BI Norwegian Business School - campus Oslo

GRA 19703

Master Thesis

Thesis Master of Science

Extracting Sentiment of Selected Twitter Accounts and Considering Its Relationship with the S&P 500 Index

Navn:	Henrik Sveen, Sepehr Cyrusian

Start:	15.01.2021 09.00
Finish:	01.07.2021 12.00

Sepehr Cyrusian Henrik Sveen

BI Norwegian Business School Master Thesis

Extracting Sentiment of Selected Twitter Accounts and Considering Its Relationship with the S&P 500 Index

Supervisor:
Professor Dag Morten Dalen

Hand-in date:
01.07.2021

Campus:
BI Oslo

Examination code and name:
GRA19703 Master Thesis

Program:
Applied Economics

Acknowledgments

To my supervisor, professor Dag Morten Dalen for his help, time, attention, and sustaining with me till the end of this project.

To Mr. Maximilian Schröder for reviewing the thesis and his constructive comments.

To BI Norwegian Business School for awarding me the presidential scholarship and providing me with the chance of receiving high-quality education.

To Henrik and my other friends for making the best two years of my life.

To my family and specially my father, Kamran, for his never stopping, invaluable supports despite the long distance between us.

And, to anyone who helped me leave the country in which I was born.

Sepehr Cyrusian

Thank you to my supervisor, professor Dag Morten for the feedback, help, and advice for the thesis.

Thank you to BI for these five incredible years.

Thank you to Sepehr, my family, and my friends for the support throughout the Master program.

Henrik Sveen

Abstract

Twitter is a source of streaming data. In this thesis, we examine whether and to what extent we can find a relationship between the sentiment of selected Twitter accounts and the S&P 500 index. This thesis uses data from 18 most-followed Twitter accounts and 20 accounts of those who tweet about financial markets in 50 months from January 2017 to March 2021. The sample period encompasses about 1.1 million uncleaned tweets from most-followed accounts and 0.6 million tweets from traders' accounts. We find that the Granger causality between the most-followed accounts sentiment and S&P suggests that while the most-followed accounts sentiment Granger causes the S&P 500, the S&P 500 Granger causes the traders sentiment. Also, we find a significant long-run effect of the net positivity first difference on the S&P 500 index first difference, which is intensified after replacing the most-followed accounts sentiment with the traders' sentiment. Our results show that using an error correction time series model; it is possible to explain 62 to 64 percent of the variation in the first difference of the S&P 500 index by the first difference of the net positivity index and the lagged values of two indices. Finally, we examine the possibility of the predictability power of the sentiment index added to a model consisting of topic probabilities as explanatory variables on the S&P 500 index.

Keywords: sentiment analysis, time series, latent Dirichlet allocation, forecasting, Opinion mining, Machine learning, Lexicon-based

Contents

1. Introduction

Sentiment analysis is a flourishing research field and studying its relationship with economic and financial variables is a burgeoning area of study within economics and computational science.

Unstructured textual data available in social media carry valuable information for financial and economic analysis. The generated textual data in Twitter has a high frequency. By aggregating it on daily intervals, it might be possible to explain the daily fluctuations of the stock market indices. By extracting, processing, and transforming the large volumes of textual data into numbers, we obtain insights about the sentiment embedded in the Twitter accounts. (Algaba et al., 2020) have coined the word *sentometrics*, a portmanteau of sentiment and econometrics, to refer to this emerging field.

In this thesis, we compute the sentiment and transform it from qualitative data into numbers. The numerical sentiment index, obtained after filtering and aggregation, lets us follow sentiment evolution over time. Explaining the S&P 500 index fluctuations using the sentiment index paves the path to predict its future fluctuations. We study sentiment to investigate its explanatory power for S&P 500 index and whether quantifying the tweets' topics and adding them as input to the sentiment index in our econometric model can improve the prediction of S&P 500 returns.

The effect of the COVID-19 pandemic on the news and the shock affecting numerous economic variables in March 2020 and afterward suggests the advantage of having additional indicators to predict the economic and financial variables using the news and massive textual data available on social media.

We investigate the existence of long-term and short-term relationships and find an estimated coefficient of 0.063 and 0.62 adjusted R-squared for the long-run relationship between the first difference of our two variables in daily frequency over a horizon of 50 months.

We combine topic recognition and sentiment analysis to explain and predict the changes in the S&P 500. We use the LDA machine learning technique (Blei et al., 2003) to quantify the topics and the lexicon-based sentiment computation method to estimate the sentiment.

We set up multiple VAR systems and simulated multiple predictions for the stock index. Our finding suggests that the net positivity and eight of the seventeen topics predict the S&P 500. With multiple predictions and as done in (Huang et al., 2005), we will use model averaging to predict and use the random walk model to compare our forecasting results. Our in-sample predictions suggest that the predictions for the S&P 500 outperform the random walk.

In recent years, combining sentiment analysis and machine learning methods has been the subject of some research mainly published in the engineering field. For example, (Ren et al., 2019) combines the sentiment analysis and SVM machine learning method showing an accuracy of 89.93% in forecasting the direction of the SSE 50 index with a rise of 18.6% after introducing the sentiment variables. We, in our thesis, take an approach from an economic point of view to this topic. We modify a lexicon developed for the finance domain and use high-frequency textual data in our econometrics models.

We use Twitter accounts that are among the most-followed ones and estimate the sentiment according to their 1048576 cleaned tweets in the last four years. Because of the suspension of Donald Trump's Twitter account and removing the contents of official accounts affiliated with the 45th president of the United States, we have used the archived accounts regarding his presidency period; as a result of that, our list of most-followed Twitter accounts consists of 18 accounts.

The media influences agents' perception of reality, and agents affect reality (Borovkova et al., 2017). As described by (Algaba et al., 2020), there are various definitions of sentiment used in the field, sentiment can be defined as the disposition of an entity like news media or individual toward another, communicated via a medium.

(M. Baker & Wurgler, 2007) list some potential sentiment proxies for measuring investors' sentiment like investor surveys, trading volume, and IPO first-day returns. However, compared to traditional sources of sentiment extraction like surveys, it is faster and cheaper to obtain sentiment from Twitter at large volumes in real-time and without the risk of facing the Hawthorne effect (Allen & Davis, 2011). As argued by (Kearney & Liu, 2014), there are two main types of sentiment. The first one is the investor sentiment which, as discussed by (M. Baker & Wurgler, 2007) is subjective, and the second type is text-based, which measures the degree of positivity or negativity of texts; hence, it is more objective. In our thesis, our sentiment proxy is the textual data available on Twitter. It is an almost untapped source of information as many similar articles have used newspapers as their data source.

As (Garz, 2014) describes, the evidence shows a bias in the number of reports about unemployment associated with the process of news production and not a result of different interpretations of the economic results. There are some famous articles in the literature that rely on just one data source. For example, (Calomiris & Mamaysky, 2019), (Glasserman & Mamaysky, 2019), (Borovkova et al., 2017), and (Heston & Sinha, 2017) use Thomson Reuters Corp as their data source, (Tetlock, 2007) uses the content of a Wall Street Journal column, and (GARCÍA, 2013) bases his work on two columns of New York Times financial news. We criticize such practice by arguing that it might lead the researchers to estimate an inherently biased index derived by the self-interest of their sentiment source creator. Moreover, relying on one source increases the chance of systematically missing some information, even if the creators of the source do not intend to provide biased data.

We, in our thesis, use two sets of accounts, one with 18 and the other with 20 Twitter accounts which is more diverse than the papers which use just one data source and this can reveal new aspects of the sentiment and be more representative of the actual sentiment, than that derived from traditional textual data sources. Also, due to their high number of followers, they address and potentially affect a broader audience. In addition, Twitter accounts have a higher

publishing frequency than traditional textual data sources like newspapers providing more data at each point in time.

Also, our data cleaning process and domain-specific lexicon are superior to many other works. We have removed more than 120 stop words, and the lexicon that we used has 360 words labeled as positive and 2369 words as negative. (see, e.g., (Picault & Renault, 2017) and (LOUGHRAN & MCDONALD, 2016)

There are two main types of methods for computing the sentiment: lexicon-based approach and machine learning approach (Kolchyna & Tharsis T. P. Souza, 2015). Lexicons are usable at any text level; however, given the application, domain-specific lexicons must be used to obtain the optimal accuracy in estimating the sentiment (Täckström & McDonald, 2011).

In order to select a feature source for the sentiment analysis, we use the lexicon-based approach by using the *Loughran-McDonald* sentiment word list 2018 after modifying it to include COVID-19 related words. (Loughran & Mcdonald, 2011) (Bodnaruk et al., 2015) (LOUGHRAN & MCDONALD, 2016)

2. Literature Review

While conventional indicators, like GDP, can be used to gain insight toward the state of the economy, the existence of obstacles like difficulty in gathering the data and the low frequency in releasing the reports suggests that using a new data source without such limitations may improve economic agents and policymakers' perception of the economy's performance at each point in time hence improving predictions, decisions, and prescriptions. There are numerous articles in the literature supporting this claim, for example, (Borovkova et al., 2017) find that sentiment-based risk indicator carries new information regarding the systematic risk that cannot be derived from traditional risk indicators. Also, (Larsen & Thorsrud, 2019) show that some topics discussed in a newspaper can predict key economic variables in quarterly intervals.

While conventional finance theory posits that sentiment does not affect stock returns and stock prices reflect all the information (Fama, 1965), various works

provide evidence against that hypothesis. As described by (Algaba et al., 2020), since the seminal work of (Keynes, 1936), economists have wondered whether and if so, to what extend sentiment influences economic decision-making at the micro and macro level in economic theory.

For example, (Tetlock, 2007) measures the interaction between the stock market and the content of a Wall Street Journal column and finds that pessimism can predict negative market returns. (Larsen & Thorsrud, 2019) investigate the role of news topics in predicting and explaining economic fluctuations. To do so, they decompose textual data in a Norwegian business newspaper according to the topics using an LDA model (Blei et al., 2003).

To mention more works in this regard, we refer to the paper *News versus Sentiment: Predicting Stock Returns from News Stories* (Heston & Sinha, 2016) in which the authors use 0.9 million news stories to predict the stock returns, finding the daily news can predict the returns for 2-1 days.

(Calomiris & Mamaysky, 2019) develop an atheoretical approach to study news through word flow measures like sentiment, frequency, entropy, and the topical context. They capture dynamic changes in coefficients to improve out-of-sample forecasts finding that news forecasts the returns one year earlier, implying that word flow captures "collective unconscious" aspects of the news, which might affect the economy. (Shiller, 2017) and (BAKER & WURGLER, 2006) run a regression with the dependent variable being the monthly return in a long-short portfolio and the independent being sentiment lagged for one period. They find that the cross-section of future stock depends on proxies of sentiment in earlier periods.

The article *Twitter as a tool for forecasting stock market movements: A short-window event study* by Nisar and Yeung has collected more than 60000 tweets and performed "a collection of correlation and regression analyses to compare daily mood with" price changes of the FTSE 100 at the market level. However, their study did not acquire statistically significant results regarding the

relationship between Twitter chatter and stock market movements. (Nisar & Yeung, n.d.)

The article “Forecasting stock market movement direction with support vector machine” by (Huang et al., 2005) discusses the complexity and difficulty of predicting the stock market. The paper argues that the stock prices are not random but rather behave dynamically and non-linear manner. Further, the article suggests model averaging techniques to improve predictive performance. Moreover, they use a random walk model as a benchmark to evaluate the forecasting ability of their prediction.

Economic sentiment can be seen as an index that reflects the information about events that have already materialized or a source containing fundamental information. Hence, it can act as a self-fulfilling prophecy. (Petropoulos Petalas et al., 2017)

In a similar vein as for (Petropoulos Petalas et al., 2017) and by noticing the relationship between sentiment and expectations, (Beber & Brandt, 2010) mentions that investors update their expectations of economic variables as they receive new information, so they study the effect of macroeconomic announcements on the bond returns. They find that the information content of the announcements has the greatest effect on the bond returns when it contains bad news in the expansionary periods.

The existence of long-term effects of news show inconsistency with the efficient market hypothesis (Fama, 1965). (Kräussl & Mirgorodskaya, 2017) hypothesize that the media sentiment translates into investor sentiment. They investigate the potential long-term effects of media sentiment on the performance of financial markets. They study two VAR models to analyze whether changes in media pessimism affect future changes in the market returns level.

They find that the log change of the (BAKER & WURGLER, 2006) investor sentiment index exhibits positive and strongly statistically significant contemporaneous relation with S&P 500 index at monthly frequency. The

estimated coefficient is 0.008 for the log change of the (BAKER & WURGLER, 2006) investor sentiment index, and the adjusted R-squared is 0.606. They conclude that despite that previous literature suggests a negative association between media pessimism and contemporaneous market returns (Antweiler & Frank, 2004);(GARCÍA, 2013);(Goetzmann et al., 2016);(Tetlock, 2007), finding that, over their three year study horizon which is longer than previous studies, the media pessimism is associated with the market performance in the long run.

Another article investigating the long-term relationship is (Kleinnijenhuis et al., 2013), in which the researchers measure market sentiment based on six newspapers and, in doing so, narrow down the words into two groups of emotions, namely “fear” and “hope,” and calculate fear-related words minus the number of references to hope on a monthly basis. They present a model in which the change in stock market value at the close of the Amsterdam exchange market (AEX) day depends on the change of the amount of news associating a bank to the financial crisis on the same day and the day before; hence, their study suggests that financial news do affect markets.

Depending on the goal of the work, specific types of sentiment can be estimated as a proxy for another hard-to-measure variable such as company reputation (Saleiro et al., 2017) or uncertainty, (S. R. Baker et al., 2016) develop an index of economic policy uncertainty based on newspaper coverage frequency for the United States by relying on 10 newspapers and confirm previous works on negative economic effects of uncertainty shocks.

(Borovkova et al., 2017) use VAR to study the behavior of a sentiment-based risk indicator with respect to macroeconomic indicators. In order to do so, they investigate the impulse response functions and granger causality relations finding that sentiment-based risk indicator carries new information about information risk which cannot be derived from traditional risk indicators.

3. Pre-processing

3.1 Legal and ethical terms

Twitter gives its users some control over their data, where users can set their accounts to private or public. The accounts considered in this thesis are public.

The accounts that we have considered in this study have chosen to publish the tweets and make them public. When they publish a tweet, it is accessible and read by people worldwide.

3.2 Data collection

We have used two datasets in this thesis, one for calculating the net positivity score and the other for measuring the S&P 500 closing price.

At the start of the data collection phase in the months before the U.S.A election, it was impossible to scrape the data using the TWINT package as Twitter had blocked it. As a result, the oldest tweets possible to download were those published in July 2020, however after the end of the election period in the U.S.A the Twitter lifted the restriction on the package, and we used the TWINT (OSINT team, n.d.) scraping tool to collect the tweets, and Twitter accounts information. TWINT is written in Python by the OSINT team, and its main advantage is that it circumvents twitter's API limitation, enabling us to extract tweets from the 1st of Jan of 2017 till the 16th of Mar 2021.

The parsing algorithm that we use returns various sorts of metadata, including, but not limited to: the account ID, date of tweet creation, time zone, tweet, language, username, and handle of each account.

Another metadata that we generate using LDA, an unsupervised machine learning technique in NLP, is topics of the tweets along time (El-Amir & Hamdy, 2019).

Our daily stock price dataset is not seasonally adjusted, closing stock prices of the Standard and Poor's 500 companies from the 1st of Jan of 2017 till the 16th of

Mar of 2021 (*S&P Dow Jones Indices LLC, S&P 500 [SP500], Retrieved from FRED, Federal Reserve Bank of St. Louis*, 2021).

3.3 List of accounts

This thesis focuses on the 18 most followed accounts, including their archived accounts on Twitter. In the end, we consider 20 accounts of those with a high number of followers tweeting with a focus on the financial markets. We have chosen these lists, which comprise the accounts that the influence of the owners and the popularity of their account might suggest a relationship between their content and the S&P 500 index and possibly other financial and economic variables.

To choose the 18-account list we exclude accounts of singers, actors, and entertainment industry public figures. Most members of the list are politicians and news agencies. The list of accounts for the most followed accounts is shown in Table 1A (Number of followers is as of 12th of Jun 2021).

After cleaning the textual data, we are left with 1048576 tweets from our list of 18 Twitter accounts (including the archive accounts). After calculating their sentiment, we aggregated the net positivity in daily intervals.

As the S&P 500 index (*S&P Dow Jones Indices LLC, S&P 500 [SP500], Retrieved from FRED, Federal Reserve Bank of St. Louis*, 2021) data is available only for working weekdays, we exclude the corresponding values from the net positivity data.

The data formats we have used in this thesis are xlsx and CSV. Due to consistency, in the parsing step, the CSV file format has been converted from a JSON file format containing tweets and their corresponding Twitter account information. The programming language that we have used is Python. We utilized the numerous packages, modules, and libraries represented in Table 2A.

As the data obtained from Twitter is unstructured with much noise, it is vital to clean it, improve the analysis, and decrease the dimensionality of data. The next phase can be summarized in the following steps:

3.4 Tokenization

In the tokenization step, we split tweets into individual words.

3.5 Redundant word removal

Removing special characters (for example, hashtags and foreign language letters), URLs, the account handles, emojis, and excessive repetitive characters.

Removing stop words: top words are words (for example, “the,” “to,” and “a”) that do not carry much information and are not informative.

Also, some of the Twitter accounts considered in our thesis included highly repetitive words and expressions that were more similar to click-baits than organic news, so we excluded those tweets by removing the tweets which included any of the expressions mentioned in the appendix.

3.6 Converting to lowercase

In order to solve potential case-sensitivity problems, we convert all the letters to lowercase form.

3.7 Lemmatization

The last step is lemmatization. In the process, the part of speech of each word is recognized, and the roots substitute the corresponding words. For example, as a result of lemmatization, “worse” and “worst” will both be converted to “bad.” For the most followed accounts, the word cloud of the words with greater than three characters has been shown in Figure 1A, and the frequency distribution of the top 20 words is visible in Figure 2A.

4. Feature engineering

The goal of feature selection is to remove irrelevant features and be left with what describes the characteristics of the data in order to reduce the dimensionality of data to improve machine learning performance (Liu, 2010).

One feature engineering method is to create a co-occurrence matrix nut because of the vast vocabulary that we face; it would lead to computation problems.

To find the importance of a word, on the one hand, the high occurrence of a word can be a sign of its importance, but on the other hand, many words do not carry much meaning and appear in most of the documents.

4.1 TF-IDF

To overcome the problem mentioned earlier and give a score to each word in the documents, we convert the text to feature using the TF-IDF. The TF-IDF is a statistical measure intended to reflect how important a word is to a document in a corpus (Swamynathan, 2019); hence normalizing words appeared frequently in all the documents (Leskovec et al., 2014).

TF-IDF stands for term frequency-inverse document frequency, and based on each word's relative importance; it assigns a normalized score to the words that appeared in the documents (Rickard Nyman et al., n.d.).

TF-IDF is calculated as the product of the term frequency and the inverse document frequency. The term frequency shows the importance of a term in each document. Term frequency is usually defined as the ratio of the number of times that term t appears in document d to the length of document d. Thus, the TF captures the importance of the word irrespective of the total number of documents.

Inverse document frequency shows the importance of a term relative to the entire corpus. The inverse document frequency increases in value the more uncommon a term is across the corpus, as it measures each word's rareness. If a word is prevalent in all documents, then that word does not have much importance and is of no use in information retrieval. IDF nullifies this problem.

Inverse document frequency:

$$\text{Idf}\,(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \qquad (1)$$

Where:

- N is the number of documents in the corpus
- $|\{d \in D: t \in d\}|$ is the number of documents where the term t appears (Swamynathan, 2019)

The terms with the highest TF.IDF scores are often the terms that best characterize the document's topic (Leskovec et al., 2014).

We calculate the term frequency and inverse document frequency by splitting each document and finding the unique words in them. However, we can use the class provided by the sklearn machine learning library to get the results faster as sklearn has already implemented various optimization methods.

As we look into the 15 words with the highest rank for each month's tweets, we get the following tuple list, where the first element of each tuple is the word and the second element is its TF-IDF score in the respective month. So, for example, the result for November 2020, while the corpus consists of all the tweets from the most-followed accounts from September 2020 till the month mentioned above, is shown in the appendix.

4.2 Latent Dirichlet allocation

In order to summarize and compress the information content of our dataset by transforming it onto a new feature subspace of lower dimensionality, we use the generative statistical model latent Dirichlet allocation (Blei et al., 2003) with the online variational Bayes algorithm provided by Scikit-learn python package (Pedregosa et al., 2011).

As described in (Nimark & Pitschner, 2019), LDA models are one of the most common tools in NLP, letting us recognize and quantify the topics. As its name suggests, it describes a latent form that could have generate the tweets according to probabilistic rules. We choose the number of topics which will then be used by the model to endogenously discover the topics as the outputs of the estimated model. The advantage of the LDA is that it measures both the absolute and

relative importance of each topic over time; however, human input has to associate topics with specific events.

Latent Dirichlet allocation assumes a fixed number of topics containing a set of words. It maps documents to topics so that the topics capture each documents' words. LDA in natural language processing (NLP) is an unsupervised technique, so there is no need for labeled samples (El-Amir & Hamdy, 2019).

By trial and error, we find the number of topics covered in the 18-account list by running the algorithm for over 100 iterations and then subjectively evaluating the results. Given our infrastructure, this process is highly time-consuming. Having more than 17 topics leads to a very similar set of words categorized as different topics while having lower than 17 topics leads to losing information about them. We use 3000 features, and each of the 17 topics is represented by three expressions, each of which has three terms. The model topics have been endogenously estimated. The LDA has recognized the 17 topics along the research horizon, and the topic probability of them is calculated. This result helps us infer the content of the 1048576 tweets along the research horizon. The results are represented in Figure 1.

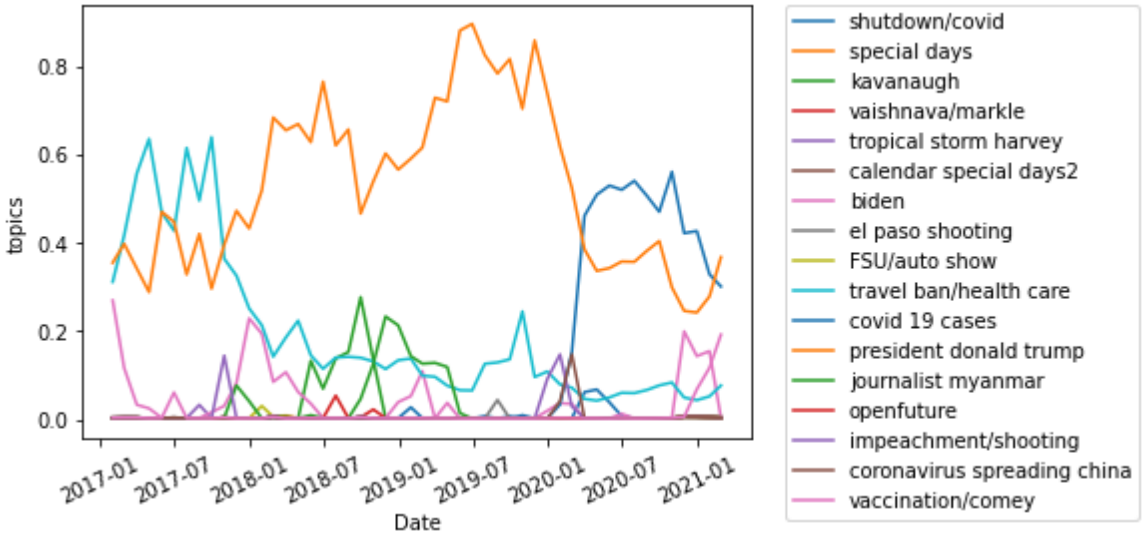

Figure 1 Recognized topics

Based on the words included in each topic, we subjectively give names to the topics as follows: partial gov shutdown/covid spread, calendar special days, court nominee Kavanaugh, Vaishnava Jana/Meghan Markle, tropical storm Harvey, calendar special days2, biden, el Paso shooting, first state union/Detroit auto show, travel ban/health care, covid 19 cases, president Donald Trump, journalist imprisoned Myanmar, openfuture video contest, trump impeachment/la vega shooting, coronavirus spreading china, covid 19 vaccination/James Comey testimony.

Several topics are easily and intuitively identifiable, like covid cases, Trump, and vaccination. However, some others are not associated with a single event or person like travel ban/health care topic, which disentangling it is not possible based on the model estimates. The complexity of the existence of those difficult-to-interpret topics is common in LDA models (Chang et al., 2009). One hundred sixty-eight out of the total 3000 features used in the LDA are presented in the appendix.

5. Feature source for the sentiment analysis

In order to select a feature source for the sentiment analysis, we use the lexicon-based approach by using the Covid-modified *Loughran-McDonald* sentiment word list 2018 (Loughran & Mcdonald, 2011) (Bodnaruk et al., 2015) (LOUGHRAN & MCDONALD, 2016)

There are various general and domain-specific lexicons; for example, both the Henry lexicon (Henry, 2008) and the Loughran-McDonald lexicon (Loughran & Mcdonald, 2011) are created to handle texts in the finance domain. However, as summarized by (LOUGHRAN & MCDONALD, 2016) the Loughran-McDonald lexicon has two main advantages over other word lists frequently used in the accounting and finance literature. First, compared to the Henry [2008] list (Henry, 2008), Loughran-McDonald lexicon is comprehensive. Second, it has been created with financial communications in mind. Recently, it has become one of the most widely applied lexicons used in the literature to compute the tone of business communications (Kearney & Liu, 2014).

As the coronavirus has had an undeniable effect on the content of tweets as our data source, we modified the *Loughran-McDonald* Master Dictionary to include *vaccine, Pfizer, Moderna, Johnson, inoculation, vaccination,* and *pandemic* as positive and *virus, lockdown, quarantine, infection, infectious, spread, outbreak, strain,* and *infected* as negative words. In total, we have 360 words labeled as positive and 2369 words as negative in the lexicon. As the words have been selected by care, this approach is highly effective. However, it is not entirely automated and is highly time-consuming (Birjali et al., 2021).

6. Sentiment classification

To tackle the c.hallenge of quantifying textual data into a numerical sentiment index, we use the lexicon-based sentiment computation approach. As summarized by (Algaba et al., 2020), all sentiment measures are proxies for the actual sentiment; hence they need to be estimated. Given the fact that sentiment is a latent variable and is not readily visible, we have to measure it from tweets texts as a qualitative data source and transform that data into numbers to analyze whether it can explain fluctuations in the stock market and be a timely driver of S&P 500 index in our forecasting model.

There are two main types of methods for computing the sentiment: lexicon-based approach and machine learning approach (Kolchyna & Tharsis T. P. Souza, 2015). In order to quantify the already observed sentiment, we define the net positivity score for each tweet to measure the sentiment as follows:

$$Tweet's\ Sentiment = \#\ Positive\ tokens - \#\ Negative\ tokens \qquad (2)$$

This score is calculated as the difference between the number of words categorized as positive and the number of words categorized as negative in the modified *Loughran-McDonald* sentiment word list 2018. As the behavior and possible trend of the sentiment compared to S&P 500 index is about to be analyzed, no normalization factor was implemented in calculating the net positivity.

We apply cross-sectional aggregation at a daily frequency on the net positivity of tweets. Obtaining daily time series makes it possible to work with the net positivity time series and the S&P 500 index closing daily price. We have obtained the overall sentiment of the tweets in each day, as shown in Figure 2, in which the horizontal axis shows the number of working days since the first working day of Jan 2017.

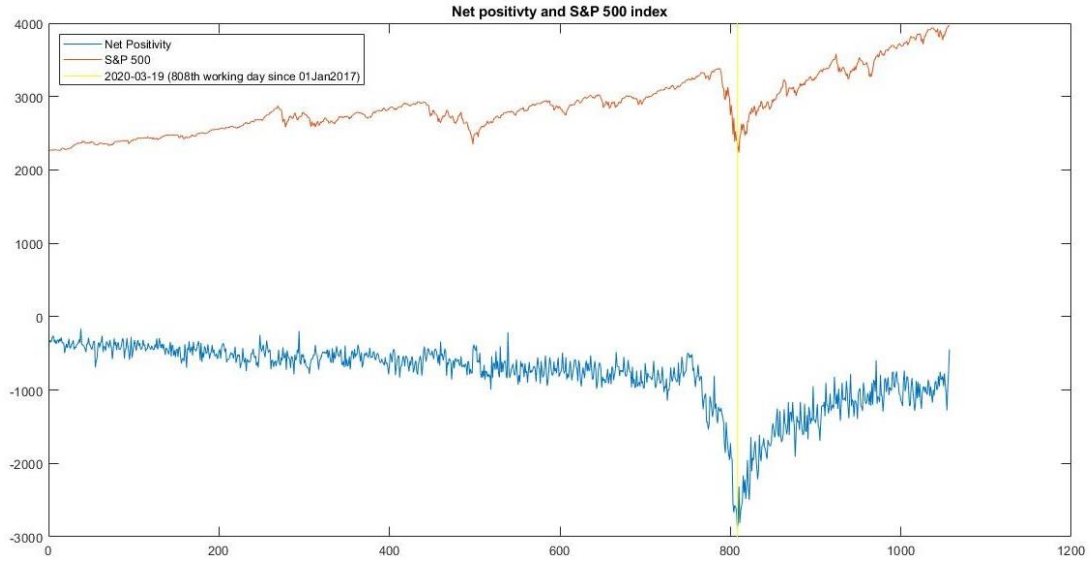

Figure 2 Net positivity and S&P 500 index

Each year consists of about 250 working days as the market is idle on the weekends and U.S. holidays. We observe that the net positivity plummets in March 2020 and reaches its lowest value on the 19th of Mar 2020 or 808th working day after the 1st working day of Jan 2017, while S&P500 reaches its minimum two working days after twitter's most pessimistic day, on the 23rd of Mar 2020. Therefore, we want to see whether and to what extent it is possible to explain and predict the change in the S&P 500 index based on the changes in the sentiment and its lags.

The advantage of the lexicon-based approach is the fact that it does not require any data training but the disadvantage of it is the domain dependency (Birjali et al., 2021)

By using the modified *Loughran-McDonald* sentiment word list 2018 and considering the Twitter accounts, which mainly use official language, we have tried to mitigate this problem.

7. Methodology

7.1 Estimating the long-term and short-term relationship between sentiment index and S&P 500 index

Long-term effects of news show inconsistency with the efficient market hypothesis (Fama, 1965). We want to investigate whether our estimated sentiment suggests the existence of such inconsistency

In order to check for stationarity, we run an augmented Dickey-Fuller test (Hill et al., 2018). The unit root tests determine whether variables are I(0) or I(1).

In a Kripfganz et al. model without independent variables, the bounds test collapses to the augmented Dickey-Fuller unit root test (Kripfganz & Schneider, 2018). So to find the number of lags of first differences, we have used the lags based on information criteria obtained from the Mata-based algorithm (Kripfganz & Schneider, 2018) (Kripfganz & Schneider, 2020)

We run an augmented Dickey-Fuller test with three lags and on the sentiment values, while the alternative hypothesis is that the sentiment index is stationary around a non-zero value. We have chosen such the test because the net positivity in figure 3A suggests that the series is not oscillating around a zero mean.
The results are represented in the “Results” section.

7.2 Forecasting

Despite explicitly mentioned, in all the forecasting models in this thesis, we use data in 49 months from January 2017 to February 2021. We will investigate whether net positivity and the topics of the tweets predict the S&P 500. We will use the Granger causality test and test the impulse response function on multiple VAR systems containing the relevant variables.

We will predict the S&P 500 using the estimation from the VAR models to investigate how different systems predict the stock index. We will use the predicted values to make a single prediction using model averaging.

7.2.1 VAR framework

To make the prediction, we have chosen to employ the VAR approach for our dataset. As we work with time-series data, we choose the VAR model that allows for multiple endogenous variables and makes examining dynamic effects possible without imposing strict restrictions. Furthermore, the model expands on the autoregression model as the variable depends on its own and other endogenous variables' lagged values.

We believe that the stock index, net positivity, and the topics might affect each other. When the stock market goes well, it will affect the people's sentiment, and when people talk more positively, it could be a sign that people are optimistic, which could lead to people buying more in the stock market. The topics people talk about will properly affect the sentiment, and when people are unhappy, they may focus on tragic topics. Also, events can change people's expectations and affect the stock market. If the stock index falls, it could lead to speculation about what caused it, like the coronavirus. Therefore, it suggests for a simultaneous equation and all the variables to be endogenous. We do not believe that all the topics affect each other. For example, we do not see the context that the Covid 19 affects the El Paso shooting.

Instead of dropping variables containing valuable information, we will set up multiple VAR systems where we believe that all the variables are endogenous. The variables that do not significantly predict the S&P500 will be dropped from the forecasting model.

In the first setup, the endogenous variables are the S&P 500 and net positivity. It allows examining the dynamic effect net positivity has on S&P 500 and how accurate the net positivity predicts the stock index when used as the only other variable.

In the second setup, eight VAR-systems with S&P 500 and a topic are set as the endogenous variables where the aim is to examine the topic's effects on the S&P 500.

In the third setup, we combine the first and second set up by setting up eight VAR-systems with the S&P 500, the net positivity, and one topic as the endogenous variables. As in the two first setups, we are interested in the dynamic effects between the variables and how the topics and net positivity predict the stock index. We will also observe how the dynamic effects change and if the prediction becomes more accurate when adding a variable to the second setup compared with the first. In total, there will be an estimated 17 VAR models, where each model with its corresponding endogenous variables is presented in tables 3A, 4A, and 5A.

Below are the mathematical representations of the VAR model for each setup in matrix form. Where t is time, S is S&P 500, P the net positivity, and T the topic, α is the intercept, β is the coefficients of the lags of the endogenous variables, and ε is the error terms.

$$\begin{bmatrix} S_t \\ P_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \sum_{i=1}^{k} \begin{bmatrix} \beta_{11}^i & \beta_{12}^i \\ \beta_{21}^i & \beta_{22}^i \end{bmatrix} \begin{bmatrix} S_{t-i} \\ P_{t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{S_t} \\ \varepsilon_{P_t} \end{bmatrix} \qquad (3)$$

$$\begin{bmatrix} S_t \\ T_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \sum_{i=1}^{k} \begin{bmatrix} \beta_{11}^i & \beta_{12}^i \\ \beta_{21}^i & \beta_{22}^i \end{bmatrix} \begin{bmatrix} S_{t-i} \\ T_{t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{S_t} \\ \varepsilon_{T_t} \end{bmatrix} \qquad (4)$$

$$\begin{bmatrix} S_t \\ P_t \\ T_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \sum_{i=1}^{k} \begin{bmatrix} \beta_{11}^i & \beta_{12}^i & \beta_{13}^i \\ \beta_{21}^i & \beta_{22}^i & \beta_{23}^i \\ \beta_{31}^i & \beta_{32}^i & \beta_{33}^i \end{bmatrix} \begin{bmatrix} S_{t-i} \\ P_{t-i} \\ T_{t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{S_t} \\ \varepsilon_{P_t} \\ \varepsilon_{T_t} \end{bmatrix} \qquad (5)$$

Forecasting

We are going to predict future values of the S&P 500 using the estimated VAR models. All the models will be estimated using the data from 01.01.2017 – 01.02.2021. To estimate the prediction, we use the methods from (Herwartz &

Kholodilin, 2011). We are going to perform dynamic forecasting, which produces predictions for periods ahead. We are going to achieve short-term in-sample forecasting from the period 01.01.2021 to 01.02.2021. It allows gathering information like the RMSE, which we can use to compare the forecast's performance with other predictions.

With multiple predictions from the VAR models, we use model averaging to combine numerous predictions into a single prediction. It allows us to use a large number of variables and obtain better-fitted models. According to (Montero-Manso et al., 2020) the combination of forecasts is often superior to their individual counterparts.

This thesis uses the Bates and Granger (1969) approach as the model average technique. As stated by (Eklund & Karlsson, 2007), the Bates and Granger forecast combination is a highly successful forecasting strategy. Their approach is that the weighting for a prediction relay on the root means square deviation (RMSE) from each model. The lower the RMSE is for a model, the more weighted their prediction is in the final forecast. We will estimate the RMSE for the weighting using the predicted value from the models and the actual S&P 500 data. The weighting estimator follows the formula below, where the W is the estimated weight, and $\hat{\sigma}$ is the RMSE.

$$W_m = \frac{\hat{\sigma}_m^{-2}}{\sum_{i=1}^{M} \hat{\sigma}_i^{-2}} \qquad (6)$$

After all the predictions obtained from the VAR models and their corresponding estimated weighting are obtained, we will estimate the final prediction.

W is the estimated weight for each model, y is the predicted output at time t, and $y_{avg,t}$ is the final prediction at time t. The formula below explains that summing over the multiplication of models' predictions with their corresponding weights will equal the Bates and Granger predictions.

$$y_{avg,t} = W_1 y_{1,t} + W_2 y_{2,t} + \ldots + W_M y_{M,t} \qquad (7)$$

To evaluate the performance of the predicted value, we will compare it with a random walk model and use that as a benchmark. According to (Malliaris, 1994), the stock price time series has a non-random underlying structure in the market. For the predicted values of the S&P 500 to be valid, they need to outperform a random process. To estimate the random walk, we use the (Nau, 2014) as our guideline where the model assumes that the model takes a random step from its previous value. We decided to use the random walk model with drift since the S&P 500 increased over our sample period. As shown in the equation below:

$$\hat{Y}_{n+k} = Y_n + k\hat{d} \qquad (8)$$

To forecast the most accurate prediction, we will estimate the drift for the period our data is collected, using the following equation

$$\hat{d} = \frac{Y_n - Y_1}{n-1} \qquad (9)$$

To estimate the error terms for the random walk, we use the following equation,

$$SE_{fcst(1)} = STD(Y_{DIFF1}) \qquad (10)$$

Where takes the standard deviation of the first difference of the S&P 500. We will compare the predictions using values like the MAE and RMSE.

7.2.2 Stationarity

To make a prediction and analyze the dynamic effects of the VAR models, all the variables must be stationary. Visual inspection of the series does not suggest that either S&P 500 or the net positivity have a constant mean or standard deviation. We will test for unit roots to check if our variables are stationarity using the Augmented Dickey-Fuller-test (ADF-test) at a 5%significant level. To make the time series stationary, we will perform the first difference in our variables.

$$d_t = x_t - x_{t-1} \qquad (11)$$

The formula above shows that the first difference is to differentiate the current period's value from the previous one. Figures 4A and 5A represent S&P 500 and net positivity of the most followed accounts sentiment after applying the first difference on the original series, and visually these series look stationary. We use (Schwarz, 1978) Information Criteria (SBIC) to determine the lag length. Summarized in table 6A the ADF-test tells us that all the variables are stationary, with P-values close to zero.

7.2.3 Lag length

To determine the optimal lag length for the models, we will simulate multiple lag selection tests to determine the optimal lag length for each of the VAR models. The lag selection test is (Akaike, 1969) Finale prediction error (FPE), (Akaike, 1974) Information Criterion (AIC), (Hannan & Quinn, 1979) Information Criterion (HQIC), and (Schwarz, 1978) Information Criteria (SBIC).

Table 7A summarizes the results for the lag selection for each model. As we can see in the table, the test shows different results. In general, we do not see any reason to include that large number of lags in our model. The problem with using many lags is that it could lead to over-parametrization. Setting few variables could lead to little information for the regression and result in a poor fit for the autoregression. In general, the FPE and AIC suggest more lagged variables than HQIC and SBIC.

We decided with the lag value of 14 for the first. In general, for the second setup, the tests did suggest the option between 33 or 1 lagged value. Since the lagged value of 1 led to a high RMSE and low regular and adjusted R-squared, we decided on the lag value of 33. The only exception was the value "Coronavirus spreading in China" where the same problem occurs for the rest of the model in the second setup that lag value of 1 is too small, and we decided for the HQIC value of 46. In the third setup, we decided to use the HQIC, which suggested 14 for every model. As in the second setup, the only exception was the model with the topic "Coronavirus spreading in China" where the HQIC suggested 46 lags. We decide to use the SBIC test and the lag value of 8.

7.2.4 Autocorrelation

Autocorrelation measures the correlation between lagged values, where it measures variables current against their past values. If there is any correlation in the residuals, then there is some pattern left to be explained by the model. To test for autocorrelation, we use the (Durbin & Watson, 1950) Statistic, a widely used tool to test serial correlation. The Durbin-Watson Statistic will produce a value between zero and four. The value is close to two means that we cannot reject the null hypothesis of no autocorrelation. On the other hand, a value close to zero and four suggests a positive and negative autocorrelation for the variable. Thus, the test measures the relationship between the error terms.

$$DW = \frac{\sum_{t=2}^{T}((e_t - e_{t-1})^2)}{\sum_{t=1}^{T} e_t^2} \qquad (12)$$

The S&P 500 variable from every model will run the formula above and tested for autocorrelation with regards to the hypothesis:

Durbin-Watson Statistic hypothesis:
Ho: No evidence of autocorrelation
Ha: Evidence for positive/negative autocorrelation

The results, which are the furthest away from the value of two, was the third setup with the topic “coronavirus spreading in China” with 1.9625. The rest of the models showed results from 1.99 and 2. Thus, all the results are remarkably close, and we cannot reject the hypothesis (H0: No evidence of autocorrelation) for all the models. The results are summarized in Table 8A.

7.2.5 Causality

To test if the variables predict the S&P 500 and analyze causality between the variables, we perform two causality tests. We use the methods presented in the paper (Lütkepohl, 2013). First, the Granger Causality test and impulse response functions (IRF) will be tested at a significant level of 5% and explained in the results.

7.2.6 Var Regression Output

The equation Below represents the output for the first setup of the VAR model.

$$S_t = \alpha_1 + \beta_{11,1}S_{t-1} + \beta_{12,1}P_{t-1} \dots \beta_{11,14}S_{t-14} + \beta_{12,14}P_{t-14} + \varepsilon_{S,t} \quad (13)$$

$$P_t = \alpha_2 + \beta_{21,1}S_{t-1} + \beta_{22,1}P_{t-1} \dots \beta_{21,14}S_{t-14} + \beta_{22,14}P_{t-14} + \varepsilon_{P,t} \quad (14)$$

The equation below represents the output for all the models in the second setup (except for the model with the topic "Coronavirus spreading in China")

$$S_t = \alpha_1 + \beta_{11,1}S_{t-1} + \beta_{12,1}T_{t-1} \dots \beta_{11,14}S_{t-33} + \beta_{12,14}T_{t-33} + \varepsilon_{S,t} \quad (15)$$

$$T_t = \alpha_2 + \beta_{21,1}S_{t-1} + \beta_{22,1}T_{t-1} \dots \beta_{21,14}S_{t-33} + \beta_{22,14}T_{t-33} + \varepsilon_{T,t} \quad (16)$$

The equation below represents the output for all the models in the third setup (except the model with the topic "Coronavirus spreading in China")

$$S_t = \alpha_1 + \beta_{11,1}S_{t-1} + \beta_{12,1}P_{t-1} + \beta_{13,1}T_{t-1} \dots \beta_{11,14}S_{t-14} + \beta_{12,14}P_{t-14} + \beta_{13,14}T_{t-14} + \varepsilon_{S,t} \quad (17)$$

$$P_t = \alpha_2 + \beta_{21,1}S_{t-1} + \beta_{22,1}P_{t-1} + \beta_{23,1}T_{t-1} \dots \beta_{21,14}S_{t-14} + \beta_{22,14}P_{t-14} + \beta_{23,14}T_{t-14} + \varepsilon_{P,t} \quad (18)$$

$$T_t = \alpha_3 + \beta_{31,1}S_{t-1} + \beta_{32,1}P_{t-1} + \beta_{33,1}T_{t-1} \dots \beta_{31,14}S_{t-14} + \beta_{32,14}P_{t-14} + \beta_{33,14}T_{t-14} + \varepsilon_{T,t} \quad (19)$$

8. Results

8.1 Forecasting results

Despite explicitly mentioned, in all the forecasting model in this thesis we use data in 49 months from January 2017 to February 2021. The table 9A represents the regression for SP500 from each model. As we can see, the RMSE values are close to each other across the models. The lowest value of RMSE is the model with the topic "Coronavirus spreading China" from the second setup. The third setup with the lowest RMSE values is the models with the topics "Covid 19" and "Partial gov shutdown/Covid 19 spread", indicating the models that fit the best.

Most of the R-squared values are around 13-14%, and the adjusted R-squared values are around 9-10%; however, there are some outliers among the models. The second and third setup models with the topic "Partial gov shutdown/Covid spread", and the third setup with the topic "Covid 19" see higher regular and adjusted R-squared values than the rest.

The highest and lowest values are for the model with the topic "Coronavirus spreading China" in the second and third setup with an R-squared of 37.83% and 7.93% and adjusted R-squared values of 33.55% and 6.43%, respectively. It is not surprising, consider these is the models with the most and most minor variables. A low value for regular and adjusted R-squared in all the models would indicate a poor fit for the models.

8.1.1 Granger Causality

To test the causality between the variables, we perform causality tests on the models; the first test we perform is the Granger Causality test. The Granger causality tests for correlation between other variables and the variable's current value and past values. It does not give any insight into the shock effect and its causes or other variables over time. However, even if the variables do not have Granger causality between them, there can still be a causality relationship. We will test for the Granger Causality relationship between the variables using the setup of the VAR models with their selected lag values. Tables 1 and 10A to 25A in the summarize all the Granger causality tests for all the models.

8.1.1.1 Net Positivity

We will first look at the variable net positivity, where it occurs in the first setup and all the models in the third setup. Table 1 summarizes the results for the first setup VAR model, where we observe that the S&P 500 and net positivity mutual Granger causal each other. We also test the Granger causality relationship between the variables in all the models in the third setup. The results are similar to the first setup, with a p-value close to zero for all the models. Meaning the variable S&P 500 and net positivity mutual Granger cause each other in all the models.

Table 1: The p-values of the Granger causality tests for the first setup

Dependent / Independent	S&P 500	Net Positivity
Net Positivity	0.00*	-
S&P 500	-	0.00*

8.1.1.2 Topics

Tables 10A-25A summarize the models where we observe a Granger causality between the S&P 500 and the topics, with their selected lagged value. We observe that the models with the topics "partial gov shutdown/Covid spread" and "Covid 19" and the third setup with "Coronavirus spreading China" mutually Granger causes each other. The second setup with "Coronavirus spreading China" is the only variable where the Granger causality test for a significant causality to the S&P500. The remaining variables did not show any evidence that they have any Granger causality relationship with the SP500. However, this does not necessarily mean the variables in the remaining models do not have a causal relationship with each other. We also investigate impulse responses.

8.1.2 Impulse responses

We will also look at the impulse responses function (IRF) to observe whether there is a relationship at individual lag. The IRF is a reaction of an impulse (like a shock) to a dynamic system and its variable response. Every variable in every model will be hit by a shock of one standard deviation and observe its effects on the other variables over a given period. We will perform orthogonalized impulse responses.

The response of the variables own shock has on itself is often intense and will not be reported. We are neither interested in the effect of a shock to the S&P 500 has on the other Variables. We examine the effect on the S&P 500 due to a shock on the net positivity and topics. We have also chosen only to include 15 periods since very little happens after that. The IRFs are shown in tables 26A-28A in the appendix.

8.1.2.1 Net Positivity

The variable net positivity is in the model from the first setup and in the eight models in the third setup. We hit a shock to the net positivity and investigate its effect on S&P 500 for every system that contains the variable. Table 26A summarizes the impulse responses for the variable net positivity. We observe a significant relationship for lag 2 for every system. Lags 4 and 13 also show a significant relationship for most of the system, and lag 3 and 12 occur in some systems.

8.1.2.2 Topics

Tables 27A and 28A summarize the response of S&P 500 (as a percentage change in S&P 500) when the topics are hit by a shock. Table 27A includes the VAR systems with net positivity, and table 28A includes the models without net positivity of 18 most-followed accounts. By comparing the second setups to the third setups, we observe that topics fluctuate similarly. There are periods in which we observe significant S&P 500 responses (as a percentage change in S&P 500) in response to a shock to each topic. The tables show that the corona-related topics have multiple periods with a significant response shown as the percentage change of S&P 500, and the rest of the topics have just one significant period.

8.2 Predicted S&P 500

The figure 3 represents the predicted S&P 500 values from all the VAR models. We see that all the predictions have similar movements. The two models with the variable “Covid 19” predict the highest value for the stock. The model with the topic “Calendar special days2” in the second setup does predict the lowest value for the S&P 500.

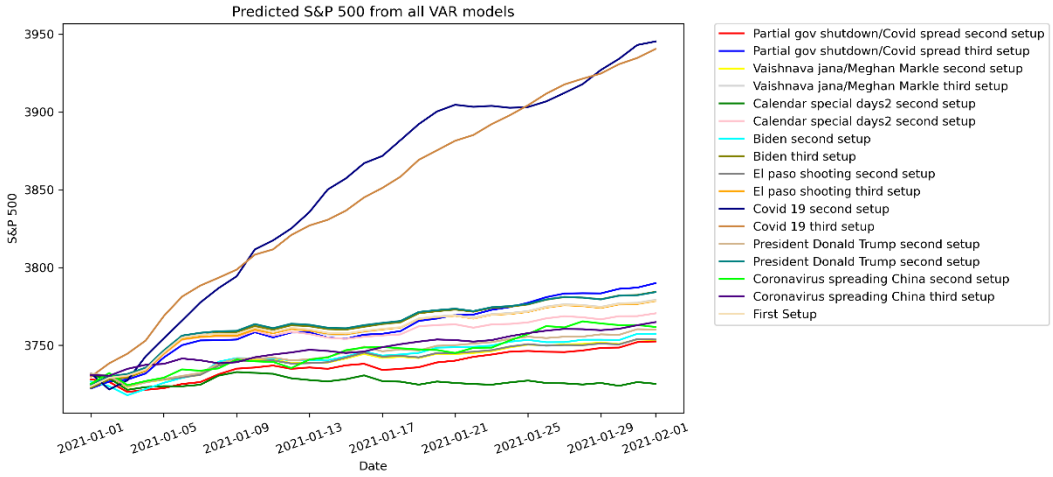

Figure 3: Predicted S&P 500 from all the VAR model's

8.2.1 Weighted average estimates

We see in table 29A that the prediction with the lowest accuracy according to RMSE is for the models with the topic “Covid 19” for both setup and “Calendar special day2” in the second setup. Where the best prediction, according to RMSE, is from the model with the topic “President Donald Trump” in the second setup. In general, the first and third setups show a lower RMSE than the second setups, indicating that models containing the net positivity perform better than the models without the variable. The weighted scheme is estimated using the RMSE for the output for the S&P 500 forecasting for each model, and figure 4 shows each model’s represented weighing in percentage for the predictions.

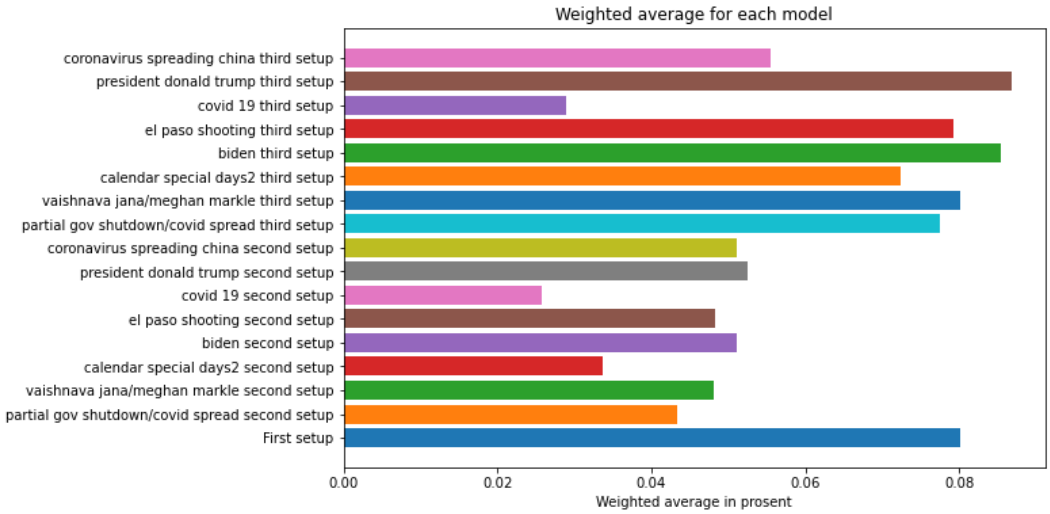

Figure 4: Weighted average for each model

8.2.2 Forecasting performance

Figures 5 and 6 represent the forecasted value after the model averaging and the random walk model prediction. Also, the actual value of the S&P 500 in percent change is shown for comparison. The actual S&P 500 is much more volatile compared with the model averaging forecast. Table 30A in the appendix summarized the performance of the predicted values, where the model averaging does predict more accurately than the random walk according to all the forecasting performance values estimated in the table.

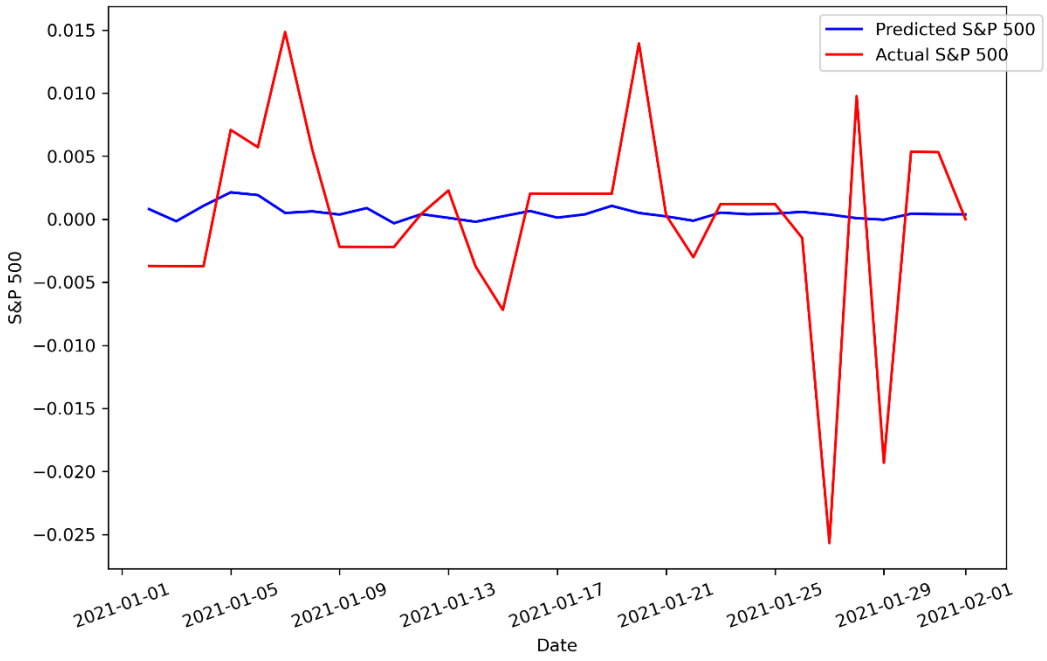

Figure 5: Actual and the most-followed accounts sentiment prediction of S&P 500 in percentage change

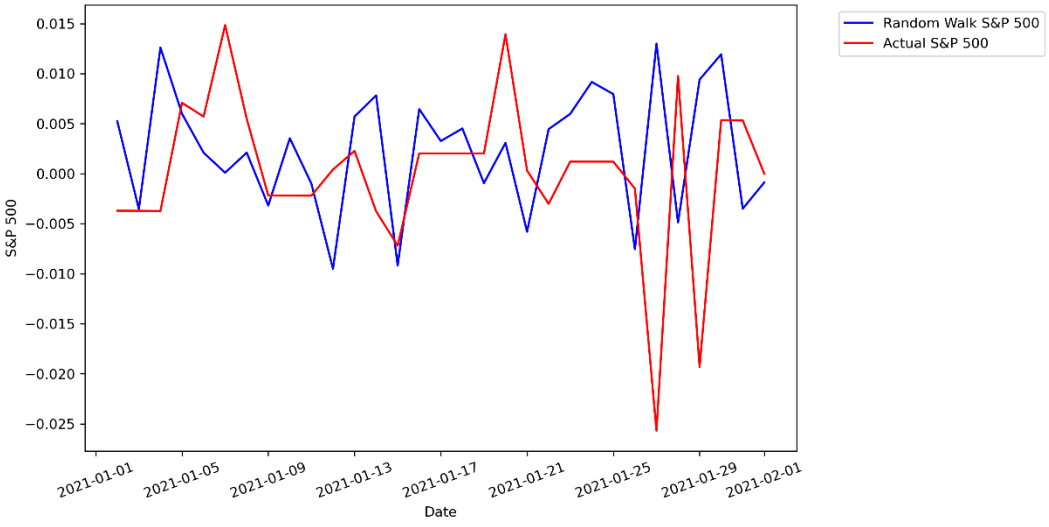

Figure 6: Actual and the random walk prediction of S&P 500 in percentage change

Tables 31A and 32A in the appendix summarize the estimated forecasting performance for the model averaging and the random walk for each week. In

weeks two and three, the random walk does perform better in forecasting than the predicted value from the model averaging. However, the model averaging does perform better in weeks one and four, where the random walk does its most considerable miss calculation in the last week.

8.3 Relationship between sentiment index and S&P 500 index (error correction representation)

By choosing the augmented Dickey-Fuller test with two lagged first-differences, the ADF test suggests that we have a stationary series with an intercept that the value for the coefficient of the first lag of the dependent variable in the ADF equation is -0.0245, which implies the coefficient of the first lag of the dependent variable in DF equation is 1-0.0245=0.9755.

The ADF test of the S&P 500 index with 2 lagged first-differences and a trend component suggests that we cannot reject the null hypothesis of a random walk so there is insufficient evidence that it is trend stationery.

In the ADF test of sentiment index, the test statistic, compared to 1% critical value- is not very large, and the estimated coefficient of the first lag of the dependent variable in the ADF equation suggests that the value of the coefficient of the first lag of the dependent variable in DF equation is close to one. Also, in the ADF test of the S&P 500 index, we could not reject the null hypothesis of a random walk; therefore, we check for cointegration of the two series.

we check for cointegration between the two series by performing an ADF test, with two lagged first-difference and no intercept, on the residuals of the first-difference regression. (Engle & Granger, 1987) (Hamilton, 1994). By looking at Table 33A and comparing the ADF test statistic of -1.24 with the Critical Value for the cointegration Test of −2.76, suggests that we do not have enough evidence to reject the null hypothesis of nonstationary residuals, so we do not have enough evidence to reject the null hypothesis of no cointegration at the 5% level.

To model the relationship between the S&P 500 index and net positivity, we use an ARDL model in error correction form. The reason for choosing such a model is

that it allows us to separate the short-term and long-term relationship between the two series.

Given the stationarity of the first difference of the sentiment index and S&P 500 index (Table 33A) we estimate the ARDL model of the two stationary series based on the error correction representation.

Our goal is to estimate the long-term and short-term effect of the net positivity first difference on the S&P 500 index first difference (Hill et al., 2018). Also, we will be able to use a bounds testing procedure to draw conclusive inference without knowledge about whether variables are I(0) or I(1) (Pesaran et al., 2001) (Engle & Granger, 1987) (Hassler & Wolters, 2006).

To formally analyze the relationship between sentiment and stock index, we postulate a model similar to what Kripfganz et al. discussed. Presented in Table 2. we obtain the estimation, which shows that the long-term coefficient θ is 0.063 representing the effect of change in the sentiment on the change on the S&P 500 index. It suggests that there is a significant, however small in magnitude, long-term relationship between the two series. The α is 1.089, which shows how quickly the distortions from equilibrium are corrected.

Table 2 ARDL(2,2) results derived from Kripfganz, S., and D. C. Schneider (2018) ARDL model

	D.dsp	Coef.	Std. Err.	t	P > |t|	[95% Conf. Interval]	
ADJ							
	dsp						
	L1.	−1.09	.05	−22.74	0.00	−1.18	−1
LR							
	dsen	.06	.02	3.79	0.00	.03	.1
SR							
	dsp						
	LD.	−.12	.03	−3.86	0.00	−.18	−.06
	dsen						
	D1.	−.05	.014	−3.74	0.00	−.08	−.02
	LD.	−.02	.007	−3.17	0.002	−.04	−.009

Sample: 1/1/2017 − 16/3/2021

Number of observations = 1,052
R-squared = 0.62
Adjusted R-squared = 0.62
Root MSE = 34.86

The short-term coefficients of ψ_{yi} and ψ_{xi} account for short-term fluctuations which are not due to deviations from the long-term equilibrium. The short-term coefficients suggest that the lagged first difference of the sentiment score has a positive effect on the S&P 500 index in the current period. However, the estimated coefficient of the first difference of the net positivity score in the current period has a negative effect on that of the S&P 500 index (Kripfganz & Schneider, 2018) (Kripfganz & Schneider, 2020).

In order to test for the existence of a long-term relationship, we use the Pesaran, Shin, and Smith (2001) bounds test, using Kripfganz and Schneider (2018) critical values.

The F-statistic test the joint null hypothesis

$H_0^F: (\alpha = 0) \cap \left(\sum_{j=0}^{q} \beta_j = 0\right)$

versus the alternative hypothesis

$H_1^F: (\alpha \neq 0) \cup \left(\sum_{j=0}^{q} \beta_j \neq 0\right)$

We observe in Table 34A that the H_0^F is rejected, so we test the single hypothesis of

$H_0^t: \alpha = 0$

versus

$H_1^t: \alpha \neq 0$

we observe that the H_0^t is also rejected.

$$\Delta y_t = -\alpha(y_{t-1} - \theta \mathrm{x}_t) + \sum_{i=1}^{p-1} \psi_{yi} \Delta y_{t-i} + \sum_{i=0}^{q-1} \psi_{xi} \Delta \mathrm{x}_{t-i} + u_t \qquad (20)$$

$$\Delta \mathrm{d}SP_t = -1.09 * (dSP_{t-1} - 0.06 * \mathrm{dSEN}_t) + 0.12 * \Delta dSP_{t-1} - 0.05 * \Delta \mathrm{dSEN}_t - 0.25 * \Delta \mathrm{dSEN}_{t-1} + u_t \qquad (21)$$

We find that the estimated coefficients are statistically significant, but not economically. One potential explanation can be that because of news accounts on our list, we face reverse causality, as the news account tweet about the events that have already happened.

9. Traders accounts

So far in the thesis, we estimated the sentiment based on the accounts with high followers, which included mainly news agencies and politicians. We found a mutual Granger causality between the S&P 500 index and most-followed accounts sentiment.

Now we want to check whether there is a causality relationship in one direction between the S&P 500 index and a new sentiment index derived from the famous accounts which tweet mainly about financial markets, we call them “traders” throughout the thesis. As we aim to find the explanatory power of sentiment on the S&P 500 index and the previous forecasts and estimations were not economically significant, we decide to re-run some of our previous models based on traders' accounts sentiment to see whether it can improve our results.

In order to select the account with chose 20 accounts active in the finance field with a high number of followers. The list of accounts and their number of followers as of Jun 30th is presented in table 35A.

By looking at figure 7 we can see that the sentiment obtained from the new set of 20 accounts is less volatile. As we focus on the patterns and changes in the values of the sentiment indices, just for the sake of visual comparison, we have shifted the traders’ sentiment graph by 2400 units and that of most-followed accounts by 2300 units.

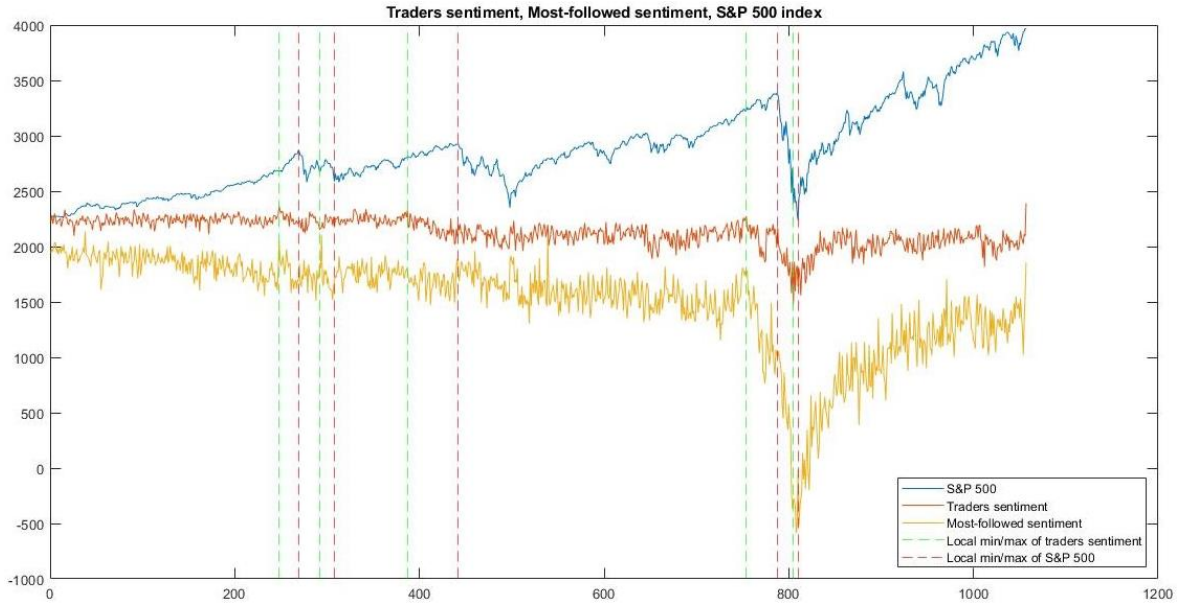

Figure 7 Sentiments and the S&P 500 index

The dashed yellow lines represent the local minimum (maximum) of the sentiment index. The red dashed lines represent the local min/max of the S&P 500, which materialized days after that of the corresponding sentiment index value. The graphs and our model results suggest a great potential for recognizing the most relevant Twitter accounts as the data source to improve the explanatory power of the sentiment index. As we see in the traders' sentiment graph, the effects of sentiment might vary across time, so future work can investigate changes in different periods.

In order to observe the graphs in more detail, we focus on two periods, one from November 2017 till April 2018 (from 230th working day to 320th working day since 1st working day of January 2017) and the other from December 2019 till April 2020.

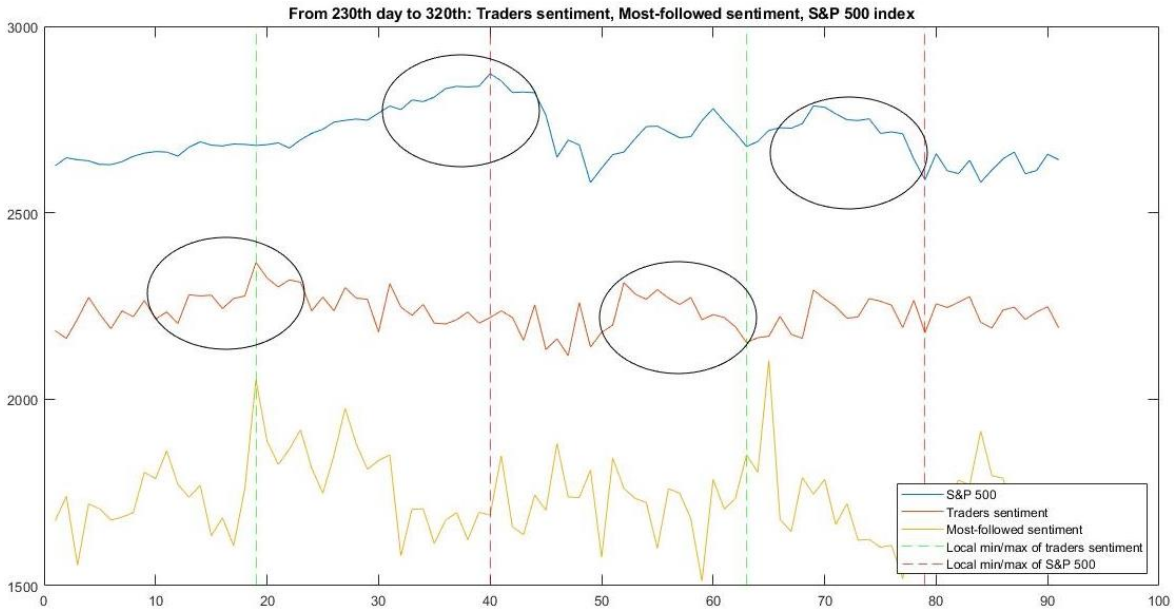

Figure 8 November 2017 till April of 2018

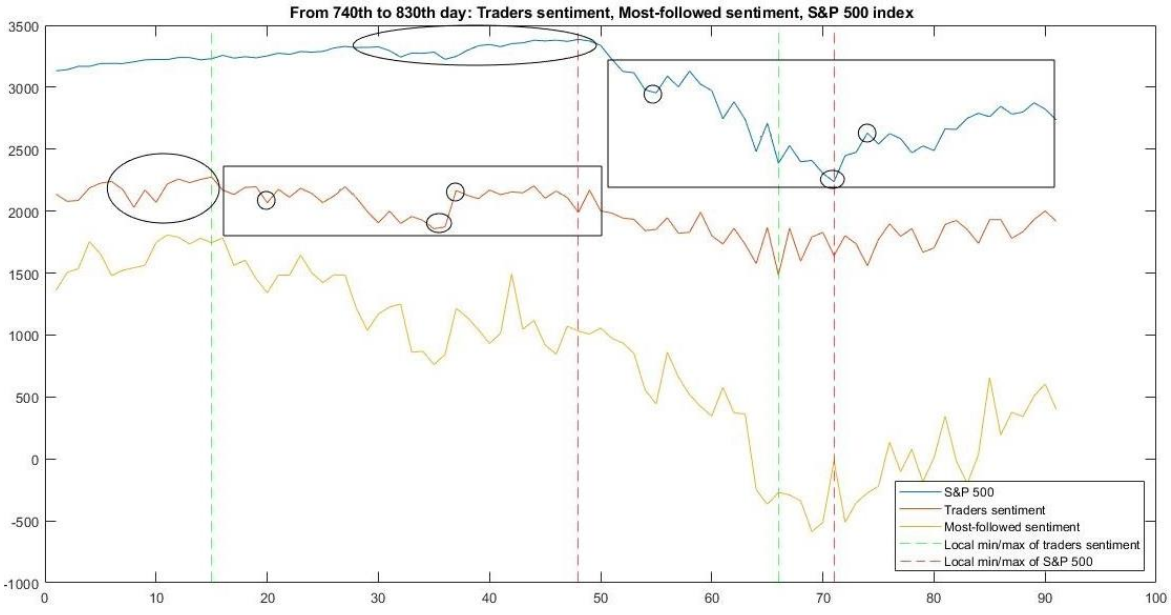

Figure 9 December 2019 till April 2020

In Figure 8 depicting the series from late November 2017 till early in the April 2018, we see that similar patterns occur between the stock market index and the trader accounts sentiment index within an almost 20-day span. More interestingly,

in the period from December 2019 till April 2020, depicted in figure 9, we can see that it looks like the sentiment index is a compressed series of the stock market, so the fluctuations in the sentiment index occur almost a month before that of the stock market index.

The efficient market hypothesis maintains that all the information is reflected in the stock market index (Fama, 1965). However, we see that despite what efficient market hypothesis states, we can see that similar patterns can be shaped in the sentiment index weeks before appearing in the stock market index.

We calculate the correlation coefficient between the S&P 500 index values and our new sentiment index and their first differences up to 40 lags. The results are shown in Table 36A.

The correlation between S&P 500 and the sentiment based on the traders reaches its maximum lag 3 in December 2019 till April 2020 period and lag 17 for November 2017 till April of 2018 period. While when we base the sentiment on the most-followed accounts, the correlation for December 2019 till April 2020 period decreases steadily as the number of lags increase and in November 2017 till April of 2018 period it reaches its maximum at 13^{th} lag with a value of 0.43 which is less than 0.51 which was the case when we had calculated the correlation coefficient based on traders' sentiment.

Regarding the first-differences we do not see any discernible pattern in the correlation values as they switch between negative and positive values in different lags.

To investigate the relationship between sentiment index and S&P 500 index in error correction representation, as shown in table 37A, we estimate an ARDL(4,1) model and observe that the coefficient for the long-term coefficient θ is 0.16, which is greater than 0.063 estimated coefficient in the model based on the most-followed accounts sentiment. The increase in the value of that coefficient and higher value of the adjusted R-squared suggest that as the set of accounts used in the analysis becomes more related to the specific domain at hand, here the stock

market, our model can explain the variation better and the long-term relationship between the two series becomes more evident and economically significant.

By considering the 50-month period from January 2017 to March 2021, we investigate the granger causality test, and, as shown in Table 3, we observe that at 10% significance level, the S&P 500 Granger causes the traders sentiment while it does not suggest the causality relationship in the opposite direction.

Table 3 The p-values of the Granger causality tests

Dependent / Independent	S&P 500	Traders' sentiment
Traders' sentiment	0.88	-
S&P 500	-	0.07

Compared to the Granger causality between most-followed accounts sentiment and S&P our results suggest that while the most-followed accounts sentiment Granger causes the S&P 500, we cannot say there is such relationship between the traders' sentiment and S&P 500. The causality channel is reversed when we consider the traders' sentiment.

We use the same method to predict the S&P 500 using VAR models, models averaging, and in-sample forecasting for the traders' sentiment, as we did previously for the most-followed accounts' sentiment. From figures 10 and 11, we observe that the Trader's sentiment forecast has similar movements to the most-followed accounts prediction.

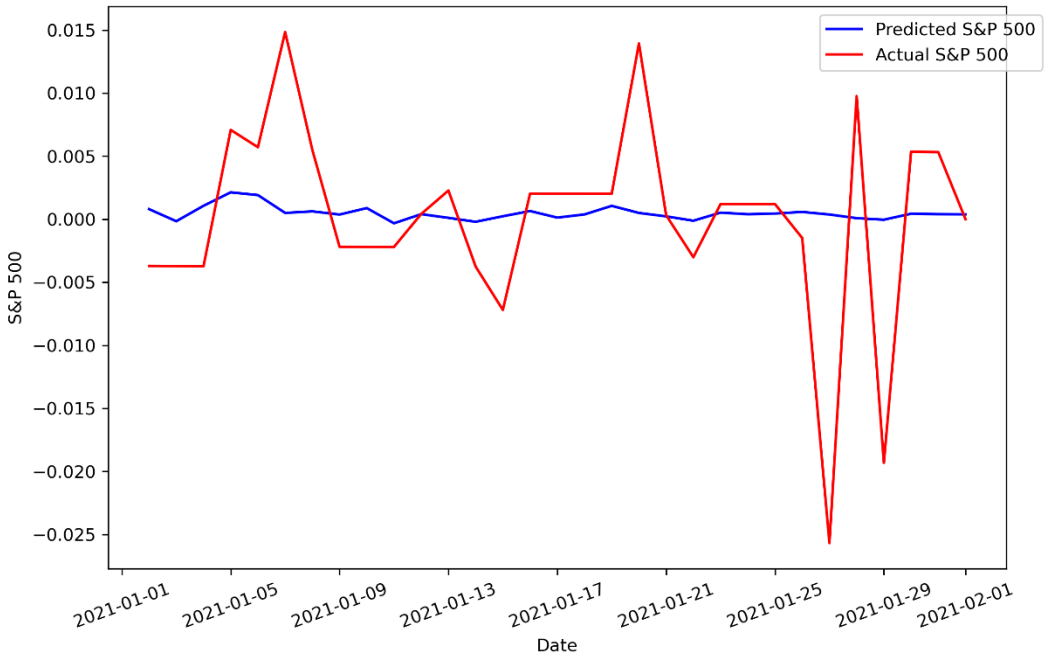

Figure 10: Actual and the most-followed accounts sentiment prediction of S&P 500 in percentage change

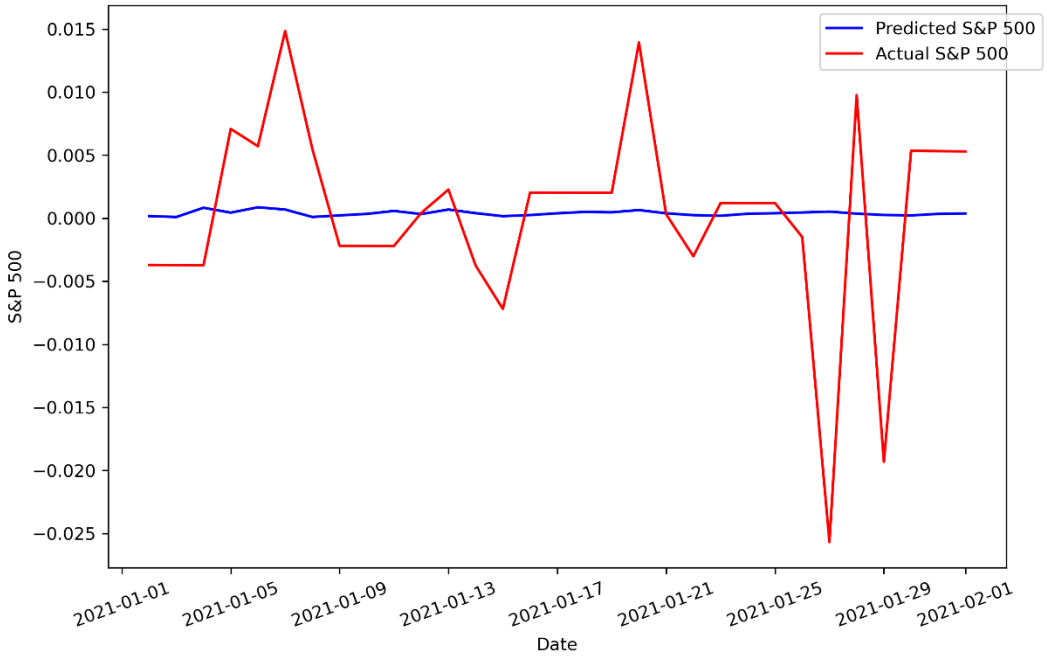

Figure 11: Actual and Trader's sentiment prediction of S&P 500 in percentage change

The results from table 30A suggest the trader's sentiment performs better in predictions. By comparing tables, 31A and 38A, we see the estimated forecasting performance values indicate that the most-followed accounts sentiment performs better predicting weeks one and four. According to the estimated performance values, in weeks two and three, the Trader's accounts' predicted value is much more accurate and closer to the accuracy in the random walk in these periods. In general, the results suggest that the Trader's sentiment is more accurate to predict than the two other predictions but cannot capture the variation of the S&P 500.

What we summarized from the traders' sentiment suggests a great potential for more extensive work in the future.

10.Discussion and future research

The current finding of us about the higher visual similarity between the traders' sentiment and the S&P 500 index hints at rewarding results in future research so performing more and deeper econometrics analysis using the traders' sentiment index and running models in different periods is encouraged.

The advantage of looking into most followed accounts like news agencies and politicians is that after deleting click-baits and cleaning the data, there is a lower chance of facing spam or sarcasm. Hence spam detection and sarcasm detection challenges would be mitigated in the data at hand. However, the presence of news accounts on our list can lead to a reverse causality as the news account tweet about the events that have already happened. Nevertheless, our estimated coefficients in the error correction model suggest significant but quantitatively small results.

Given that the tokenization has been done at the word level and that we have used a lexicon-based approach, we still face the challenge of handling negation and the intensity of the negation. For example, the tweets "a vaccine has been produced, but it is not good" and "no vaccine has been developed yet, bad news for the world's greatest economy" have different negation scopes while our lexicon-based approach, which cannot cope with the context in which words appear, will not be

able to catch neither their scope nor the polarity of them. Advanced methods can recognize valence shifters like negators and amplifiers that change the direction or intensity of the sentiment (Polanyi & Zaenen, 2006).

Another limitation of our approach is its inability to recognize the tweets' subjectivity, which will be an issue in tweets that include direct quotations on controversial topics.

The process of creating a dictionary and annotating it for a lexicon-based approach is tedious and time-consuming *Lexicon-Based Methods for Sentiment Analysis* (Taboada et al., 2011). As an alternative method, numerous articles try to utilize machine learning techniques that do not require a dictionary and have a higher precision at the cost of requiring more time and more dependency on the domain. ML approach is divided into supervised or unsupervised learning techniques wherein the supervised learning algorithm can apply what it has learned from the past to the new data. Unsupervised learning does not require labeled data it aims is to model the data so that it can learn the most from it (Jindal & Aron, 2021).

Future research can be done using subjectivity detection or machine learning techniques that require more time but can increase the precision and accuracy of the results of sentiment estimation. The authors of *Stock Prediction Using Event-Based Sentiment Analysis* (Makrehchi et al., 2013), generate training data based on stock markets events and use it to build a classifier for assessing the tweet sentiment. They also create an autoregressive model to account for the historical dependence of the series. The novelty of their approach is that they have been able to generate the training data automatically

Patterns linking sentiment and topics with S&P 500 index may vary over time, and future work can investigate changes in different periods. For example, the outcomes may differ between the period before March 2020 and the period after it.

In sum, future research can be done on both the computational and econometric aspects of our thesis to improve the forecasting power of the estimation accuracy of the sentiment index.

11. Conclusion

In this thesis, we collected and cleaned almost 1.1 million uncleaned tweets from most-followed accounts and 0.6 million tweets from traders' accounts. After pre-processing, we evaluated the net positivity of each tweet and aggregated the results on daily intervals for the most-followed accounts and those who tweet mainly about financial markets. Furthermore, we investigated the existence of long-term and short-term relationship between the net positivity score and daily closing stock prices of the Standard and Poor's 500 companies from the 1st working day of Jan of 2017 till the 16th of Mar of 2021.

We find that the granger causality between most-followed accounts sentiment and S&P suggests that while the most-followed accounts sentiment granger (mutually) causes the S&P 500, it is the S&P 500 which Granger causes the traders sentiment.

We built a forecasting model to predict the S&P 500, using variables extracted from Twitter, i.e., net positivity and topics. We compared the results with a random walk model and showed that the predictions from the model averaging perform better than the random walk. However, we saw that the forecasting model is less accurate than the random walk model in specific weeks. As various variables affect the S&P 500, predicting with higher precision requires a more extensive model. We saw that the models with most-followed sentiment and random walk perform worse than the models with traders' sentiment in predictions.

We investigated the effect of net positivity on the S&P 500. The Granger causality test suggests a mutual Granger causality between the net positivity based on the 18 accounts and S&P 500. We also showed that corona-related topics significantly affected the S&P 500.

We can confirm the result of (Kräussl & Mirgorodskaya, 2017), in their paper, they have mentioned that despite that previous literature suggests a negative association between media pessimism and contemporaneous market returns (Antweiler & Frank, 2004);(GARCÍA, 2013);(Goetzmann et al., 2016);(Tetlock, 2007), they find that over their three-year study horizon the media pessimism is associated with market performance in the long run.

We found that even though the estimated coefficients are statistically significant, but they are quantitatively small in magnitude. The reason for these results of our estimated error correction model might be related to the presence of news accounts on our list. The content on the Twitter accounts of news agencies in many cases might imply a reverse causality as the news account tweet about the events which have already happened.

Further research with better hardware and more complicated deep learning algorithms can be used to capture nuance sentiment of tweets and improve the results with and perform precise predictions about the movements of the stock market indices.

12. REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*(1), 243–247. https://doi.org/10.1007/BF02532251

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). ECONOMETRICS MEETS SENTIMENT: AN OVERVIEW OF METHODOLOGY AND APPLICATIONS. *Journal of Economic Surveys*, *34*(3), 512–547. https://doi.org/10.1111/joes.12370

Allen, R. L., & Davis, A. S. (2011). Hawthorne Effect. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (pp. 731–732). Springer US. https://doi.org/10.1007/978-0-387-79061-9_1324

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, *59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

BAKER, M., & WURGLER, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Baker, M., & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, *21*(2), 129–152. https://doi.org/10.1257/jep.21.2.129

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Beber, A., & Brandt, M. W. (2010). When It Cannot Get Better or Worse: The Asymmetric Impact of Good and Bad News on Bond Returns in Expansions and Recessions. *Review of Finance*, *14*(1), 119–155.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K Text to Gauge Financial Constraints. *Journal of Financial and Quantitative Analysis*, *50*(4), 623–646.

Borovkova, S., Garmaev, E., Lammers, P., & Rustige, J. (2017). *SenSR: A sentiment-based systemic risk indicator* [DNB Working Papers]. Netherlands Central Bank, Research Department. https://EconPapers.repec.org/RePEc:dnb:dnbwpp:553

Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics*, *133*(2), 299–336. https://doi.org/10.1016/j.jfineco.2018.11.009

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in*

Neural Information Processing Systems (Vol. 22). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, *37*(3/4), 409–428. JSTOR. https://doi.org/10.2307/2332391

Eklund, J., & Karlsson, S. (2007). Forecast Combination and Model Averaging Using Predictive Measures. *Econometric Reviews*, *26*(2–4), 329–363. https://doi.org/10.1080/07474930701220550

El-Amir, H., & Hamdy, M. (2019). *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow*. Apress. https://books.google.no/books?id=2lyRyAEACAAJ

Engle, R., & Granger, C. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, *55*(2), 251–276.

Fama, E. F. (1965). Random Walks in Stock-Market Prices. *Financial Analysts Journal*, *21*, 55–59.

GARCÍA, D. (2013). Sentiment during Recessions. *The Journal of Finance*, *68*(3), 1267–1300. https://doi.org/10.1111/jofi.12027

Garz, M. (2014). Good news and bad news: Evidence of media bias in unemployment reports. *Public Choice*, *161*(3), 499–515. https://doi.org/10.1007/s11127-014-0182-2

Glasserman, P., & Mamaysky, H. (2019). Does Unusual News Forecast Market Stress? *Journal of Financial and Quantitative Analysis*, *54*(5), 1937–1974. https://doi.org/10.1017/S0022109019000127

Goetzmann, W., Kim, D., & Shiller, R. (2016). *Crash Beliefs From Investor Surveys* (NBER Working Papers No. 22143). National Bureau of

Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:22143

Hamilton, J. D. (1994). *Time Series Analysis* (1st ed.). Princeton University Press. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0691042896

Hannan, E. J., & Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 190–195. JSTOR.

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hassler, U., & Wolters, J. (2006). Autoregressive distributed lag models and cointegration. *AStA Advances in Statistical Analysis*, *90*(1), 59–74.

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *The Journal of Business Communication (1973)*, *45*(4), 363–407. https://doi.org/10.1177/0021943608319388

Herwartz, H., & Kholodilin, K. A. (2011). *In-Sample and Out-of-Sample Prediction of Stock Market Bubbles: Cross-Sectional Evidence* (Discussion Papers of DIW Berlin No. 1173). DIW Berlin, German Institute for Economic Research. https://ideas.repec.org/p/diw/diwwpp/dp1173.html

Heston, S. L., & Sinha, N. R. (2016). News versus Sentiment: Predicting Stock Returns from NewsStories. *Finance and Economics Discussion Series*

2016-04 Ashington: Board of Governors of the Federal Reserve System. http://dx.doi.org/10.17016/FEDS.2016.048

Heston, S. L., & Sinha, N. R. (2017). News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal*, *73*(3), 67–83. https://doi.org/10.2469/faj.v73.n3.3

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of Econometrics*. Wiley. https://books.google.no/books?id=UdVSDwAAQBAJ

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jindal, K., & Aron, R. (2021). A systematic study of sentiment analysis for social media data. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.048

Kearney, C., & Liu, S. (2014). Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis*, *33*. http://dx.doi.org/10.2139/ssrn.2213801

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. Macmillan.

Kleinnijenhuis, J., Schultz, F., Oegema, D., & Atteveldt, W. van. (2013). Financial news and market panics in the age of high-frequency sentiment trading algorithms. *Journalism*, *14*(2), 271–291. https://doi.org/10.1177/1464884912468375

Kolchyna, O., & Tharsis T. P. Souza, T. A., Philip Treleaven. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. *ArXiv*.

Kräussl, R., & Mirgorodskaya, E. (2017). Media, sentiment and market performance in the long run. *The European Journal of Finance*, *23*(11), 1059–1082. https://doi.org/10.1080/1351847X.2016.1226188

Kripfganz, S., & Schneider, D. C. (2018). ARDL: Estimating Autoregressive Distributed Lag and Equilibrium Correction Models. *Proceedings of the 2018 London Stata Conference*.

Kripfganz, S., & Schneider, D. C. (2020). *Response Surface Regressions for Critical Value Bounds and Approximate p-values in Equilibrium Correction Models*.

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203–218. https://doi.org/10.1016/j.jeconom.2018.11.013

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets* (Second). Cambridge University Press. http://mmds.org

Liu, H. (2010). Feature Selection. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 402–406). Springer US. https://doi.org/10.1007/978-0-387-30164-8_306

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. https://doi.org/j.1540-6261.2010.01625.x

LOUGHRAN, T., & MCDONALD, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Lütkepohl, H. (2013). Vector autoregressive models. In N. Hashimzade & M. A. Thornton (Eds.), *Handbook of Research Methods and Applications in Empirical Macroeconomics* (pp. 139–164). Edward Elgar Publishing. https://ideas.repec.org/h/elg/eechap/14327_6.html

Makrehchi, M., Shah, S., & Liao, W. (2013). Stock Prediction Using Event-Based Sentiment Analysis. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 337–342. https://doi.org/10.1109/WI-IAT.2013.48

Malliaris, M. E. (1994). Modeling the behavior of the S P 500 index: A neural network approach. *Proceedings of the Tenth Conference on Artificial Intelligence for Applications*, 86–90. https://doi.org/10.1109/CAIA.1994.323688

Mayr, J., & Ulbricht, D. (2007). *VAR Model Averaging for Multi-Step Forecasting* (Ifo Working Paper No. 48). ifo Institute - Leibniz Institute for Economic Research at the University of Munich. http://hdl.handle.net/10419/73799

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. https://doi.org/10.1016/j.ijforecast.2019.02.011

Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R., Corvellec, M., Medina, J., Dai, Y., Petrushev, B.,

Langner, K. M., Hong, Alessio, Ozsvald, I., vkolmakov, Jones, T., Bailey, E., … Mai, F. (2018). *amueller/word_cloud: WordCloud 1.5.0* (1.5.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1322068

Nau, R. (2014). Notes on the random walk model. *Fuqua School of Business*, *1*, 1–19.

Nimark, K. P., & Pitschner, S. (2019). News media and delegated information choice. *Journal of Economic Theory*, *181*, 160–196. https://doi.org/10.1016/j.jet.2019.02.001

Nisar, T. M., & Yeung, M. (n.d.). *Twitter as a tool for forecasting stock market movements: A short-window event study*. https://doi.org/10.1016/j.jfds.2017.11.002

OSINT team. (n.d.). *TWINT*. OSINT team. https://github.com/twintproject/twint

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). *Bounds testing approaches to the analysis of level relationships*.

Petropoulos Petalas, D., van Schie, H., & Hendriks Vettehen, P. (2017). Forecasted economic change and the self-fulfilling prophecy in economic decision-making. *PLOS ONE*, *12*(3), 1–18. https://doi.org/10.1371/journal.pone.0174353

Picault, M., & Renault, T. (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance*, *79*(C), 136–156.

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text*.

Ren, R., Wu, D. D., & Liu, T. (2019). Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Systems Journal*, *13*(1), 760–770. https://doi.org/10.1109/JSYST.2018.2794462

Rickard Nyman, Paul Ormerod, Robert Smith, & David Tuckett. (n.d.). *Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis*.

Saleiro, P., Rodrigues, E. M., Soares, C., & Oliveira, E. (2017). TexRep: A Text Mining Framework for Online Reputation Monitoring. *New Generation Computing*, *35*(4), 365–389. https://doi.org/10.1007/s00354-017-0021-3

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. JSTOR.

Shiller, R. J. (2017). Narrative Economics. *American Economic Review*, *107*(4), 967–1004. https://doi.org/10.1257/aer.107.4.967

S&P Dow Jones Indices LLC, S&P 500 [SP500], retrieved from FRED, Federal Reserve Bank of St. Louis. (2021). https://fred.stlouisfed.org/series/SP500

String python module. (n.d.). https://docs.python.org/3/library/string.html

Swamynathan, M. (2019). *Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python* (2nd ed.). Apress.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Täckström, O., & McDonald, R. T. (2011). Discovering Fine-Grained Sentiment with Latent Variable Structured Prediction Models. *ECIR*.

team, T. pandas development. (2020). *pandas-dev/pandas: Pandas* (latest) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance*, *62*(3), 1139–1168.

TextBlob: Simplified Text Processing. (n.d.). https://textblob.readthedocs.io/en/dev/

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

12. Appendix

12.1 Figures

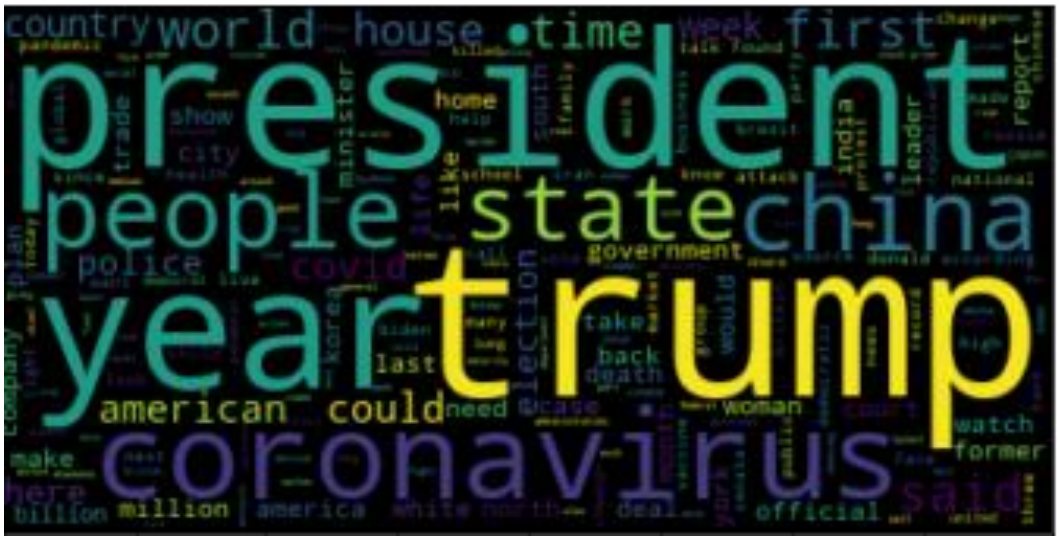

Figure 1A WordCloud

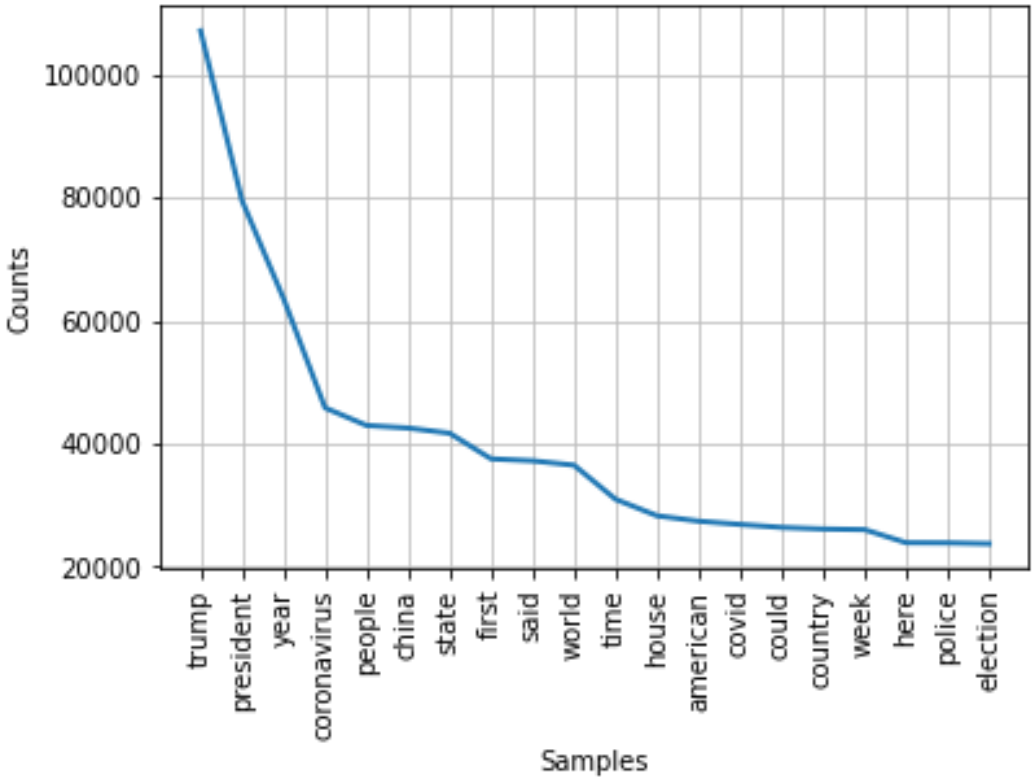

Figure 2A Frequency distribution

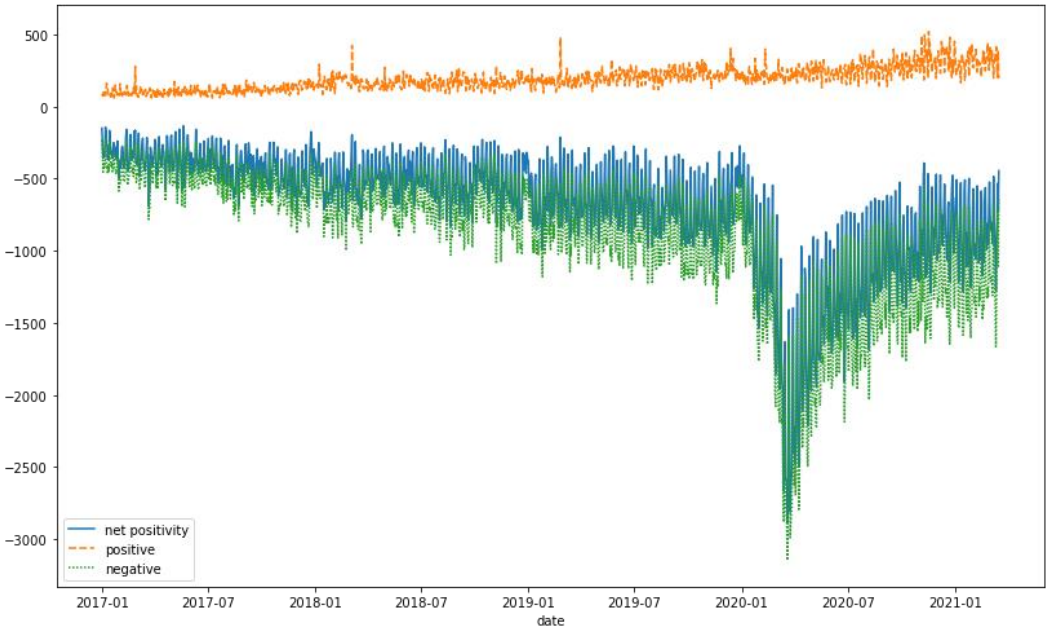

Figure 3A Net positivity

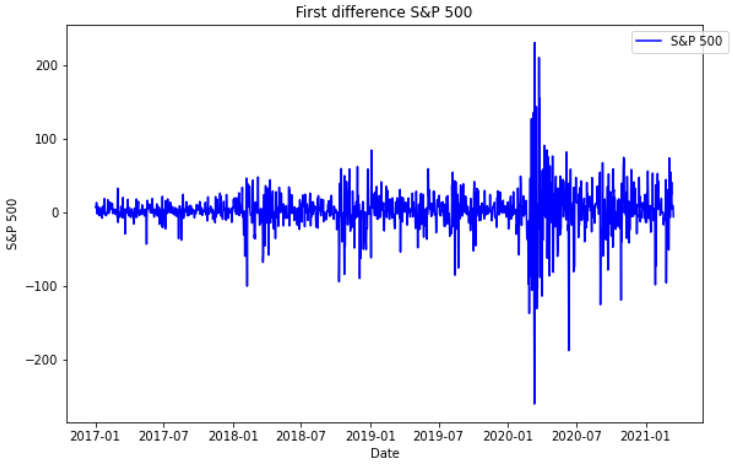

Figure 4A: First difference of S&P 500

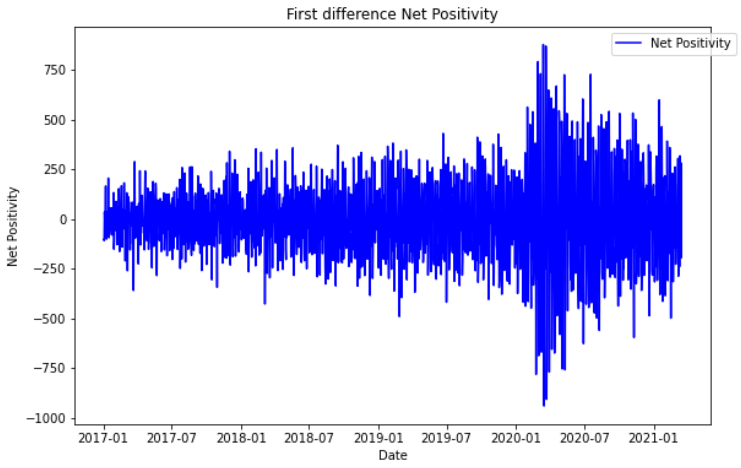

Figure 5A: First difference of Net Positivity

12.2 Tables

Table 1A List of accounts

Account	**Description**	**Number of Followers** **(Numbers in parathesis represent the number of followers before the change of government in the United States of America or account suspension)**
Barack Obama	The Twitter account of an American politician for the democrats and the former President of the United States	129.5M
Donald Trump	The Twitter account of an American business owner, politician, and the former President of the United States	- (88.7M)
CNN BRK	The Twitter account of the multinational news-based pay television covering breaking news	61M
Narendra Modi	The Twitter account of the prime minister of India.	68.8M

CNN	The Twitter account of a multinational news-based pay television	53.8M
The New York Times	The Twitter account of an American newspaper	49.8M
BBC BRK	The Twitter account of the national broadcaster of the United Kingdom covering breaking news	47.8M
Elon Musk	The Twitter account of an entrepreneur and CEO of multiple corporations	56.9M
PMO India	The Twitter account of the Prime Minister of India	42.6M
POTUS	The Twitter account of the President of the United States	11.7M (33.4M)
BBC World	The Twitter account of the national broadcaster of the United Kingdom covering news from all over the world.	32M

Hillary Clinton	The Twitter account of an American politician and the former secretary of state.	30.8M
The White House	The Twitter account belonging to the US government	5.4M (26.1M)
The Economist	The Twitter account of the international weekly newspaper based in London	25.6M
Joe Biden	The Twitter account of the current President of the United States.	30.5M
Reuters	The Twitter account of an international news organization	23.4M
President Trump 45 Archived	An archived Twitter account of Trump's Administration account, maintained by the National Archives and Records Administration.	32M
The White House 45 Archived	An archived Twitter account of Trump's Administration account, maintained by the	26M

	National Archives and Records Administration.	

Table 2A Python packages, modules and libraries

Name	Utilization
TWINT (OSINT team, n.d.)	To extract the tweets and accounts information
Pandas (team, 2020)	To manage and manipulate our dataset, together with sorting and aggregating (McKinney, 2010)
String module (*String Python Module*, n.d.)	To handle punctuations
NLTK (Bird et al., 2009)	To perform multiple operations like removing stop words that are included in that package. Based on our data, we later removed some other commonly used words that do not carry much information
TextBlob (*TextBlob: Simplified Text Processing*, n.d.)	To perform lemmatization
WordCloud (Mueller et al., 2018), seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007)	To create the word cloud and figures
scikit-learn (Pedregosa et al., 2011)	To perform machine learning algorithms
Numpy (Harris et al., 2020)	To do numerical computations

Table 3A: First setup for the VAR model

Model	Endogenous variable 1	Endogenous variable 2
First setup	S&P 500	Net Positivity

Table 4A: Second setup for the VAR models

Second setup model	Endogenous variable 1	Endogenous variable 2
partial gov shutdown/covid spread	S&P 500	partial gov shutdown/covid spread
vaishnava jana/meghan markle	S&P 500	vaishnava jana/meghan markle
calendar special days2	S&P 500	calendar special days2
biden	S&P 500	biden
el paso shooting	S&P 500	el paso shooting
covid 19	S&P 500	covid 19
president donald trump	S&P 500	president donald trump
coronavirus spreading china	S&P 500	coronavirus spreading china

Table 5A: Third setup for the VAR models

Model	Endogenous variable 1	Endogenous variable 2	Endogenous variable 3

partial gov shutdown/covid spread	S&P 500	Net Positivity	partial gov shutdown/covid spread
vaishnava jana/meghan markle	S&P 500	Net Positivity	vaishnava jana/meghan markle
calendar special days2	S&P 500	Net Positivity	calendar special days2
biden	S&P 500	Net Positivity	biden
el paso shooting	S&P 500	Net Positivity	el paso shooting
covid 19	S&P 500	Net Positivity	covid 19
president donald trump	S&P 500	Net Positivity	president donald trump
coronavirus spreading china	S&P 500	Net Positivity	coronavirus spreading china

Table 6A: ADF Unit Root Test Results

Variable	P-Value	Lag length
S&P 500	0.00	13
Net Positivity	0.00	8
Partial gov shutdown/Covid spread	0.00	1
Vaisnava Jana/Meghan Markle	0.00	1
Calendar special days 2	0.00	1
Biden	0.0001	1
El Paso shooting	0.00	1
Covid 19	0.0004	1
President Donald Trump	0.00	1
Coronavirus spreading China	0.00	1

Table 7A: Optimal lag length

Model	FPE	AIC	HQIC	SBIC
First setup	42	42	14*	14*
Partial gov shutdown/Covid spread third setup second setup	46	46	33*	1
vaishnava jana/Meghan Markle second setup	33*	33*	33*	1
calendar special days2 second setup	33*	33*	1	1
Biden second setup	33*	33*	1	1
el paso shooting second setup	33*	33*	33*	32
Covid 19 second setup	45	45	33*	1
President Donald Trump second setup	33*	33*	1	1
coronavirus spreading china second setup	48	48	46*	1
Partial gov shutdown/Covid spread third setup third setup	42	42	14*	14*
vaishnava jana/Meghan Markle third setup	33	33	14*	8
calendar special days2 third setup	33	33	14*	8
Biden third setup	33	33	14*	8
el Paso shooting third setup	33	33	14*	8
Covid 19 third setup	43	43	14*	13
President Donald Trump third setup	33	33	14*	8
coronavirus spreading china third setup	48	48	46	8*

Table 8A: Durbin-Watson Statistic results

Model	Durbin-Watson Statistic value
First setup	1.9916
Partial gov shutdown/Covid spread second setup	2.0024
vaishnava jana /Meghan Markle second setup	1.9975
calendar special days2 second setup	1.9968
Biden second setup	1.9958
el paso shooting second setup	1.9963
Covid 19 second setup	2.0024
President Donald Trump second setup	1.9976
coronavirus spreading china second setup	2.0068
Partial gov shutdown/Covid spread third setup	2.0014
vaishnava jana /Meghan Markle third setup	1.9911
calendar special days2 third setup	1.9908
Biden third setup	1.9909
el paso shooting third setup	1.9932
Covid 19 third setup	2.0018
President Donald Trump third setup	1.9928
coronavirus spreading china third setup	1.9625

Table 9A: VAR regression Output

Model	R-Squared	RMSE	Adj R-Squared
First setup	0.1216	25.3907	0.1040
Partial gov shutdown/Covid spread third setup second setup	0.2411	24.0674	0.2045
vaishnava jana/Meghan Markle second setup	0.1423	25.5849	0.1009
calendar special days2 second setup	0.1197	25.812	0.0955
Biden second setup	0.1311	25.7518	0.0892
el paso shooting second setup	0.1378	25.6521	0.0963
Covid 19 second setup	0.1867	24.6341	0.1475

President Donald Trump second setup	0.1308	25.7557	0.0889
coronavirus spreading china second setup	0.3783	22.0902	0.3355
Partial gov shutdown/Covid spread third setup third setup	0.1841	24.6613	0.1596
vaishnava jana/Meghan Markle third setup	0.1314	25.3715	0.1053
calendar special days2 third setup	0.1313	25.3732	0.1052
Biden third setup	0.1207	25.3995	0.0943
el paso shooting third setup	0.1335	25.3397	0.1075
Covid 19 third setup	0.1290	24.6613	0.1546
President Donald Trump third setup	0.1793	25.4059	0.1028
coronavirus spreading china third setup	0.0792	25.9073	0.0634

Table 10A: The p-values of the Granger causality tests for the model with the topic "Partial gov shutdown/Covid spread" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.00*	-
S&P 500	-	0.00*

Table 11A: The p-values of the Granger causality tests for the model with the topic "Partial gov shutdown/Covid spread" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.00*	0.0003*	-
Net Positivity	0.00*	-	0.0125*
S&P 500	-	0.00*	0.0013*

Table 12A: The p-values of the Granger causality tests for the model with the topic "Covid 19" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.00*	-
S&P 500	-	0.0018*

Table 13A: The p-values of the Granger causality tests for the model with the topic "Covid 19" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.00*	0.0006*	-
Net Positivity	0.0001*	-	0.0372*
S&P 500	-	0.00*	0.0157*

Table 14A: The p-values of the Granger causality tests for the model with the topic "Coronavirus spreading in China" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.00*	-
S&P 500	-	0.1074*

Table 15A: The p-values of the Granger causality tests for the model with the topic "Coronavirus spreading in China" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.0008*	0.0712	-

Net Positivity	0.0003*	-	0.043*
S&P 500	-	0.00*	0.0043*

Table 16A: The p-values of the Granger causality tests for the model with the topic "Vaishnava Jana/Meghan Markle" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.2188	-
S&P 500	-	0.9992

Table 17A: The p-values of the Granger causality tests for the model with the topic "Vaishnava Jana/Meghan Markle" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.00*	0.9315	-
Net Positivity	0.0003*	-	0.5927
S&P 500	-	0.00*	0.9907

Table 18A: The p-values of the Granger causality tests for the model with the topic "Calendar Special days 2" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.6005	-
S&P 500	-	0.9328

Table 19A: The p-values of the Granger causality tests for the model with the topic "Calendar Special days 2" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic

Topic	0.3143	0.9867	-
Net Positivity	0.00*	-	0.9936
S&P 500	-	0.00*	0.9855

Table 20A: The p-values of the Granger causality tests for the model with the topic "Biden" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.9532	-
S&P 500	-	0.1607

Table 21A: The p-values of the Granger causality tests for the model with the topic "Biden" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.5268	0.1567	-
Net Positivity	0.00*	-	0.0035*
S&P 500	-	0.00*	0.3248

Table 22A: The p-values of the Granger causality tests for the model with the topic "El Paso shooting" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.5374	-
S&P 500	-	0.9998

Table 23A: The p-values of the Granger causality tests for the model with the topic "El Paso shooting" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.1369	0.6041	-
Net Positivity	0.00*	-	0.4917
S&P 500	-	0.00*	0.7751

Table 24A: The p-values of the Granger causality tests for the model with the topic "President Donald Trump" in the second setup

Dependent / Independent	S&P 500	Topic
Topic	0.9599	-
S&P 500	-	0.9971

Table 25A: The p-values of the Granger causality tests for the model with the topic "President Donald Trump" in the third setup

Dependent / Independent	S&P 500	Net Positivity	Topic
Topic	0.5844	0.7018	-
Net Positivity	0.00*	-	0.5389
S&P 500	-	0.00*	0.9375

Table 26A: the percentage change of S&P 500 in response to net positivity shocks for the first and third setup with the variables containing the topics

Lag	First setup	Partial gov shutdown/Covid spread third setup	vaishnava jana/Meghan Markle	calendar special days2	biden	el paso	Covid 19	President Donald Trump	coronavirus spreading china
1	.53	0.24	0.59	0.59	0.45	0.53	0.16	0.43	-0.67
2	2.69*	2.62*	2.70*	2.67*	2.78*	2.69*	2.33*	2.75*	2.71*
3	1.22	0.79	1.28*	1.20	1.19	1.33*	0.98	1.25*	0.89
4	1.41*	1.12	1.48*	1.49*	1.30*	1.39*	1.09	1.42*	0.85
5	-.87	-0.95	-0.82	-0.84	-0.79	-0.94	1.09	-0.83	0.36
6	-.77	-1.03	-0.71	-0.78	-0.78	-0.76	-0.72	-0.83	0.17
7	.91	0.66	0.92	0.89	0.91	0.87	0.35	0.93	0.35
8	1.03	0.78	1.07	1.05	0.97	1.04	0.64	0.98	-0.39
9	.89	0.89	0.84	0.87	0.86	0.79	0.81	0.81	-0.09
10	.29	0.28	0.21	0.27	0.29	0.21	0.01	0.29	0.53
11	.38	0.26	0.33	0.33	0.40	0.34	0.23	0.44	-0.08
12	.30	0.03	0.26	0.28	0.27	0.27	-0.06	0.29	-0.8*
13	1.07*	1.03*	1.07*	1.09*	1.05*	1.09*	1.04*	1.06*	-0.09
14	.55	0.69	0.53	0.57	0.59	0.55	0.67	0.56	0.25

Table 27A: the percentage change of S&P 500 in response to net topics shocks for the third setup with the variables containing the topics

lag	Covid 19	Partial gov shutdown/Covid spread third setup second setup	President Donald Trump	vaishnava jana/Meghan Markle	calendar special days2	Biden	Covid spread in China	El paso
1	.51	2.35*	.49	.24	.82	-.82	-1.39*	.66
2	-1.46*	-2.60*	.94	.91	.36	-.63	.64	.95
3	1.57*	1.08	.22	.24	.36	-1.22	-1.60*	.78
4	1.17	.19	-.24	.03	-.91	-.60	-.53	-.91
5	-.27	-.31	-.12	.55	.57	.04	-.43	-.13
6	.55	.19	.67	.44	1.60*	-1.40*	-.32	-.90
7	.81	-.40	-.22	-.11	-.17	.69	-.51	.21
8	-2.94*	-2.24*	1.38*	1.67*	-.56	.58	.65*	.13
9	1.34	.36	-.24	1.04	-.09	.30	.78*	.31
10	3.62	3.50*	-.27	-1.07	.22	-.22	.52*	.33

11	-2.54*	-3.06*	.65	-.05	-1.17	-.25	.97*	-1.14
12	1.49	-.13	-.61	-.42	-.50	.01	.82*	1.46*
13	1.09	1.46*	-.46	.18	-.74	.24	.83*	-.44
14	-.87	-.36	.18	-.35	.10	.02	.73*	-.08
15	-1.02*	-1.36*	-.45	.05	-.02	-.19	.80*	-.31

Table 28A: The percentage change of S&P 500 in response to net topics shocks for the second setup with the variables containing the topics

lag	Covid 19	Partial gov shutdown/Covid spread third setup second setup	President Donald Trump	vaishnava jana/Meghan Markle	calendar special days2	Biden	Covid spread in China	El paso
1	.76	2.01*	.48	.01	.92	-.92	-.98	1.07
2	-1.28*	-2.3*	.84	.64	.39	-.48	-.24	.74
3	1.42*	.91	.08	.19	.45	-1.30*	-.97	.66
4	1.28*	.57	-.18	-.09	-.77	-.66	.34	-.71
5	-.47	-.69	.02	.41	.38	.02	.93	-.08
6	.23	-.04	.70	.35	1.44*	-1.24	-.35	-.71
7	.78	-.10	-.18	-.23*	.05	.52	.32	.168
8	-2.96*	-2.40*	1.41*	1.63	-.43	.46	1.50*	.23
9	1.41*	.50	-.09	.83	-.019	.27	-1.25*	.13
10	3.34*	3.08*	-.23	-1.15	.45	-.37	-.67	.40
11	-2.72*	-2.82*	.62	.13	-1.03	-.15	1.01	-.94
12	1.49*	-.37	-.73	-.30	-.22	.09	.31	1.10*
13	1.49	1.33*	-.40	.20	-.46	.25	-1.07	-.157
14	.97	.95	.13	-1.19	.68	.01	.71	-.75
15	-2.51*	-2.72*	.34	.038	.35	-.07	.35	-.35

Table 29A: Estimated RMSE from each predicted model's

Model	RMSE
First setup	49.3983
partial gov shutdown/covid spread second setup	67.0911
vaishnava jana/Meghan Markle	63.7975
calendar special days2 second setup	76.1754
biden second setup	61.8862
el paso shooting second setup	63.6337
covid 19 second setup	87.2518
president donald trump second setup	61.0145
coronavirus spreading china second setup	61.8675
partial gov shutdown/covid spread third setup	50.2438
vaishnava jana/meghan markle third setup	49.4055
calendar special days2 third setup	51.9706
biden third setup	47.8629
el paso shooting third setup	49.6738
covid 19 third setup	82.1629
president donald trump third setup	47.4448
coronavirus spreading china third setup	59.3443

Table 30A: Forecasting Performance

Forecast sample:	Predicted value	RW value	Predicted value 2
MAPE	0.0113	0.0193	0.0101
ME	-29.0802	70.6687	-9.0602
MAE	43.1614	72.6608	38.2592
MPE	-0.0076	0.0188	-0.0023
RMSE	50.5838	110.6842	43.952

Table 31A: Forecast performance for the most-followed accounts sentiment prediction in weeks

	MAPE	ME	MAE	MPE	RMSE
Week 1	0.0067	-10.6183	25.4958	-0.0027	35.2488
Week 2	0.0106	-40.39	40.39	-0.0106	43.7763
Week 3	0.0158	-60.625	60.625	-0.0158	65.5769
Week 4	0.011	-0.1046	41.6876	0.0001	49.6292

Table 32A: Forecast performance for the random walk prediction in weeks

	MAPE	ME	MAE	MPE	RMSE
Week 1	0.0107	37.7725	39.7484	0.0101	51.4338
Week 2	0.0064	20.5021	24.1414	0.0054	30.1531
Week 3	0.0109	39.6756	41.7797	0.0104	47.6599
Week 4	0.055	206.7356	206.7356	0.055	221.9406

Table 33A Autoregressive model without independent variable & ADF tests

	# Variable lags	# Lagged First Difference	Trend Variable	Suppressed Constant	τ Statistic	5% Critical Value	1% Critical Value

Variables							
Sentiment Index	1	3	*N*	N			
ADF test for unit root					-2.52	-1.65	-2.33
S&P 500 Index	1	2	Y	*N*			
ADF test for unit root					-2.28	-3.41	-3.96
Residuals of Regressing S&P on Sentiment	1	2	N	Y			
ADF test for unit root					-1.24	-2.76	-3.4
FD of sentiment index	1	3	N	Y			
ADF test for unit root					-30.41	-1.94	-2.56
FD of S&P 500 index	1	1	N	Y			

ADF test for unit root	-22.33	-1.94	-2.56

Table 34A Post estimation results derived from Kripfganz, S., and D. C. Schneider (2018) ARDL model

	10%		5%		1%		p - value	
	I(0)	I(1)	I(0)	I(1)	I(0)	I(1)	I(0)	I(1)
F	2.43	3.28	3.14	4.10	4.79	5.96	0.00	0.00
t	−1.62	−2.27	−1.94	−2.60	−2.56	−3.23	0.00	0.00

Pesaran, Shin, and Smith (2001) bounds test

H0: no level relationship

F = 259.432

t = −22.741

Kripfganz and Schneider (2018) critical values and approximate p-values

Do not reject H0 if both F and t are closer to zero than critical values for I(0) variables

Reject H0 if both F and t are more extreme than critical values for $I(1)$ variables

Table 35A List of traders' accounts

Username	Number of Followers
paulkrugman	4.6 M
JustinWolfers	205.9 K
peterschiff	529.2 K
investorslive	173.7 K
RedDogT3	146.2 K
jimcramer	1.7 M
elerianm	396.4 K
DailyFXTeam	127.9 K
zerohedge	1 M
Trader_Dante	101.9 K
AsennaWealth	100.4 K
Rayner_Teo	120.2 K
Schuldensuehner	175 K
LizAnnSonders	189.8 K
NicTrades	101.2 K
TheStalwart	279.3 K
steve_hanke	338.9 K
SJosephBurns	376.2 K
peterlbrandt	538.7 K
ritholtz	184.7 K

Table 36A Correlation coefficients

# Lags	Traders' Sentiment and S&P 500 December 2019 till April 2020	Traders' Sentiment and S&P 500 November 2017 till April of 2018	Most-followed Sentiment and S&P 500 December 2019 till April 2020	Most-followed Sentiment and S&P 500 November 2017 till April of 2018	FD of Most-followed Sentiment and FD of S&P 500 November 2017 till April of 2018	FD of Most-followed Sentiment and FD of S&P 500 December 2019 till April 2020	FD of Traders' Sentiment and FD of S&P 500 December 2019 till April 2020	FD of Traders' Sentiment and FD of S&P 500 November 2017 till April of 2018
1	0.78	-0.01	**0.88**	-0.05	-0.01	0.11	-0.22	-0.2
2	0.81	0.04	0.88	0.01	0.02	0.08	0.03	0.02
3	**0.82**	0.09	0.87	0.06	-0.07	-0.01	0.08	0.07
4	0.82	0.1	0.86	0.14	0.13	0.01	0.05	0.05
5	0.8	0.08	0.85	0.16	-0.10	-0.04	-0.03	-0.08
6	0.8	0.1	0.84	0.23	0.04	0.18	-0.03	-0.01
7	0.8	0.12	0.82	0.29	0.06	-0.06	**0.25**	0.1
8	0.75	0.09	0.81	0.34	0.02	0.05	-0.24	-0.11
9	0.75	0.12	0.79	0.36	0.07	0.05	0.18	-0.01
10	0.71	0.17	0.76	0.35	-0.12	-0.18	-0.15	-0.09
11	0.7	0.25	0.75	0.39	0.04	0.14	0.05	0.05
12	0.68	0.31	0.73	0.42	0.02	-0.04	0.05	0.04
13	0.65	0.35	0.71	**0.43**	0.06	0.03	-0.1	0
14	0.64	0.4	0.69	0.41	0.09	-0.04	0.21	0.12
15	0.59	0.38	0.67	0.34	0.00	-0.05	-0.32	-0.29
16	0.59	0.51	0.65	0.26	-0.16	0.15	0.3	**0.28**
17	0.55	**0.51**	0.63	0.28	**0.18**	-0.24	-0.14	0.13
18	0.53	0.45	0.62	0.19	0.07	0.2	0	-0.07
19	0.51	0.41	0.6	0.06	-0.28	-0.2	0.09	-0.02

20	0.47	0.38	0.6	0.1	0.16	-0.01	-0.11	0.02
21	0.45	0.35	0.59	0.05	0.02	0.14	0.07	0.1
22	0.41	0.27	0.58	-0.01	0.01	-0.31	-0.08	0.08
23	0.39	0.17	0.6	-0.07	-0.11	0.21	0.05	-0.06
24	0.37	0.06	0.59	-0.06	-0.01	-0.26	-0.2	-0.13
25	0.37	0.02	0.62	-0.05	0.17	**0.18**	0.24	0.17
26	0.34	-0.08	0.63	-0.14	-0.03	0.04	-0.24	-0.04
27	0.35	-0.18	0.64	-0.18	-0.19	-0.07	0	-0.27
28	0.36	-0.16	0.67	-0.13	-0.10	0.07	0.05	0.15
29	0.36	-0.21	0.66	-0.02	0.33	-0.33	-0.21	0.02
30	0.4	-0.26	0.7	-0.1	-0.24	0.13	0.17	-0.14
31	0.41	-0.23	0.71	-0.05	0.06	0	-0.1	-0.13
32	0.45	-0.14	0.73	-0.02	0.09	0.05	0.1	0.12
33	0.45	-0.1	0.75	-0.07	-0.10	0.08	-0.16	0
34	0.5	-0.05	0.76	-0.05	0.21	-0.11	0.11	0.04
35	0.54	-0.02	0.77	-0.13	-0.15	0.27	0.13	0.04
36	0.53	0.01	0.73	-0.12	0.12	-0.05	0	-0.06
37	0.53	0.08	0.73	-0.21	-0.16	0.2	0.14	-0.01
38	0.5	0.15	0.69	-0.18	-0.16	-0.16	-0.13	-0.01
39	0.5	0.24	0.65	-0.06	0.12	0.06	0.13	0.16
40	0.45	0.27	0.6	0.01	-0.12	-0.04	-0.14	-0.05

Table 37A ARDL(4,1) results derived from Kripfganz, S., and D. C. Schneider (2018) ARDL model based on the traders accounts' sentiment

	D.dsp	Coefficient	Std. err.	t	P> |t|	[95% conf. interval]	
ADJ							
	dsp						
	L1.	−1.07	.06	−16.76	0.000	−1.19	−.94
LR							
	dsen	.16	.026	6.30	0.000	.11	.21
SR							

dsp						
LD.	−.12	.056	−2.08	0.037	−.23	−.01
L2D.	.038	.047	0.80	0.42	−.054	.13
L3D.	.09	.03	2.94	0.003	.029	.15
dsen						
D1.	−.05	.02	−3.22	0.001	−.08	-.012

Sample: $1/1/2017 - 16/3/2021$

Number of observations $= 1{,}052$
R-squared $= 0.64$
Adjusted R-squared $= 0.64$
$Root\ MSE = 33.87$

Table 38A: Forecast performance for the Trader's sentiment prediction in weeks

	MAPE	ME	MAE	MPE	RMSE
Week 1	0.0094	11.1004	35.1948	0.0031	38.7531
Week 2	0.007	-24.0543	26.5221	-0.0063	30.4357
Week 3	0.011	-42.1746	42.1746	-0.011	49.0703
Week 4	0.013	18.8877	49.1453	0.0052	53.7688

12.3 Excluded expressions

Tweets that included the following expressions were removed:
icymi, most read article, popular video 2016, popular video 2017, popular video 2018, popular video 2019, popular video 2020, Reuters poll, looking back 2017, looking back 2018, looking back 2019, looking back 2020, look back, one most read, follow latest, subscribe daily, print subscribe, subscribe economist, online subscribe

12.4 TF-IDF scores for November 2020

November: ('covid', 0.24125), ('president', 0.23857), ('trump', 0.21731), ('19', 0.20449), ('biden', 0.20401), ('new', 0.19951), ('rt', 0.18821), ('say', 0.17969), ('election', 0.17423), ('state', 0.16016), ('vaccine', 0.14982), ('joe', 0.14657), ('year', 0.14513), ('coronavirus', 0.14101), ('first', 0.10502)

12.5 Feature names

168 out of total 3000 features: ['000 american died', '000 coronavirus case', '000 coronavirus death', '000 covid 19', '000 first time', '000 new case', '000 new coronavirus', '000 people died', '000 square foot', '000 year ago', '000 year old', '10 000 people', '10 000 year', '10 comment week', '10 et pt', '10 year ago', '10 year old', '10 year prison', '100 000 people', '100 day office', '100 million dos', '100 year ago', '100 year old', '10p et pt', '11 year old', '12 2017 follow', '12 2017 see', '12 week access', '12 week subscription', '12 year old', '125 crore indian', '13 year old', '130 crore indian', '14 day quarantine', '14 year old', '15 year old', '16 year old', '17 year old', '18 year old', '19 case rise', '19 case surge', '19 death toll', '19 vaccine candidate', '19 vaccine dos', '19 vaccine trial', '19 year old', '1990 28 spain', '20 million people', '20 year ago', '20 year old', '20 year prison', '200 000 people', '2008 financial crisis', '2015 nuclear deal', '2016 presidential campaign', '2016 presidential election', '2016 trump tower', '2017 follow case', '2017 see coverage', '2017 see full', '2018 may bring', '2018 midterm election', '2018 world cup', '2020 democratic presidential', '2020 presidential bid', '2020 presidential campaign', '2020 presidential candidate', '2020 presidential election', '2020 presidential race', '2020 presidential run', '21 year old', '21st century fox', '22 year old', '23 year old', '24 hour coronavirus', '24 year old', '25 roundup no', '25 year ago', '25 year old', '26 year old', '27 year old', '28 spain forested', '28 year old', '29 year old', '30 year ago', '30 year old', '31 year old', '33 year old', '34 year old', '347 322 0415', '35 year old', '40 year ago', '50 year ago', '50 year old', '500 year old', '52 place go', '60 year old', '65 year old', '70 year old', '71 year old', '737 max aircraft', '737 max crash', '737 max flight', '737 max grounding', '737 max jet', '737 max plane', '75 year old', '80 year old', '90 year old', '92 year old', '93 year old', '94 year old', '9p et pt', 'abu bakr al', 'according cnn affiliate', 'according court document', 'according john hopkins', 'according national hurricane', 'according new analysis', 'according new cnn',

'according new data', 'according new poll', 'according new report', 'according new research', 'according new study', 'according newly released', 'according reuters ipsos', 'according source familiar', 'according state medium', 'accused president trump', 'accused sexual harassment', 'accused sexual misconduct', 'across middle east', 'across political spectrum', 'across united state', 'acting attorney general', 'acting chief staff', 'acting homeland security', 'acting white house', 'action climate change', 'activist greta thunberg', 'activist joshua wong', 'actor jussie smollett', 'actress lori loughlin', 'actual effect trump', 'adam schiff say', 'administration official said', 'administration official say', 'administrator scott pruitt', 'adviser jared kushner', 'adviser john bolton', 'adviser kellyanne conway', 'adviser michael flynn', 'adviser president trump', 'adviser roger stone', 'affordable care act', 'affordable health care', 'afghan capital kabul', 'africa week picture', 'african american woman', 'african swine fever', 'ag jeff session', 'ahead 2020 election', 'ahead president trump', 'air force base', 'air force one', 'air france klm', 'air new zealand']