45919106	0.3492089	0.4708854	0.2471578	0.3437784	0.4549303	0.46334667	1.0000000

Figure 1: Table 1

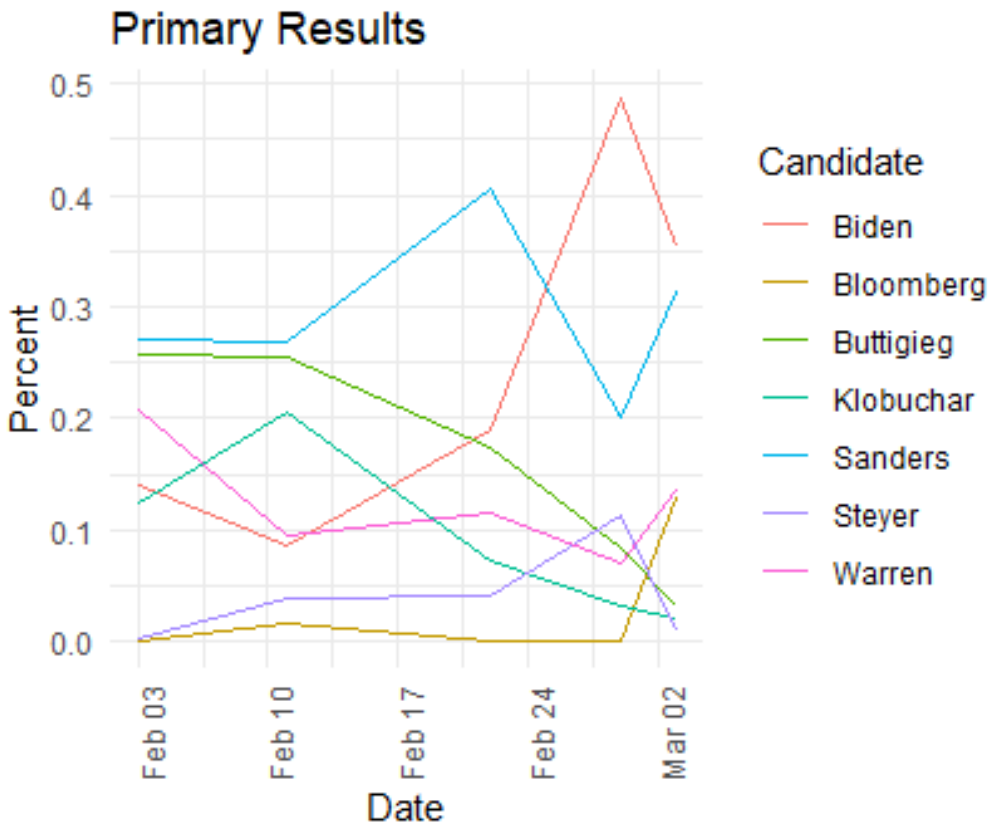

Figure 2: Chart 1

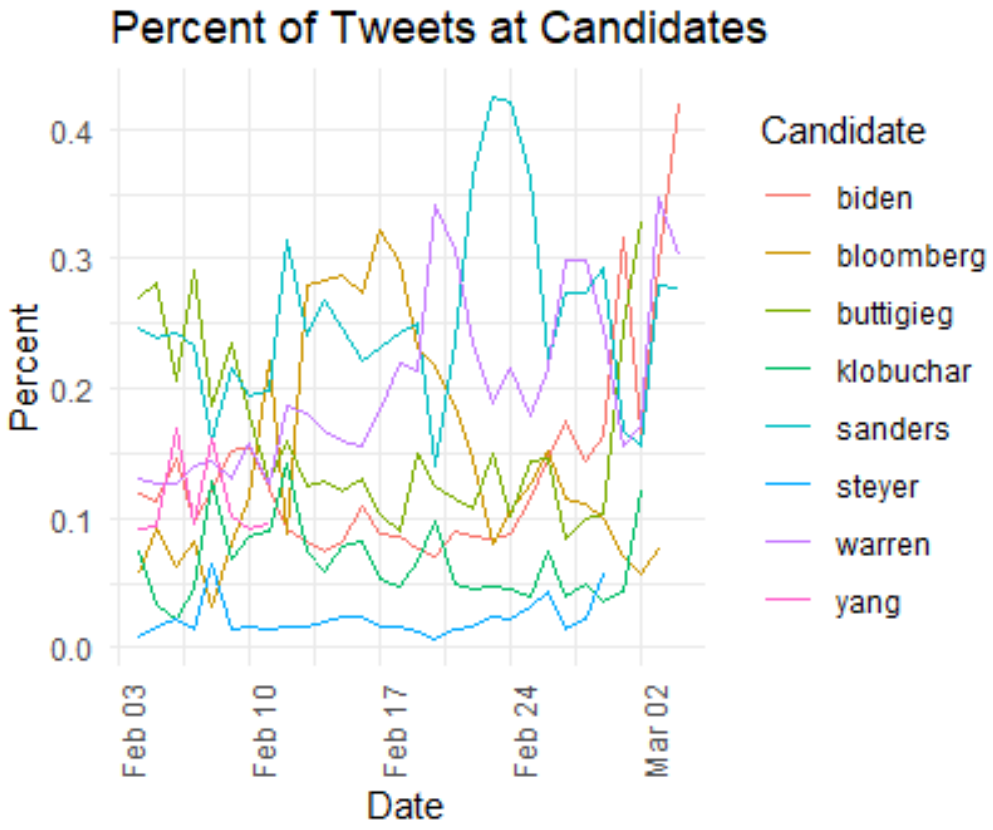

Figure 3: Chart 2

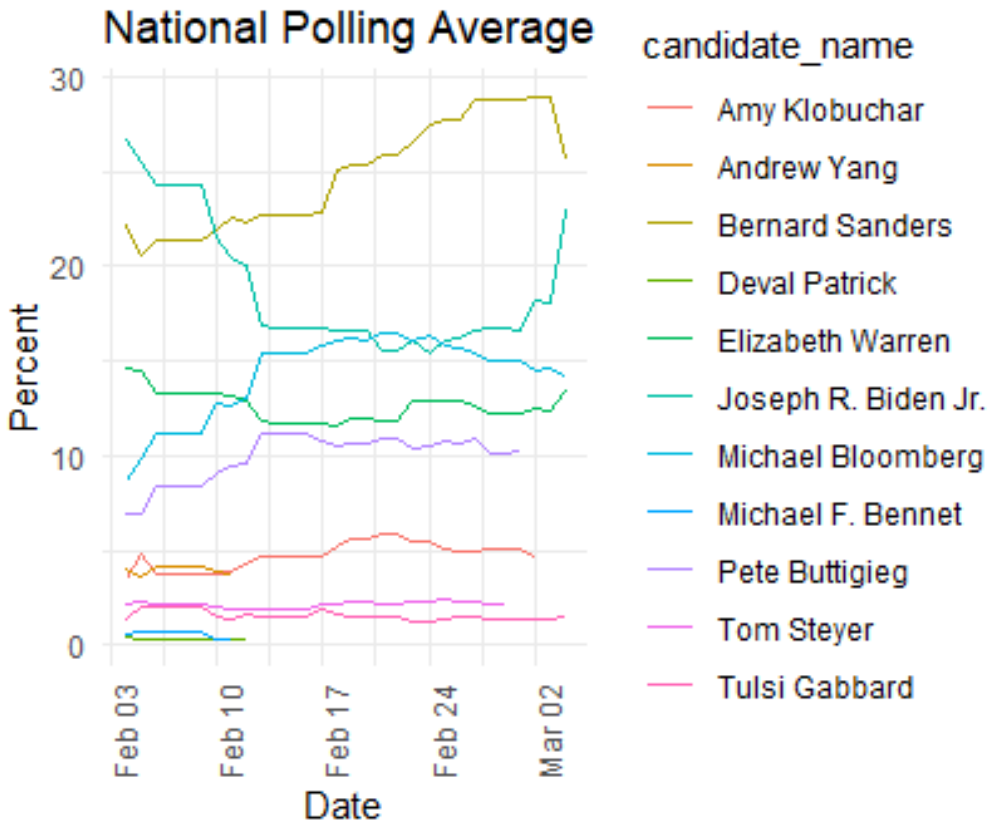

Figure 4: Chart 3

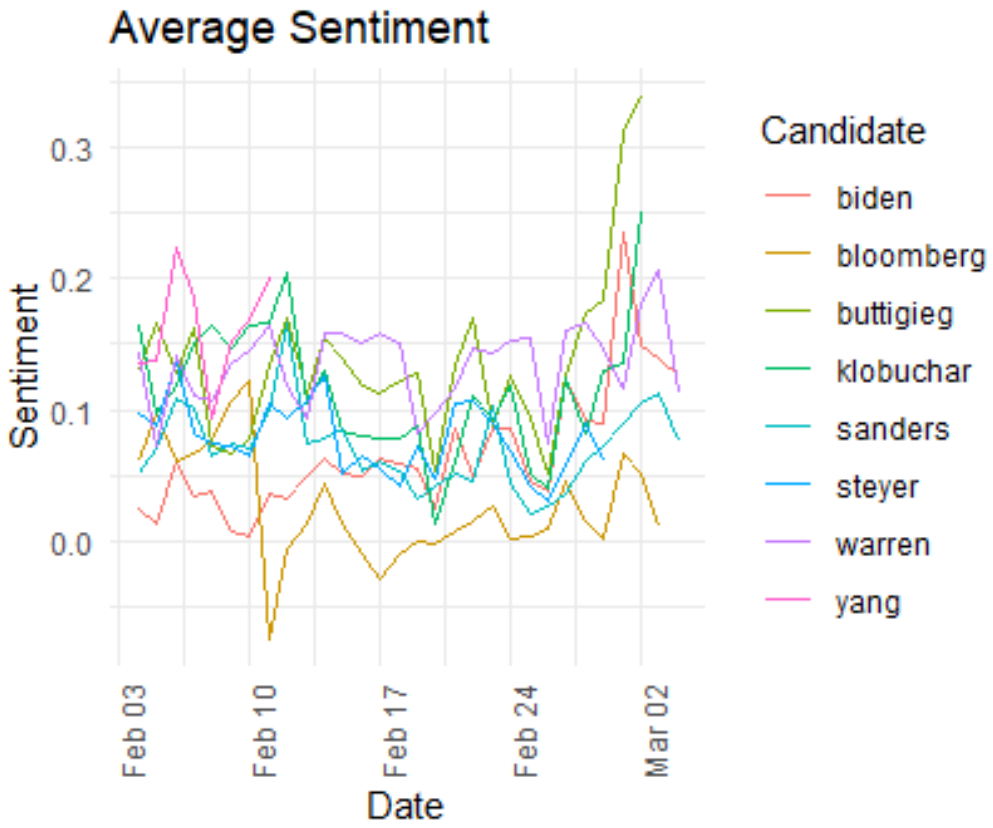

Figure 5: Chart 4