NEWS SENTIMENT AND NETWORK EFFECTS

ALMA GRACIC

ABSTRACT. Timely understanding of public perception is of vital importance for policymakers, as this measure influences how they respond to public issues. I use the wealth of data available from Twitter to measure public opinion that is traditionally captured with time consuming and expensive surveys. In particular, this study focuses on improving forecasting performance of social and economic activities, such as daily presidential approval ratings, through machine learning techniques. It measures media sentiment shocks and evaluates the network effects via both active (retweets) and passive (favorites) network activities. Additionally, I build a neural network to account for nonlinearity in emotion-based predictions. I train the sentiment index measured from news and its network response on measures available from polls, and find that using social media data can establish a presidential approval index in real-time and with more precision. I also find that shocks in media sentiments have a more immediate effect, while network predictors have a longer lasting impact. Granger analysis shows that the past values of the media sentiment and its network effects are important beyond past values of the approval index alone. Results indicate that the model can successfully classify the public's opinions and emotions. In the second part of the paper I evaluate the effects of "self-promotion". Results suggest that signals found in the sentiment and the network response to the President's own tweets are predecessors to the same effects in media, and consequently to the President's approval index.

Keywords: Social Media, Social Networking, Beliefs, Network Effect, Forecasting, User Behavior, Big Data, Data Mining, Text Sentiment Analysis, Neural Networks.

JEL Classification: C45, C55, C81, D83, E17.

UNIVERSITY OF MIAMI, Department of Economics, Coral Gables, FL; e-mail: agracic@bus.miami.edu
This is a draft of the first chapter of my thesis. The paper was presented at the Institute of New Economic Thinking YSI North America Convening, USC Dornsife, Los Angeles, CA on February 23, 2019, the Missouri Valley Economics Association Conference in Kansas City, MO on October 10, 2019, and at the Workshop in Applied and Theoretical Economics in Gainesville on October 25, 2019; and will be presented at the Southern Economic Association 89th Annual Meeting in Fort Lauderdale, FL in November 2019.
The author declares no funding source and no conflicts of interest for this research. Any errors that remain are mine alone.

1. Introduction

Public opinion polls play an important role as they represent a link between citizens and policymakers. They serve as a tool for citizens to express their views on certain events and policies. Polls also help politicians focus their attention on issues that are most valued by the public and pass legislations a majority of citizens agree with. Elections are events in which opinion polls have the greatest measured effect. During election cycles, the most visible impact of polling is on media, as candidates who poll well generally receive significantly more media coverage. As media puts such high emphasis on polling, it is not surprising that opinion polls earn a bad reputation easily in cases of high-profile inaccuracies in forecasting major events. In the past few years, some of the flaws of using survey methods for measuring public opinion have intensified. This is mostly due to declines in response rates, as traditional survey approaches still rely on phone-calling methods. Nevertheless, election polls and polls tracking movement in public opinion remain quite accurate, and as long as we are aware of their limitations and what the polls are telling us, they can be trusted.

This paper therefore considers the importance of evaluating opinion changes following certain events, and the role of social networks as an alternative way of measuring public opinion. The importance of identifying and understanding public sentiment on major policy debates is well understood, yet many empirical questions remain as to how the public receives the information, how they perceive such information, and most importantly, how they react to it. With increasing tensions over evaluating the impacts of mainstream media bias, and constantly increasing social media usage, it is important to not just look at public opinion outcomes but also to quantify the effects of news diffusion through social networks and their role in belief formation and public opinion changes.

In this paper, I use social media as an alternative method of measuring public sentiment. This social media approach is used as a supplement to traditional survey polling measures, with particular attention devoted to social network analysis and the memory effect of the information received. I focus specifically on the perception of information, rather than starting from the assumption that people are perfectly informed. I perform text sentiment analysis on breaking news collected from social media, and analyze its trends with added network effects, both active (retweets) and passive (favorites). I account for the fact that sentiment of news articles may have counter-effects, and that participants' engagement carries significant weight. Through the availability of high-frequency and real-time data from Twitter, I construct daily news sentiment measures. I collect all of the tweets from major mainstream media Twitter accounts in the interest of examining the relationship between the media and the public. The tweets are aggregated daily and trained on presidential polling

data using machine learning algorithms. Through the analysis I discuss the benefits of using data mining techniques over traditional survey approaches. Additionally, I evaluate whether the signals found in network activities of the President's own tweets contain any predictive capabilities.

The sentiment-based approach performs well, and it emerges that network effects, especially active network effects, are of highest significance. This quantitatively demonstrates the importance of online social networks and information dissemination in sentiment-based modelling. Additionally, it points to the importance of considering public perception of information rather than solely focusing on media and its possible biases. Undoubtedly, big data usage presents a valuable alternative, with the potential to overcome the limitations of survey approaches in upcoming years. It is cheaper, faster, and in-real time, and as such opens up possibilities for improving forecasting accuracy.

This work combines two diverse streams of literature: public opinion measurement, and news media and assessment of partisanship attitudes. Polls are traditionally conducted through surveys. Some of the most predominant flaws of survey approaches are last minute opinion preference changes and unrepresentative samples. Additionally, even well-constructed surveys are not completely immune to subjects' as well as interviewers' biases. Lastly, and most importantly, opinion polls are extremely fragile and can change quickly depending on events and information the public receives. Since surveys take time to construct, they do not capture instant changes. It is important to realize that public opinion can be easily shifted. One way to overcome these concerns is by using machine learning techniques. With machine learning algorithms it is possible to gauge the mood expressed in a certain piece of text, and adding network analysis can help us understand possible causes of shifts in public opinion.

Social media sentiment polling emerges as a valid alternative to traditional polling methods, and it has the potential to become increasingly better with the development of new and improved machine learning techniques. Most polls have exposure to certain kind of demographics that is increasingly narrow. Social media data can overcome these limitations. In today's digital era, people relate to each other primarily through social media platforms. Digital news media continues to grow rapidly. News reaches users even when they are not actively looking for information. Content reaches a wider audience faster, and subsequently communication on social media accelerates (Heo, Park, Kim & Park (2016), Ernst, Engesser, Büchel, Blassnig & Esser (2017)). According to reports from the Pew Research Center,[1] social media usage has noticeably grown among groups that have been underrepresented in survey data. Currently, people use a variety of social media platforms to receive news, as

[1]For more information on recent trends of social media usage: http://www.pewresearch.org/

well as to engage in civic-related activities. Increasingly, Americans have been taking part in political debates and online discussions inside their online social networks.[2]

Big data allows for the analysis of a broad range of economic activities and interactions, and it helps empirically quantify activities that researchers previously could not: social networks, information exposures and flows, news utilization, personal communications and geolocation data. Online social networks directly affect economic and social outcomes, and in today's economic and political spheres it is crucially important to examine the trends of how social media influences people's beliefs and perceptions, as well as the degree of polarization it creates across different spheres of social wellbeing. However, departing from traditional techniques and using data mining methods instead comes with its own challenges. For the purpose of this study, we deploy a neural network model that can be applied to unstructured, multidimensional data. With supervised learning, such a model is able to establish a correlation between input variables and daily presidential approval ratings.

This paper builds on the newly developed literature using social media to better forecast economic activities. If proven to be a reliable predictor, information from social media has an advantage of being relatively low-cost compared to traditional survey data, as well as providing the data in real time and at high frequency. Antenucci, Cafarella, Levenstein, Ré & Shapiro (2014) use real-time Twitter data to measure labor market flows. They do so by counting the specific terms within tweets that serve as an unemployment proxy, and they show that the social media index is a good predictor of the real-time state of job loss in the market when compared to the official statistics from the Bureau of Labor Statistics. The authors then construct indexes of job searches and job postings that are not well-measured otherwise but are important metrics in the labor market.

Sentiment analysis and opinion mining using Twitter data is a field that has attracted a lot of interest from researchers in recent years. It has been utilized in a wide range of applications, including electoral forecasting (Tumasjan, Sprenger, Sandner & Welpe (2010); Burnap, Gibson, Sloan, Southern & Williams (2016)), indicating social tensions (Burnap et al. (2016)), and suggested as a replacement for traditional surveys (O'Connor, Balasubramanyan, Routledge & Smith (2010); Bollen, Mao & Zeng (2011)). O'Connor et al. (2010) take a broader approach, analyzing the relationship between Twitter sentiment and consumer confidence and political opinion. Based on their results, they suggest replacing expensive and time-consuming polling with more efficient, simple data analysis gathered from online social media platforms. Tumasjan et al. (2010) analyze Twitter messages prior to the German 2009 elections and find that tweets reflect voters' preferences and come significantly close to

[2]http://www.pewinternet.org/2018/07/11/public-attitudes-toward-political-engagement-on-social-media/

predicting election results. Within the literature of opinion mining through social media, one stream of research investigates information dissemination through social networks (Stieglitz & Dang-Xuan (2013)), while another focuses more on finding patterns from tweets in forecasting major events. Specifically, this study attempts to fill the gap between the two: social media-based forecasting and dynamics of information diffusion through social networks.

In terms of incorporating mainstream media with opinion mining methods, my methodology is most closely related to the recent paper by Shapiro, Sudhof & Wilson (2018). They use text sentiment analysis of historical economic and financial newspaper articles to construct an index of newspaper sentiment for predicting economic activities. They find that such an index improves forecasting for macroeconomic variables, specifically inflation and the federal funds rate. I argue that previous research, by using historical articles published in certain news outlets, does not capture the sentiment of articles that draw the most reader's attention. Instead of including all the news articles and topics of interests, focusing only on news articles that are being circulated through social media represents a better sample selection, as this is the information being exposed to a wider audience and further disseminated through followers' networks.

As discussed above, using social media activity as a replacement for traditional opinion polls has been exploited in recent years. The majority of studies generalize on information volume and topics distributed through social networks, and use sentiment analysis to approximate average public opinion on certain topics of interest. I introduce an emotion-based model that emphasizes network effects as a stronger predictor than the volume of information and the sentiment itself. I test the empirical model through data mining techniques and by restricting the data to mainstream social media accounts only. The idea is that regardless of the sentiment of the news itself, the way people perceive and react to the news, and their act of distributing the news through their networks leads to improvement of the model accuracy.

The second area of literature incorporated in this paper addresses media's role in society and partisanship attitudes. Bonaparte & Kumar (2013) find that politically active people spend more time on news daily. Online social media platforms help users gather information at low or no cost. Media's role in society is to inform the public, but there are rising concerns about the quality of information people are provided with. The effects of media reporting and partisanship divide are amplified in the era of social media, where the information flows instantly and more people get exposed in a shorter period of time (Newman, Fletcher, Kalogeropoulos, Levy & Nielsen (2017)). Trends in most recent years capture the largest divide ever measured in partisan attitudes of support for the news media's watchdog role. The way news is presented to an audience could have counterintuitive effects, depending on

whether the audience perceives the news sentiment to be valid and unbiased. Our model builds on the assumption that public opinion on current economic and political issues depends on not just the sentiment of the information received, but on how the information is being reacted to- this depends on a receiver's own beliefs and biases and on how active she is at sharing the information further through her own social networks. Hence, the way that media presents the news and triggers a network proliferating effect amplifies its influence. As a result, network proliferated effects emerge as predominant ones.

Literature on information diffusion in social networks and how objective, and hence reliable, such information actually is, gives us insights about using big data to measure political polarization. This is important when looking at voting turnouts and policy preferences. Media bias remains difficult to quantify. There has been a lot of research focusing on two ideological mechanisms for media bias: issue filtering and issue framing. Issue filtering relates to what topics are chosen to be spoken about by news media outlets, and issue framing deals with topic selections in articles. All prior work confirms ideological divisiveness among US news outlets, and news outlets can be ordered on a conservative-to-liberal spectrum. The most comprehensive work to date has been provided in Gentzkow & Shapiro (2010). They construct a new index of media slant by algorithmically selecting the phrases from congressional speeches that are associated strongly with either Republicans or Democrats, and then use those to measure the frequencies of selected terms in news outlets.

The vast majority of political news shared over social networks comes from professional news sources, with established news brands shared most widely. They influence the largest share of people (Chomsky (1997)). Therefore, I specifically direct my attention to mainstream media accounts and information shared through those channels. I test whether a data-driven model of opinion dynamics is able to accurately forecast public sentiment from active online participants. The news reaches users faster, but news diffusion through the networks, as well as belief updates, take time. As the model predicts that network effects are important, I improve upon survey methods by not just implementing faster, cheaper and real-time measures, but also by obtaining significantly higher forecasting accuracy.

By combining the importance of correctly identifying and measuring the public opinion with the role of news media in society and people's partisanship attitudes, this paper examines not just the content and the sentiment of the information provided, but also how the general audience responds to it. The purpose of the proposed model is to perform both, polling prediction and opinion analysis. In addition to evaluating news media channels of information, I also include an analysis of self-promotion. The network effects of a political

candidate's outreach emerge as strong signals. Self-promotion is important for quantitively assessing social network channels between candidates, news media and the public.

The study is organized as follows. In section 2, I discuss training and testing data. In section 3, I discuss methods used for acquiring data and construction of daily news sentiment measures. In section 4, I present the model. In section 5, I provide descriptive analysis of the model variables. In section 6, I present empirical findings and effects of media sentiment and networks. In section 7, I examine whether news sentiment and network indicators measures improve the forecasting performance and predictive accuracy. In section 8, I perform variable selection by implementing Lasso. In section 9, I introduce a neural network model to address nonlinearity of emotion-based predictions. In section 10, I quantify the effects of self-promotion in online context. Section 11 concludes.

2. Data

For analysis I use only news shared through social media channels, specifically Twitter. The advantages of using social media for the purpose of this research are that it provides some meaningful insights and real-time sentiments, but downsides are that we are limited in scope of what could be captured through social networks. It is extremely important to organize the data in a way that reduces multidimensionality.

Evaluating the news exposure to certain economic and political activities from several news media streams, I use the metrics of likes and retweets on the subject scaled to population. I quantitatively evaluate the amount that news media exposes active users to the information, and the tone in which the information presents itself. Furthermore, I can evaluate how active the networks are given the sentiment of the presented information.

2.1. **Twitter Data Cleaning.**

For analysis I use one of the most popular networking services worldwide, Twitter. Twitter is one of the fastest growing social media platforms that increasingly serves as a means by which people obtain some or all of their news and information. In the most recent years, Twitter serves as a platform that millions of people use daily to share their voice in regards to civic-related issues, as well as a place of increased political engagement and debate. One of the most predominant strengths of Twitter is real-time distribution of user activity, hence I argue that is the focal point for why and in what ways we could use such real-time data, and what its purpose and impact could be.

Twiter's developer platforms offer a variety of APIs *(Application Programming Interface)*, allowing for easy and diverse usage of Twitter as a data source for academic research purposes. Twitter's microblogging service allows its users to share small messages popularly known as tweets. Initially, tweets were limited to 140 characters, but as of November 7, 2017 this limit

has been doubled in length. However, this change has had little impact on the length of Twitter posts in general.

First, I extract all tweets and relevant metadata from specific media outlets' Twitter accounts that contain the key index of interest under study. I have 32 months of Twitter data that spans from the beginning of the presidency in January 2017, until September 2019. I recover tweets using Tweepy[3], Python library enabling the extraction of 316663 tweets filtered by relevant keyword of the index of interest. For the purpose of this analysis, the keyword *Trump* has been used. By using Tweepy, I am able to extract all the information and variables[4] of interest available for a specific tweet ID. Tweepy uses the publicly available Twitter Streaming API.

Second, I extract media metadata. The tweets have been collected from 25 major US news media outlets, including CNN, Fox News, WSJ, Washington Post, and others shown in Figure 1. The variables[5] of media outlets and specific tweets containing the chosen keywords have also been extracted using Twitter Streaming API. It is important to remark that media outlets have been chosen rather subjectively, with reference to Pew Research survey data study of ideological placement of news media' source audience.[6] This study simply relies on using widely-known news outlets that capture the largest share of audience on Twitter.

2.2. **Polling Data.** Survey data are not the most accurate representation of public opinion, however they are the best approximation we have at the moment. For political opinion, we choose two polls, Gallup and Rasmussen presidential approval rating. They are shown in Figures 2 and 3, respectively. These polls were chosen based on their popularity and availability. Rasmussen data is the only daily poll available publicly. Parsing of this poll data was performed using the Beautiful Soup[7] Python library. Gallup data was collected manually. One year of Gallup data was collected, as they discontinued daily reporting starting in January 2018. These polling data represent the daily approval/disapproval of President Trump's job performance collected from registered U.S. voters. Rasmussen further breaks down the approval measure into "strongly approve" and "strongly disapprove", and their approval index is calculated as the difference between the two. Gallup data captures only total approve and total disapprove. When analyzing the data, I consider the approval index for Rasmussen as the difference between total approve and total disapprove, to remain

[3]Python library used for accessing the Twitter API: http://www.tweepy.org/
[4]Appendix Table 1: Metadata of Tweets
[5]Appendix Table 2: Media outlets variables' metadata extracted using Twitter Streaming API
[6]More information about Pew Research political polarization and media bias study: http://www.journalism.org/2014/10/21/political-polarization-media-habits/
[7]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

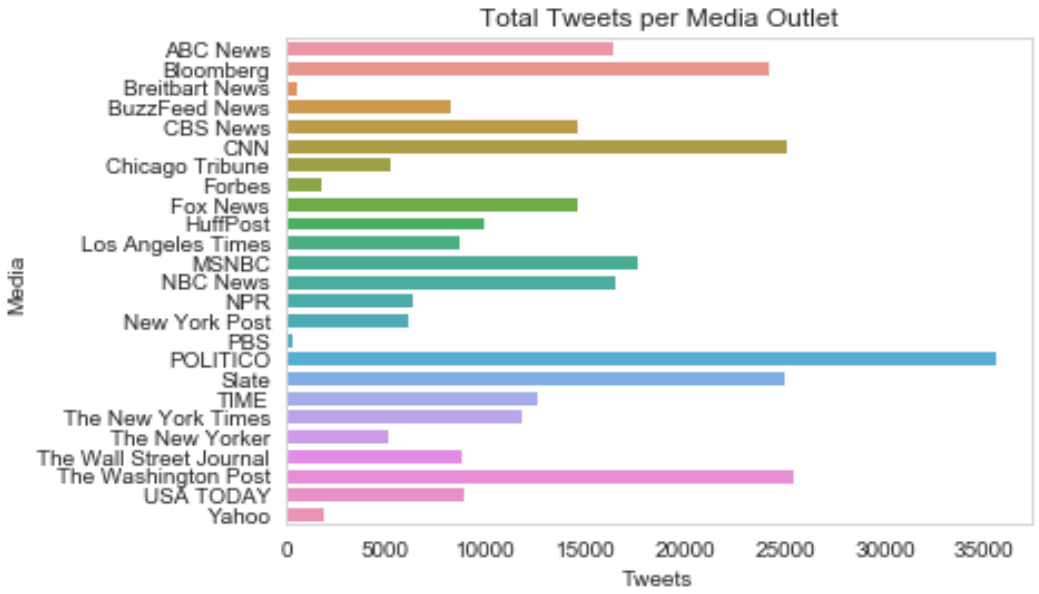

FIGURE 1. News media outlets' Twitter accounts and the total number of tweets per media outlet between January 1, 2017-September 1, 2019.

consistent between the different data sets. There is no reporting on holidays, and this has been accounted for in the process of data cleaning.

I use standard smoothing techniques of n days rolling average to suppress noise in polling data. The same smoothing technique is performed with data obtained from Twitter. I smooth aggregate sentiment ratio with n = 7 and 14 days being the number of time periods in the average.

$$\overline{y_t} = \frac{y_t + y_{t-1} + ... + y_{t-n-1}}{n}$$

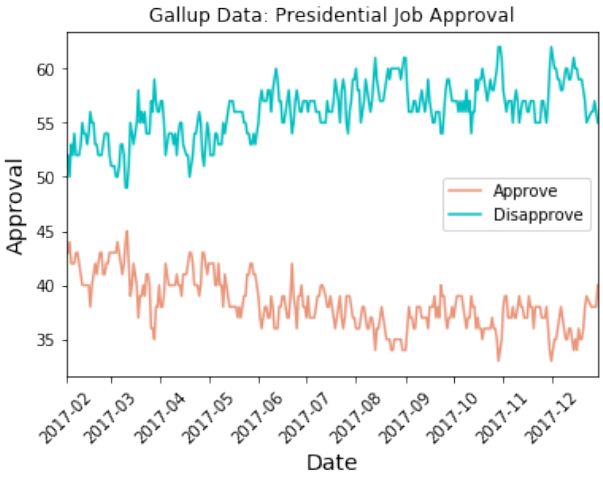

FIGURE 2. Gallup Organization: 2017 polling data

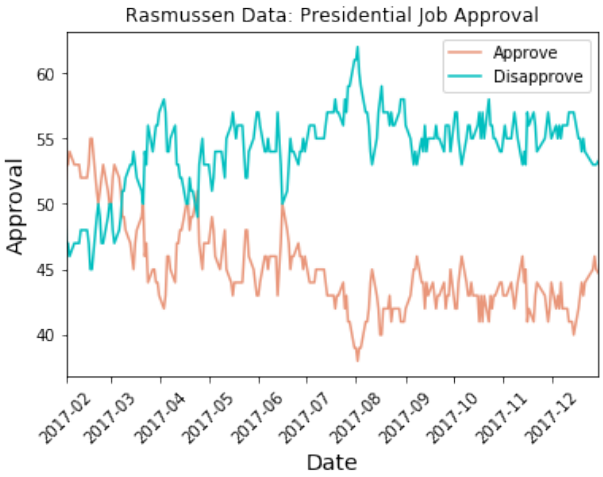

FIGURE 3. Rasmussen Reports: 2017 polling data

3. METHODS

3.1. **Message Retrieval.** Only tweets (and retweets) created and shared through the Twitter accounts of major US news media outlets have been considered. It is worth noting that chosen media outlets almost always tweet the breaking news for a given day. I implement text sentiment analysis only on tweets themselves, not accounting for sentiment of the entire article that has been linked. The assumption is that the tweet most likely conveys the message and sentiment of the given article.

I retrieved all the messages in the period between January 1st, 2017 until September 1st, 2019 that satisfy given criteria. For presidential approval, we use a topic keyword "Trump" in our search. For network analysis, the number of followers of a news media outlet is considered as a *Network Exposure Factor*. We use the metrics 'favorites' as a *Passive Indicator* and 'retweets' as an *Active Indicator*. The tweets are first cleaned using the regular expression operations module,[8] which cleans the text in a tweet by removing all links and tags. Other special characters are retained for VADER sentiment analysis, as this type of analysis takes those characters into account when scoring the text sentiment. Next, tweets are analyzed using the text sentiment tools. Further, the data is preprocessed using StandardScaler.[9] It

[8]https://docs.python.org/2/library/re.html
[9]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

standardizes inputs to have zero mean and unit variance. Standardization of a dataset is very common for machine learning estimators, since the variables are of different magnitudes.

3.2. **Text Sentiment Analysis.** A sub-field of Natural Language Processing (NLP), Sentiment Analysis, also known as Opinion Mining, is the method used for analyzing the opinions expressed in text on any subject of interest. Most sentiment analysis approaches take one of two forms: *polarity-based*, where pieces of texts are classified as either positive or negative; or *valence-based*, where the intensity of the sentiment is taken into account. Lexicons are in general time-consuming and expensive to produce. The sentiment-based approach works well when there is a good fit between lexicon and the text.

TextBlob is an easily accessible Python sentiment analyzer. It provides polarity and subjectivity scores. Subjectivity score is a float within the range [0,1] where 0 is very objective and 1 is very subjective. Once the piece of text is classified as either objective or subjective, it can be further classified based on polarity- whether it expresses positive or negative sentiment.

VADER (*Valence Aware Dictionary and Sentiment Reasoner*) sentiment analyzer is based on lexicons of sentiment-related words, and is specifically used for sentiments expressed through social media. VADER sentiment analyzer relies on a dictionary constructed using human raters from Amazon Mechanical Turk. In addition to lexical features, VADER also differentiates between sentiments by considering five heuristics: punctuation, capitalization, degree modifiers, shift in polarity due to the word "but", and tri-gram examination to identify negation. VADER produces four sentiment metrics from the word ratings. Three of those: positive, neutral and negative; represent the proportion of the text that falls into those categories. The final metric, compound score, is the sum of all of the lexicon ratings, which have been standardized to range between -1 and 1. As opposed to TextBlob, VADER does not simply match between lexicon and text. It also takes into account the way the words have been written and expressed, accounting for intensity in that aspect as well.

Since a piece of text may, and most often does, contain multiple sentiments all at once, valence-based approach is the most reliable to use. Hence, I believe that VADER provides more valid ratings, accounting for the intensity of the sentiments expressed through text.

4. Opinion Estimation Model

My main objective is to construct a daily index of presidential approval based on news sentiment, media activity, and network effects. To extract daily sentiment factor, the proposed empirical model is as follow:

$$\text{SentimentFactor}(t) = \beta_t^{sent} \sum_i s_{i,t} \frac{\omega^n}{\Omega} + \beta_t^{net} \sum_n N_{n,t}^{tweets} \frac{\omega^n}{\Omega} + \beta_t^{pas} \sum_i s_{i,t} \omega_{i,t}^{pas} + \beta_t^{act} \sum_i s_{i,t} \omega_{i,t}^{act} \tag{1}$$

where i denotes tweet index, and s_i is a sentiment score for tweet i. Hence, β^{sent} is the sentiment factor for a given tweet. β^{net} is overall media news outlet activity factor, measured by daily volume of tweets. We select n media outlets (*The Wall Street Journal, Fox News, CNN, The Washington Post, etc.*). N_n^{tweets} is the total volume of tweets for media outlet n. Both sentiment and volume are dependent on the media outlet size and exposure, and as such, tweets are weighted by the market size for both sentiment and activity factors. ω_n is the market of n news outlet for all relevant news issued. Our total market, Ω, is simply a sum of ω_n for all tweets, $\sum_n \omega_n$.

β^{pas} and β^{act} are network passivity and network activity factors, respectively. ω_i^{pas} is passive network response for tweet i, calculated by the number of likes that are averaged over the sum of tweets and aggregated daily. The same methods of aggregation are used for the number of shares, which are represented as the active network response for tweet i, ω_i^{act}. Analysis is limited to tweets with a minimum of 5 retweets and 10 favorites.

The model of media sentiment index at time t, with j representing the number of lag periods is:

$$\text{SentimentIndex}(t) = \sum_{j=0}^{p} (\text{SentimentFactor}(t-j)) - \text{Media}_{FixedEffect} + \epsilon \tag{2}$$

In equation 2, fixed effects account for media bias, calculated by the following equation:

$$\text{Media}_{FixedEffect} = \sum_n \text{b}^n \frac{\omega^n}{\Omega} \tag{3}$$

In (3), b_n is a vector of size n, one value of media fixed effect for each outlet. For the purpose of the empirical analysis, I equate media effect to zero, but the proposed model allows for expansions.

5. Descriptive Analysis

Figures 4 and 5 show plots of the two primary measures of volume of tweets and network effects over time. While crude, the volume metric still provides some intuition about general trends and popularity of the chosen topic of interest. In Figure 5 we notice that both network effects, passive (favorites) and active (retweets), tend to spike during the dates of key events, such as tax plan cuts, government shutdowns, etc.

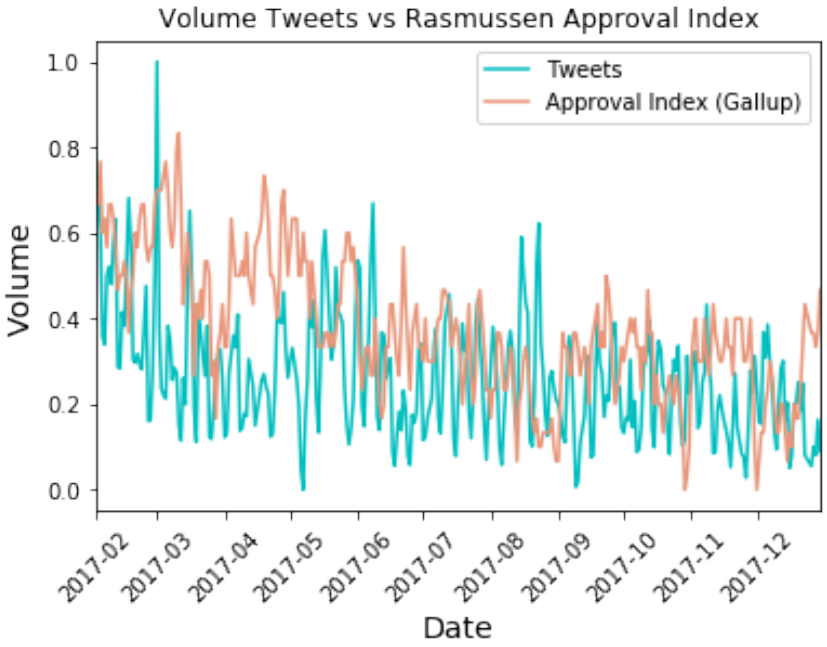

Figure 4. Presidential Approval Volume Metric

Figures 6 and 7 display some interesting dynamics between different media outlets. It shows differences between left- and right- leaning media outlets, for active and passive network effects. We can see that some media outlets such as CNN (which used to be placed closer to the center of the media bias spectrum by ideological placement of their audience[10]) trend similarly to strongly left-leaning media networks.

[10]http://www.journalism.org/2014/10/21/political-polarization-media-habits/

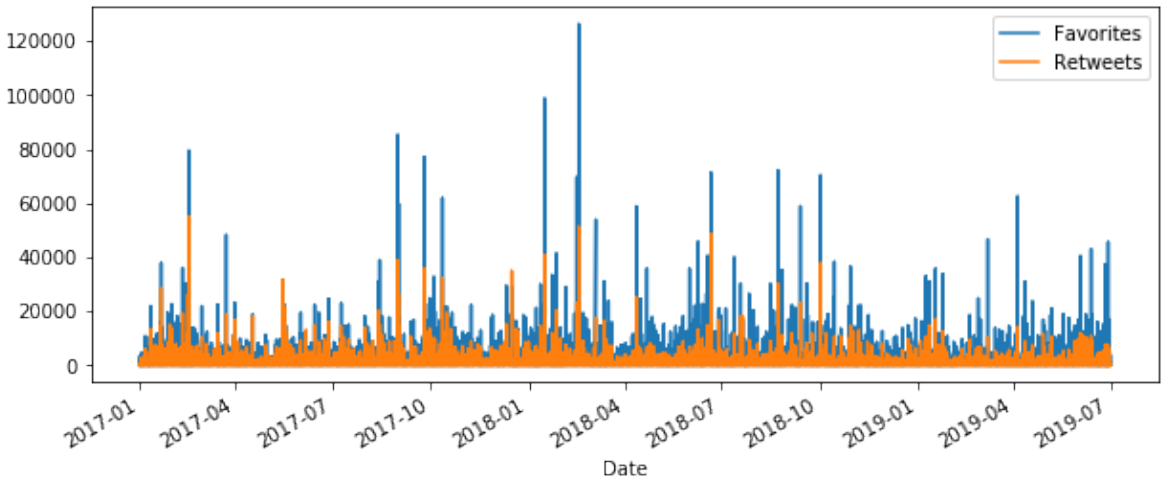

FIGURE 5. Network Indicators Time Series

Left-leaning ooutlets get more favorites for negative sentiment tweets, with opposite dynamics for right-leaning news, where positive tweets get more favorites. For the active network effect, retweets, we see similar dynamics. Washington Post negatively-scored tweets are retweeted more than positively-scored ones, while the opposite is true for Fox News. The complete breakdown of audience response to positive vs. negative sentiment tweets per media outlet can be found in the Appendix.[11]

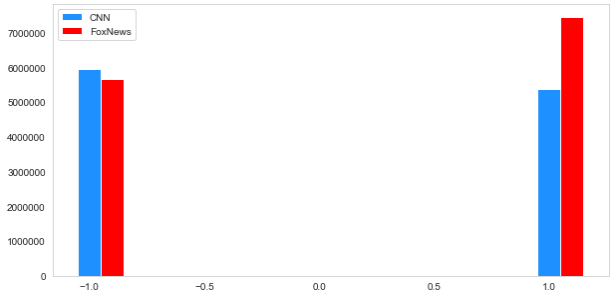

FIGURE 6. Passive Network Effect (Number of favorites): Difference between left- and right-leaning media outlets

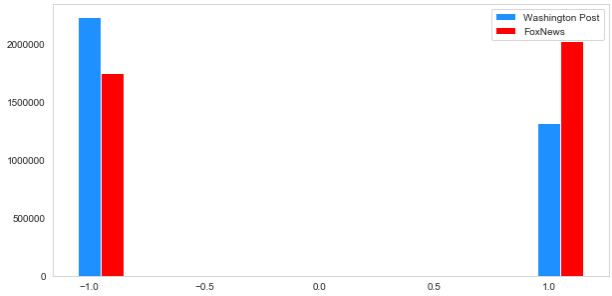

FIGURE 7. Active Network Effect (Number of retweets): Difference between left- and right-leaning media outlets

[11]Appendix Figure 13 and Figure 14: Audience response to Sentiment per Media Outlet

6. Estimating the Effects of News Sentiment and Network Effects

First I examine whether there is correlation between poll indicators and the text sentiment of breaking news reported daily, and most importantly, whether there is an impact when network effects are included as well. I look for direction and size of the impact, if any.

Table 1. Regression Estimate Results

	(1)	(2)	(3)	(4)
$\beta_{p=1}^{net}$	2.179***	0.963***	0.597	1.81***
	(0.415)	(0.329)	(0.38)	(0.573)
$\beta_{p=2}^{net}$		2.372***	2.983***	1.918***
		(0.367)	(0.418)	(0.64)
$\beta_{p=1}^{sent,n}$	-5.263***	-3.588***	-3.961***	-4.098***
	(0.470)	(0.317)	(0.431)	(0.436)
$\beta_{p=2}^{sent,n}$		-2.642***	-4.262***	-4.503***
		(0.35)	(0.482)	(0.483)
$\beta_{p=1}^{sent,p}$				-1.413***
				(0.538)
$\beta_{p=2}^{sent,p}$				1.489***
				(0.56)
$\beta_{p=1}^{pas}$	3.626***		1.474**	1.877***
	(0.620)		(0.599)	(0.605)
$\beta_{p=2}^{pas}$			6.048***	5.987***
			(0.572)	(0.61)
$\beta_{p=1}^{act}$	-4.129***		-1.777***	-1.969***
	(0.658)		(0.62)	(0.62)
$\beta_{p=2}^{act}$			-6.931***	-7.211***
			(0.598)	(0.616)
R-squared	0.383	0.400	0.539	0.549
N	588	588	588	588
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 1 reveals primary regression estimate results, using rolling averages of 7 days. Here I show a few models of approval rating, with one an two periods. This is a media-based model as the main channel of information flow in social network diffusion. The first model is one-period base model. The second model is two-period model with only the sentiment factor. In the third model I introduce network effects, passive and active. As a result of this, the R-squared significantly improves, from 0.400 to 0.539. The results show that all of the variables from our opinion estimation model are statistically significant. Scores of the

negative sentiment tweets show much stronger effects than positive sentiment scores. The direction of impact is negative, as expected. Negative sentiment breaking news might induce worry and affect public perception of the given topic of interest.

For the network factors, both passive and active exhibit up to six times stronger lagged effects. Lagged active network response has the most predominant effect, followed by lagged passive network response and negative media sentiment score. The direction of the impact of network effects provides some intuitive information. Active social media participants are more likely to favorite tweets that convey information they agree with. It is unclear whether the public is spreading the information to reach a wider audience by retweeting more news that has negative sentiment that they agree with, or whether they retweet more news with positive sentiment that they do not agree with (which could ultimately denote negative sentiment when spreading it through their own network). The intuition would be that the first case is predominant, but this has yet to be proven empirically.

The finding that lag effects are in fact predominant is not surprising, as information diffusion through social networks takes time to reach a wider population.

7. Forecasting Performance

In this exercise I seek to determine whether the model of sentiment analysis with included network effects has any predictive information about future values of the polls. The proposed empirical model includes *Granger causality*:

$$\text{ApprovIndex}_{t+h} = \alpha + \sum_{j=0}^{p} \beta_j^k \text{ApprovIndex}_{t-j} + \tag{4}$$
$$\sum_{j=0}^{p} (\beta_j^{sent} \sum_i s_{i,j} \frac{\omega^n}{\Omega} + \beta_j^{net} \sum_n N_{n,j}^{tweets} \frac{\omega^n}{\Omega} + \beta_j^{pas} \sum_i s_{i,j} \omega_{i,j}^{pas} + \beta_j^{act} \sum_i s_{i,j} \omega_{i,j}^{act})$$

where i denotes article index, j is the summation over the lagging p periods, and h indicates how many periods in the future to which the prediction applies.

From Figure 8 it is apparent that the model of sentiment ratio with active and passive network shocks captures the broad trends in presidential approval survey data. Regression results are presented in Table 2. When looking two periods in the future, including the model specification improves R-squared from 0.554 to 0.738. Hence, Granger analysis shows that the past values of the network effects are important beyond past values of the approval index alone. The only variables that are not statistically significant are positive sentiment score and market indicator of tweet volume. Interestingly, lagged coefficients of network indicators

TABLE 2. Granger Causality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\beta_{p=1}^{net}$		0.168		0.092		-0.781***		-1.212***
		(0.265)		(0.288)		(0.3)		(0.291)
$\beta_{p=2}^{net}$		0.843***		1.293***		1.474***		0.658**
		(0.273)		(0.289)		(0.294)		(0.263)
$\beta_{p=3}^{net}$								2.505***
								(0.26)
$\beta_{p=1}^{sent,n}$		-2.243***		-2.349***		-1.305***		-1.0***
		(0.311)		(0.337)		(0.359)		(0.327)
$\beta_{p=2}^{sent,n}$		-1.076***		-2.174***		-2.578***		-1.738***
		(0.318)		(0.326)		(0.331)		(0.301)
$\beta_{p=3}^{sent,n}$								-3.754***
								(0.298)
$\beta_{p=1}^{pas}$		1.974***		1.626***		1.619***		2.015***
		(0.427)		(0.465)		(0.478)		(0.425)
$\beta_{p=2}^{pas}$		1.232***		3.325***		3.621***		3.224***
		(0.447)		(0.467)		(0.48)		(0.435)
$\beta_{p=3}^{pas}$								2.455***
								(0.444)
$\beta_{p=1}^{act}$		-2.161***		-1.756***		-1.005**		-1.15***
		(0.444)		(0.484)		(0.497)		(0.441)
$\beta_{p=2}^{act}$		-1.266***		-3.483***		-3.953***		-3.286***
		(0.454)		(0.474)		(0.485)		(0.441)
$\beta_{p=3}^{act}$								-3.549***
								(0.439)
$y_{h=1}$	4.267***	3.126***						
	(0.127)	(0.144)						
$y_{h=2}$			3.921***	2.524***				
			(0.149)	(0.139)				
$y_{h=3}$					3.842***	2.5***	3.993***	1.908***
					(0.158)	(0.141)	(0.156)	(0.138)
R-squared	0.664	0.772	0.554	0.738	0.522	0.735	0.554	0.814
N	574	574	560	560	546	546	532	532

Note: *p<0.1; **p<0.05; ***p<0.01

have higher values than the past values of the approval index itself.

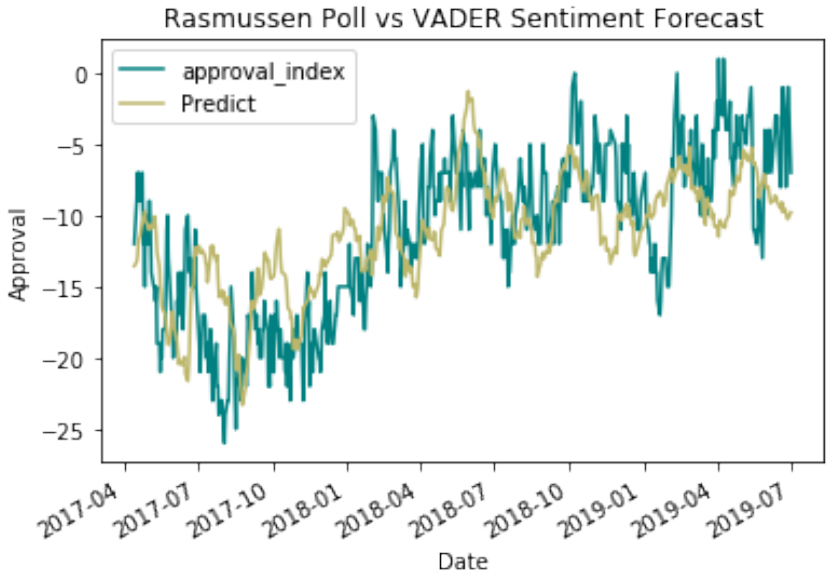

FIGURE 8. Granger Causality- Sentiment information and network factors capture broad trends in the polling data.

The comparison between Vader and TextBlob regression results are presented in Table 3. TextBlob reports polarity score, hence for the purpose of comparing the two, we use the compound score of Vader sentiment analyzer. Both of these support the hypothesis that adding network effects increases the model performance. Although both sentiment measures are consistent in the direction of the impact, they do somewhat differ in the intensity of the current vs lag effects.

TABLE 3. Forecasting Performance: Vader vs. TextBlob Text Sentiment Analyzer

	(1)	(2)	(3)	(4)
$\beta_{p=1}^{net}$	-2.556***	-2.554***	-1.679***	-1.78***
	(0.219)	(0.184)	(0.159)	(0.136)
$\beta_{p=2}^{net}$	-0.795***	-0.66***	-0.692***	-0.858***
	(0.223)	(0.179)	(0.151)	(0.132)
$\beta_{p=1}^{sent}$	0.506	1.559***	0.756***	0.642***
	(0.319)	(0.277)	(0.235)	(0.213)
$\beta_{p=2}^{sent}$	1.05***	1.047***	0.512**	0.298
	(0.329)	(0.242)	(0.234)	(0.204)
$\beta_{p=1}^{pas}$	1.071	2.252***	1.279**	2.199***
	(0.688)	(0.532)	(0.506)	(0.4)
$\beta_{p=2}^{pas}$	6.242***	3.811***	2.713***	2.03***
	(0.674)	(0.543)	(0.525)	(0.41)
$\beta_{p=1}^{act}$	0.109	-2.19***	-0.951*	-1.611***
	(0.694)	(0.512)	(0.51)	(0.383)
$\beta_{p=2}^{act}$	-6.174***	-2.66***	-2.307***	-0.85**
	(0.672)	(0.508)	(0.508)	(0.383)
$y_{h=2}$			2.847***	2.457***
			(0.145)	(0.139)
R-squared	0.4	0.55	0.695	0.768
N	588	588	560	560
Note:				*p<0.1; **p<0.05; ***p<0.01

8. Variable Selection

To determine which variables are the most prevalent in the proposed model, we implement LASSO (*Least Absolute Shrinkage and Selection Operation*). LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance prediction accuracy and interpretability:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} (Y_i - \beta_0^z - \beta^{z\top} X_i)^2 \quad \text{subject to} \quad \sum_{k=1}^{m} \lambda \, |\beta_k^z| < s \tag{5}$$

where z is the economic variable of interest, i element in the time series of predictors and Y_i and X_i are m-dimensional vectors, one dimension for each predictor variable. Lasso regression performs $l1$ regularization: it adds a penalty equivalent to the absolute value of the magnitude of the coefficient. We select the tuning parameter, *lambda*, based on minimizing cross-validated generalization error. We find our optimal lambda value is 0.003 (Figure 9).

Lasso selects some features while reducing the coefficient of others to zero. For our optimal tuning parameter, the prevailing predictors[12] emerge: lagged values of two-week moving averages for network passivity factor, two-week moving averages of network activity factor, and negative sentiment shock in current period. In fact, almost all of the model parameters emerge as relevant. The only parameter that is suppressed to zero is current period media focus (volume of tweets).

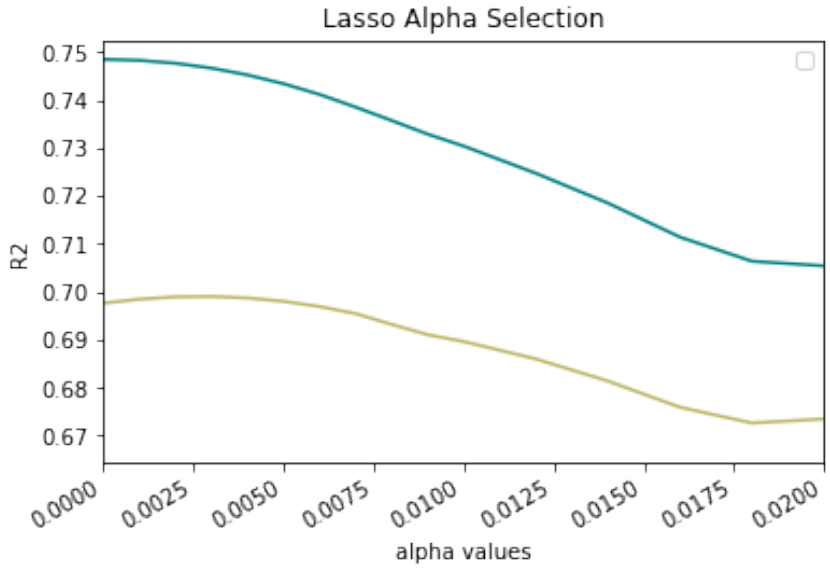

FIGURE 9. Lasso Optimal Lambda Selection

9. Non-linear Model for Emotion-based Opinion Prediction

In the last phase, I develop a multi-layer neural network model with a gradient-based learning algorithm. Gradient descent relies on the fact that it is much easier to minimize a reasonably smooth continuous function instead of a discrete function. Dependencies are not entirely linear, hence neural network regression can improve performance.

The architecture of the neural network consists of 11 layers with 4 inputs: 10 hidden layers of 10 neurons each, and one node output layer. Our input layer is a vector X that consists of 4 variables: market weight, sentiment score, and passive and active network indicators. Hidden layers are where all the computation happens and they represent so-called "activation" nodes. Each node combines the input data with a set of weights. These input-weights are summed up, and passed through the chosen activation function. If the

[12]Appendix Table 7: Lasso Variable Selection. Features with non-zero weght, and suppressed features.

node encounters sufficient stimuli, the signal is passed further through the network. All of the activated signals get stronger the further they pass through the network. Ultimately, those inside the hidden layers activations affect the output. The output layer is the predicted value of approval rating.

Our model function is as follows:

$$\mathrm{Y} = f(\sum_{i=1}^{n} w_i x_i + \mathrm{b}) \tag{6}$$

Where Y is recognized class label of pattern x_i, x_i is input pattern, w_i is the weight, b is the bias and f is the activation function.

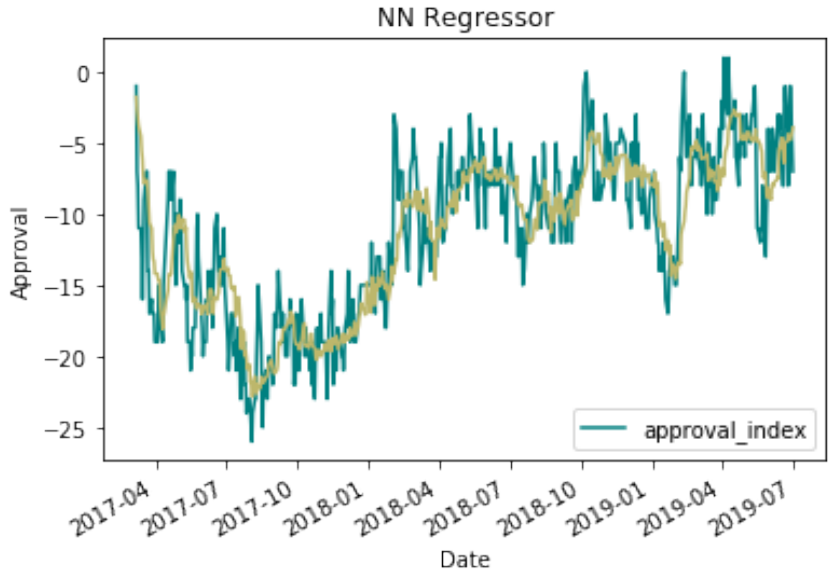

FIGURE 10. Neural Network Regresssor Performance. Yellow line represents the model predictions.

The non-linear transforms at each node determine whether the signal would be switched on, and such activation functions are usually *s*-shaped. Our chosen activation function is *tanh*, and this sigmoidal function puts all the values within the range [-1,1].

The ultimate goal is to find the parameters that minimize the error, and that is what is known as "learning". I use a mean squared error loss function. The idea is to minimize the loss function by estimating the impact of small variations in parameter values on the loss function. This is measured by the gradient of the loss function with respect to the

parameters. Usually, when overfitting occurs, the training error decreases over time, while the test error starts increasing after a certain number of iterations and passing through a minimum. We divide the data randomly, 70% for training and 30% for testing. A plot of mean squared error loss over network size when optimizing the mean squared error loss function is shown in Figure 11.
The neural network is 95% accurate in predicting daily changes in the presidential approval rating (Figure 10).

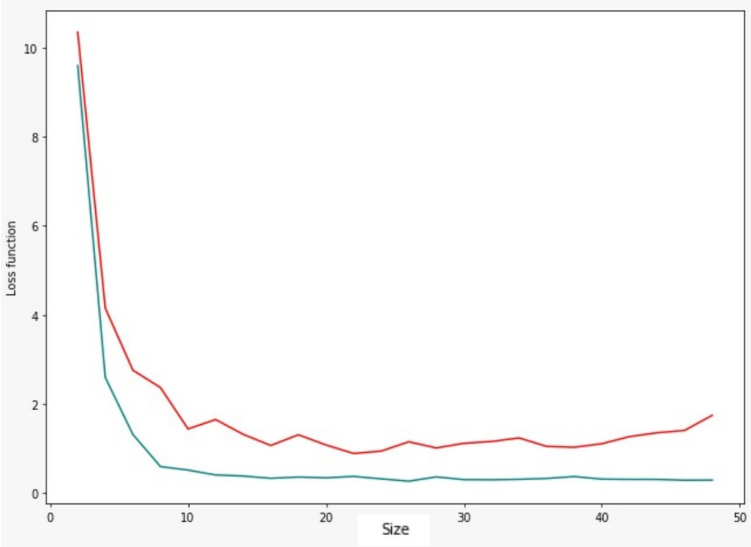

FIGURE 11. Loss Function showing the mean squared error loss over the network size (hidden layers, neurons) for the train (blue) and test (red) sets.

10. Self Promotion

The previous exercises considered news media as the main source of information. Here, I add data from the President's Twitter account as well to evaluate whether this individual information source has any predictive power for public opinion polling.

10.1. **Effects of Self-Promotion on Predictive Capabilities.** Self-promotion shows a competing effect to media influence. I show eight models in Table 4: the first two are one-period and demonstrate the improvement in prediction capability when self promotion is included into the model. Including self-promotion improves predictive capability, increasing the R-squared from 0.506 to 0.587. The effect does not seem dependent on the volume of tweeting in a one-period model. Network effects are very important, with active network effects emerging as predominant.

One can notice opposite sign betas for the active network sentiment between self and media promotion. This could be due to the fact that self-tweets are mostly promoted in their network by their supporters, therefore carrying a positive message for the subject, while media in the current sample is mostly carrying negative sentiment (Number of Tweets Negative Sentiment / Number of Tweets Positive Sentiment =1.22).

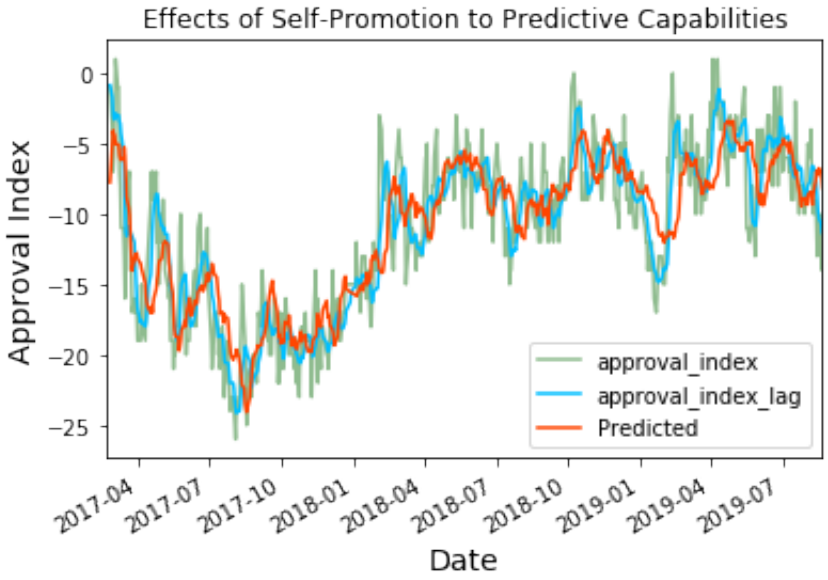

FIGURE 12. Effects of Self-Promotion on Predictive Capabilities. Red line represents the model predictions.

10.2. **Delayed Effects of Self-Promotion and Media Influence.** The predictive capabilities are significant for the effect of previous approval ratings on the current period. When including two periods, *current* (0 through 7 days) and *previous* (7 through 14 days), one can see that the previous period is important for self-promotion, while the current period is not. For media influence, the current period is important while the previous is not.
This might suggest that self-promotion and its network effects triggers media response that in turn influences the population. The effect is enhanced by media influence but triggered by self-promotion.

TABLE 4. Effects of Self-Promotion to Predictive Capabilities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\beta^{net}_{self,p=1}$		-0.804		-0.902*				-0.496
$\beta^{net}_{self,p=2}$				-0.343				-0.268
$\beta^{net}_{media,p=1}$	-0.033	-0.142		1.026**				0.062
$\beta^{net}_{media,p=2}$				-0.344				-0.749**
$\beta^{sent,n}_{self,p=1}$		-1.155***		-0.063	-0.24			
$\beta^{sent,n}_{self,p=2}$				-0.474**	-0.598***			
$\beta^{sent,n}_{media,p=1}$	-1.593***	-1.109***		-0.81***	-0.608***			-0.799***
$\beta^{sent,n}_{media,p=2}$				-0.426	-0.523**			-0.344
$\beta^{sent,p}_{self,p=1}$		-0.987*		0.318				
$\beta^{sent,p}_{self,p=2}$				0.046				
$\beta^{sent,p}_{media,p=1}$	3.303***	4.405***		-0.612*				
$\beta^{sent,p}_{media,p=2}$				0.534				
$\beta^{pas}_{self,p=1}$		-3.493**		-0.828	-0.679	0.445		-7.467***
$\beta^{pas}_{self,p=2}$				-4.082***	-3.806***	-2.605**		-2.385*
$\beta^{pas}_{media,p=1}$	3.738***	1.244**		2.635***	3.058***	3.057***		2.49***
$\beta^{pas}_{media,p=2}$				0.706	0.82*	0.446		2.745***
$\beta^{act}_{self,p=1}$		8.205***		0.676	-0.085	-1.428		7.316***
$\beta^{act}_{self,p=2}$				4.599***	4.106***	2.547**		2.663**
$\beta^{act}_{media,p=1}$	-7.463***	-4.983***		-3.687***	-3.809***	-4.36***		-2.63***
$\beta^{act}_{media,p=2}$				-0.268	-0.303	-0.331		-2.089***
$y_{h=1}$			4.409***	3.399***	3.377***	3.524***	4.409***	2.997***
$y_{h=2}$			0.297	0.761***	0.882***	0.873***	0.297	0.111
R-squared	0.506	0.587	0.745	0.827	0.822	0.813	0.745	0.811
N	617	617	624	624	624	624	624	617

Note: *p<0.1; **p<0.05; ***p<0.01

11. CONCLUSION

In this paper, I explore an alternative method for measuring public opinion that could replace time consuming and expensive survey-based measures. I present empirical results for

media effects, as well as the effect of self-promotion on public opinion through an alternative opinion estimation model. I build on improving the methods of text sentiment analysis, with the primary focus being improvement of model specification by accounting for media focus and proliferated network effects in addition to sentiment shocks. Sample selection is another feature of the model that captures important trends. The model performs well forecasting the daily approval rating up to several days in advance. Hence, this paper shows promising results for a method of measuring public opinion based on social network activities. Accounting for biases and having timely opinion-based models with better forecasting power would undoubtedly help policymakers when formulating forward-looking policies.

References

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C. & Shapiro, M. D. (2014), Using social media to measure labor market flows, Technical report, National Bureau of Economic Research.

Bollen, J., Mao, H. & Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal of computational science* **2**(1), 1–8.

Bonaparte, Y. & Kumar, A. (2013), 'Political activism, information costs, and stock market participation', *Journal of Financial Economics* **107**(3), 760–786.

Burnap, P., Gibson, R., Sloan, L., Southern, R. & Williams, M. (2016), '140 characters to victory?: Using twitter to predict the uk 2015 general election', *Electoral Studies* **41**, 230–233.

Chomsky, N. (1997), 'What makes mainstream media mainstream', *Z magazine* **10**(10), 17–23.

Ernst, N., Engesser, S., Büchel, F., Blassnig, S. & Esser, F. (2017), 'Extreme parties and populism: an analysis of facebook and twitter across six countries', *Information, Communication & Society* **20**(9), 1347–1364.

Gentzkow, M. & Shapiro, J. M. (2010), 'What drives media slant? evidence from us daily newspapers', *Econometrica* **78**(1), 35–71.

Heo, Y.-C., Park, J.-Y., Kim, J.-Y. & Park, H.-W. (2016), 'The emerging viewertariat in south korea: The seoul mayoral tv debate on twitter, facebook, and blogs', *Telematics and Informatics* **33**(2), 570–583.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. & Nielsen, R. K. (2017), 'Reuters institute digital news report 2017'.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A. (2010), From tweets to polls: Linking text sentiment to public opinion time series, *in* 'Fourth International AAAI Conference on Weblogs and Social Media'.

Shapiro, A. H., Sudhof, M. & Wilson, D. (2018), Measuring news sentiment, Federal Reserve Bank of San Francisco.

Stieglitz, S. & Dang-Xuan, L. (2013), 'Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior', *Journal of management information systems* **29**(4), 217–248.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010), Predicting elections with twitter: What 140 characters reveal about political sentiment, *in* 'Fourth international AAAI conference on weblogs and social media'.