AN6201 – Data Analytics Practicum Final Report

By

Wei Kexian (G2000380D)

Zhao Chenxi (G2000293B)

Introduction

Recently, many researchers have shown that online information obtained from the mainstream media and social media discussion such as tweets can have a significant effect on decision making by stock market investors. Particularly, in times of COVID-19 crisis, any positive or negative sentiment of public related to COVID-19 can have a ripple effect on decision making by investors in stock markets.

In this paper, we investigate the relationship between COVID-19 sentiment and stock market movement. Specifically, we wish to see if, and how well, COVID-related sentiment extracted from Twitter can help to predict future shifts in prices. To answer this question, we chose S&P Global sector indices as the research object. As it subdivides 11 sector indices of the Unites States (US) stock market, we can also analyze the relationship for different industries. For the selection of sentiment variables, we tried two data sets and multiple variables, such as sentiment score, the number of tweets, and the percentage of positive or negative sentiment. At the same time, we visualized and checked data distribution to make our analysis more accurate.

After data preprocessing, we performed Pearson Correlation Analysis on datasets, based on different lags and industries. The results show the Correlation Coefficients and corresponding p-values. Correlation however does not prove causation. We therefore use Granger causality analysis, which is a statistical hypothesis test for determining whether one time series is useful in forecasting another. By reading some relevant literatures, we learned Granger causality test is very sensitive to the choice of

time and the data volume should also be at least greater than 30. Therefore, we also tried different periods and summarized the results.

From the results of two tests above, we can select the sentiment variable that is most helpful in predicting the stock market, which is the total number of tweets. In addition, the specific industry and lag days can also be determined. Based on this information, we tried to build a predictive model to detect whether the prediction accuracy of the model with emotional indicators will be higher. Specifically, we used the autoregressive model to evaluate the gained predictive benefit of sentiment variables by comparing the accuracy of the two models.

Background

The COVID-19 virus is raging around the world, causing tragic loss of life. In order to fight it, over 100 countries had implemented some form of lockdown by the end of March, while governments ramp up spending on testing and treatment, to control the virus' spread. Social distance and lockdown make so many people to stay at home. They are devoting a considerable amount of time and effort creating and reading the messages posted on Internet media. All this attention to Internet caused us to wonder whether these texts actually effect stock prices.

For sentiment analysis, the most well-known study in this area is by Bollen and Zeng (2010). Using Twitter data, they investigated whether the collective mood states of public are correlated to the Dow Jones Industrial Index. They used a fuzzy neural network for their prediction and showed significant correlation between public mood states in twitter and the Dow Jones Industrial Index. Zhang (2013) examined correlation between stock price and significant keywords in tweets. He showed a strong negative correlation between mood states like hope, fear and worry in tweets with the Dow Jones Average Index. Lima et al. (2016) improved on the accuracy of predicting stock trends using support vector machine (SVM) by considering an overall public sentiment attribute. They showed that a day on

which the number of positive tweets exceeded the number of negative tweets is indicative of an overall positive public mood regarding the stock. Using tweeter data, Pagolu et al. (2016) also showed a strong correlation between public opinion and the Dow Jones Industrial Index.

By reading literatures, we found extensive research on sentiment analysis aims to predict stock market movement, not for causality exploration. As the ongoing COVID-19 pandemic in US was confirmed in January 2020, there is little research on the impact of COVID-19 on the US stock market. What's more, based on previous literatures and availability of data, Twitter mood seems to be the most popular and appropriate emotional indicator. Considering of these factors, we decided to investigate the cause-and-effect relationship between COVID-related twitter mood and US stock market by industry.

Statement of Problem

By investigating the links between 11 select sector indices and COVID-19 sentiment measured by sentiment score and tweet count related to coronavirus, this study provides a comprehensive overview of the relationship between COVID-19 sentiment and US stock market and how it differs by industry and time lag.

Before research, we need to figure out an important concept. That is, the difference between correlation and causality. Although correlation and causality seem deceptively similar, correlation doesn't mean causality. While causality explicitly applies to cases where action A causes outcome B.

On the other hand, correlation is simply a relationship. Action A relates to Action B—but one event doesn't necessarily cause the other event to happen.

For our research, we focus on the causality. On the one hand, we will initially explore the correlation between COVID-sentiment and stock market. Furthermore, if they have significant correlation, we will investigate their causality through Granger causality analysis.

Since our research is tentative, we will try different periods, sentiment variables and time lag in our analysis to see which works best. At last, we can select industries that have a causal relationship and corresponding time lags. By establishing the time series regression model, we can verify whether sentiment variables are truly helpful for predicting stock prices.

Method of Analysis

1. Data Collection

Initially, we found a tweet score dataset related to COVID-19 but there are some missing values. Based on the characteristics of the data set and related literature, we decided to use the average of the two days before and after to fill in the missing values to guarantee the continuity and integrality of data. What's more, we found an alternative dataset in Github which covers a longer period regarding to the count of tweet. Finally, we collected 11 sector indices provided by S&P Global over the period from 22nd of January 2020 to 18th of December 2020.

2. Data Preprocessing

We utilized python to preprocess original data and merge multiple datasets. Then we checked the distribution and skewness of data, in order to prepare data more accurate for subsequent analysis.

3. Pearson Correlation Analysis

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables (Donnelly & Abdel-Raouf, 2016). It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

We performed a two-tailed hypothesis test to determine whether the correlation coefficient between COVID Twitter sentiment and SPX500 stock price, ρ, is significantly different from 0 based on the sample data we collected. The hypothesis is shown as below:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Through statistical function library in Python, we can calculate the lagged values of correlation coefficient and the corresponding p-value by each industry.

4. Granger Causality Test

However, correlation does not prove causation. As we aim to explore the causal relationship between COVID-related twitter mood and US stock market, we therefore use Granger causality analysis, which is a statistical hypothesis test for determining whether one time series is useful in forecasting another. Similar to Pearson Correlation Analysis, the statistical function library in Python can show p-values of Granger Causality Test by each industry and time lag.

5. Time Series Regression Model

According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone. For our research, sentiment variable such as Tweet count and sentiment score is X1 and stock price is X2.

Based on the results of Granger Causality Test, we can select the most specific time lags and industries that have the best causality effect. To verify the results, we built auto regression prediction model on stock prices with and without sentiment variables. By comparing accuracy rates, we can know whether tweet count is useful for stock prediction.

Data Sources

1. Data Description

For sentiment data, we collected two datasets which contain different sentiment variables. The first one 'COV19Tweets Dataset' is generated from IEEE[1], containing IDs and sentiment scores of the tweets related to the COVID-19 pandemic from March 20 2020. The Twitter feeds are related to 90+ different keywords and hashtags that are commonly used while referencing the pandemic and are monitored in real time. The sentiment scores are defined in the range [-1,+ 1]. If a score falls between (0,+ 1], the tweet is considered to have a Positive sentiment. Similarly, a score in the range [-1, 0) represents a Negative sentiment. And the score "0" denotes a Neutral sentiment. Scores in the extremes of the range [-1,+ 1] represent strongly Negative sentiment and strongly Positive sentiment, respectively.

Another sentiment dataset is from Github[2], covering a period from January 22, 2020 to December 27, 2020. However, this dataset only counts daily coronavirus-related tweets and distinguishes positive, negative, and neutral sentiment.

For stock market data, this study uses 11 select sector indices provided by S&P Global[3]. Furthermore, we decided to download stock data in the longest period to cover COVID-19 crisis and conduct follow-up analysis.

2. Data Limitations

Since both sentiment datasets do not show the specific content of tweets, we cannot check the accuracy of sentiment scores or classification of word emotions. In addition, our research object is US stock market, but twitter data covers the whole world. Due to the Twitter protection policy, we cannot obtain Tweet information only from the United States, which may have impact on our research accuracy.

[1] https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset
[2] https://github.com/lopezbec/COVID19_Tweets_Dataset/tree/master/Summary_Hashtag/2020_01
[3] https://us.spindices.com/index-finder/

The sentiment scores of several days in the IEEE dataset are missing due to technical factors. In order to ensure the integrity of the data, we can only use the average method to fill the missing values, but we are not sure whether this approach affects our research results.

For Github dataset, we are not clear the algorithm of sentiment analysis to distinguish positive, negative, and neutral sentiment and thus if it is suitable for our research.

Technical Analysis

Because this project is to check whether COVID Twitter sentiment is useful for analyse stock market price, as mentioned in the previous sections, in order to eliminate the selection bias, we tried to find out various ways and datasets to prove whether this hypothesis is correct. We used the following three research methods, for each one, we experimented with both two datasets.

1. Pearson Correlation Test

Here we chose Python as analytical tool, in addition to do the pure Pearson Correlation Test, referring to the previous research, we also took the time lag into consideration. After doing the two-tailed test, the results clearly show that Twitter sentiment including sentiment score and tweet volume is correlated with S&P 500 stock price, and most of the p-values are less than 0.01, which means we have enough evidence to reject the null hypothesis.

1.1 Sentiment Score

Table 1: The coefficient of sentiment score and stock prices from Mar.23, 2020 to Jan.8, 2021

Sentiment Score	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag=0	0.263	0.210	0.272	0.153*	0.239	0.319	0.250	0.212	0.238	0.230	0.303	0.249
lag=1	0.259	0.201	0.256	0.134**	0.221	0.303	0.239	0.205	0.227	0.203	0.249	0.237
lag=2	0.270	0.202	0.255	0.120**	0.210	0.295	0.231	0.214	0.227	0.158	0.227	0.238
lag=5	0.296	0.225	0.255	0.167*	0.265	0.300	0.258	0.242	0.254	0.138	0.180	0.265

*: significant at 5% level
**: significant at 10% level

The test results shown in Table 1 clearly demonstrate that the sentiment score is positively related to all sectors indices, regardless of time lag. The positive coefficient represents that the stock prices are more likely to increase when people come across more positive comments on Twitter. Most of the correlation coefficients concentrate on 0.2 to 0.3. Here health care sector shows the highest correlation for all the time lags, which means the US stock market in health care sector tend to be more sensitive to social media sentiment. Among all the industries, energy sector's correlation with Twitter sentiment is the lowest before lag2 at about 0.14.

Different sectors' stock indices show totally different reaction according to time lag. Seven industries as well S&P500 have higher correlation with Twitter sentiment score at lag5 when comparing to those of time lag less than 5. Therefore, the Twitter sentiment score does not have an immediate impact on the day's investment. However, as to three sectors including utilities, consumer staples and real estate, with the increase of time lag, the correlation decreases significantly, which means these three sectors' reaction to sentiment score is immediate.

2.1 Twitter Volume

We did the Pearson Correlation Test of Twitter Volume with stock price for two datasets. Besides, we analyzed our datasets from three dimensions. Firstly, we divided Twitter Volume into three categories, which are total tweets, negative tweets and positive tweets. Secondly, we counted the number of tweets and calculated the percentage of each category respectively. Thirdly, taking the short-term effect of news into consideration, we chose March when US declared a state of emergency as a time slot to focus on. Here in this section, we will only choose partial result to explain, for more results, please refer to Appendix.

2.1.2 Long-Term

Table 2: The coefficient of total twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

total count	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.54	0.66	0.35	-0.43	-0.13	0.53	0.23	0.66	0.53	-0.1*	-0.21	0.47
lag 1	0.55	0.7	0.37	-0.42	-0.11*	0.55	0.25	0.67	0.54	-0.084*	-0.177	0.49
lag 2	0.56	0.68	0.39	-0.4	-0.1*	0.6	0.27	0.68	0.55	-0.07*	-0.16	0.5
lag 5	0.61	0.71	0.44	-0.37	-0.03*	0.61	0.33	0.71	0.59	-0.013*	-0.1*	0.55

*: Not significant at 5% level

Compared to sentiment score, twitter volume has stronger correlation with stock price, which means the US stock market tend to be more sensitive to tweets amount as well as the discussion popularity in social media. Without regard to insignificant industries for all time lags, among rest 10 sectors, half of them have high coefficient more than 0.5. Financials and Utilities show negative correlation with tweets volume, which means investors are less likely to invest in these two sectors when there exists more discussion about COVID in Twitter.

S&P500 as well as all sectors except for Utilities show a higher level of correlation across the time lags. Therefore, there is a time lag effect for impact of Twitter volume on the day's investment.

2.1.2 Short-Term

In March 2020, the COVID begain to get serious in USA, and the country declared a state of emergency in the middle of March and adopted lockdown policies in regions gradually (BBC NEWS, 2020). A state of emergency was declared in the United States after the outbreak began in March. Twitter discussion started to explode during that period of time, considering the requirements of data amount and comparison, we chose the time period from February 18 to March 23.

Table 3: The coefficient of positive twitter percentage and stock prices from Feb.18, 2020 to Mar.23, 2020

Positive Percentage	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	-0.50	-0.53	-0.47	-0.52	-0.53	-0.49	-0.54	-0.47	-0.52	-0.56	-0.53	-0.51
lag 1	-0.50	-0.50	-0.40	-0.51	-0.51	-0.46	-0.51	-0.48	-0.48	-0.50	-0.46	-0.49
lag 2	-0.41	-0.43	-0.32*	-0.47	-0.46	-0.37	-0.45	-0.40	-0.42	-0.44	-0.39	-0.42
lag 5	-0.45	-0.43	-0.49	-0.49	-0.46	-0.49	-0.47	-0.48	-0.50	-0.47	-0.47	-0.48

*: Not significant at 10% level

Compared to long term correlation, it is obvious to see that the direction and strength of short-term correlation are more clear and strong. From the table of correlation between positive tweets percentage and stock price, the result implies that the percentage of positive tweets shows the negative correlation with stock market. In addition, this result implies that the percentage of positive tweets has higher impact on investment in stock market on the same day than a few days later.

Table 4: The coefficient of negative twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

Negative Percentage	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.37	0.37	0.37*	0.34*	0.34	0.33*	0.35	0.34	0.35	0.35	0.34	0.35
lag 1	0.37	0.42	0.23*	0.38	0.35	0.32*	0.38	0.35	0.37	0.36	0.28*	0.36
lag 2	0.36	0.41	0.25*	0.40	0.36	0.35*	0.38	0.38	0.37	0.39	0.31*	0.37
lag 5	*	*	*	*	*	*	*	*	*	*	*	*

*: Not significant at 10% level

The results above show a moderate level of positive correlation between the percentage of negative tweets and stock price. The percentage of negative tweets has higher impact on the investment in consumer discretionary sector across all time lags.

2. Granger Causality Test

Granger causality Analysis tests if a variable X causes Y then changes in X will systematically occur before changes in Y (Bollen et al, 2010). Based on the results of our Granger Causality, we can reject the null hypothesis that sentiment score does not predict stock price. However, as to the twitter volumne, some sectors show significant p-values.

Table 5: The p-value of total twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

total count												
lag	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
1	0.023	0.0167	0.0407	0.657	0.1811	0.0085	0.0406	0.0318	0.0266	0.7449	0.4108	0.0243
2	0.0363	0.0115	0.0625	0.6367	0.1716	0.0202	0.0357	0.0323	0.0243	0.6726	0.3591	0.0201
3	0.0908	0.0286	0.2012	0.7205	0.5138	0.0589	0.1228	0.0709	0.0731	0.9255	0.5104	0.0829

It is observed that tweets volume has more significant Granger causality relation with communication services, consumer discretionary, health care, industrials, information technology, materials and SP500 for lags ranging from 1 to 2 days compared to 3 days.

According to the previous research, Granger Causality Test requires relatively huge data quantity, and the larger the sample size, the more precise the estimates. Therefore, even though we did the test for shorter period, the p-value turns out insignificant.

3. Time Series Regression Model

As to the time series regression model, in order to find the appropriate independent variables, we firstly referred to previous studies to learn some reliable models and factors, then did the correlation test of various factors with stock price to find the variables which has higher correlation. We finally choose the exact value of sentiment score and twitter volume rather than their difference because the correlation test result shows the difference of sentiment variables is not related with stock price (p-value is not significant).

Table 6: The coefficient and p-value of sentiment score and twitter volume for predicting stock prices from Mar.23, 2020 to Jan.8, 2021

	lag1				lag2				lag3				R^2	Accuracy with sentiment	Accuracy without sentiment	
	Sentiment Score		Twitter count		Sentiment Score		Twitter count		Sentiment Score		Twitter count					
	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value				
Communication Services	5.73074614	0.783367375	-1.25785E-06	0.03053458*	14.9543724	0.493417665	5.34695E-07	0.4521015	-11.655288	0.56681019	7.92856E-07	0.18341292	0.97984924	98.59%	98.56%	0.03%
Consumer Discretionary	19.9672214	0.868210744	-5.72772E-06	0.09044352*	15.4631498	0.902983024	2.55631E-06	0.53525048	-107.37501	0.36068543	4.36716E-06	0.20866001	0.98790039	98.59%	98.57%	0.02%
Consumer Staples	-4.0876469	0.934818148	-1.48079E-06	0.28808919	-6.820969	0.896949774	3.82029E-07	0.82243468	-19.045884	0.69766438	1.21925E-06	0.38851105	0.97107178	98.98%	98.97%	0.01%
Energy	19.9654318	0.749887493	-1.00831E-07	0.95358118	-43.198914	0.514063864	1.17421E-06	0.58010096	13.2228791	0.82778542	-1.51118E-06	0.38646312	0.91580161	96.96%	96.94%	0.01%
Financials	34.17936	0.594425057	-9.0344E-07	0.60809013	-71.999652	0.286187231	1.47067E-06	0.4967194	-7.8464309	0.90045947	-1.86431E-07	0.91729285	0.87604372	97.96%	97.95%	0.01%
Health Care	53.1491496	0.649749973	-4.89833E-06	0.13322262	-107.06465	0.382045555	3.49643E-06	0.37659523	15.0953328	0.89588783	1.46414E-06	0.654703	0.95118362	98.75%	98.74%	0.01%
Industrials	75.9567293	0.357419372	5.49892E-07	0.80684911	-97.368844	0.259746127	-4.28576E-07	0.876379	-19.392712	0.8084993	9.38814E-07	0.68302511	0.98142679	98.31%	98.30%	0.01%
Information Technology	43.4538001	0.857277435	-1.49053E-05	0.02886863*	257.472362	0.312508316	1.0292E-05	0.21629694	-254.38703	0.28167042	7.18775E-06	0.30401998	0.98233167	98.35%	98.31%	0.04%
Materials	18.5650503	0.707832101	5.82993E-07	0.67093501	-21.176974	0.684345591	-3.85685E-07	0.81841747	-25.658861	0.59641495	2.4018E-07	0.86441775	0.98202721	98.31%	98.30%	0.01%
Real Estate	19.378941	0.496956827	-1.73748E-07	0.82729615	-25.533783	0.397655362	1.72459E-08	0.98583696	-29.092878	0.29965416	9.28476E-07	0.24835502	0.86929695	98.28%	98.25%	0.03%
Utilities	-30.786874	0.418745554	2.76335E-07	0.79341853	-1.7508039	0.965417834	-9.80231E-08	0.93919606	-19.955589	0.59568781	-5.23554E-08	0.96056444	0.88640155	98.37%	98.36%	0.01%
S&P 500	143.799032	0.660212382	-1.4389E-05	0.11342904	-76.121605	0.824547497	8.76183E-06	0.43078581	-238.3337	0.45609569	7.13443E-06	0.44388994	0.98090939	98.69%	98.68%	0.01%

The above table reports estimated coefficients and p-values for lag1, lag2, and lag 3 changes in sentiment score and twitter volume from the time series regression model. We did prediction twice, one with sentiment and another without sentiment. From the table we can see even though almost none of

the p-values are significant, the prediction accuracies for every sector with sentiment variables are higher than that without sentiment.

Summary

From the whole technical analysis, we achieved the following three conclusions. Firstly, in terms of Pearson correlation test, both sentiment score and twitter volume are significantly related to S&P500 price, while the twitter volume seems has stronger relationship with stock price. Probably because important event will arouse more discussion in Twitter and trigger action in stock market, we cannot say twitter sentiment has cause and effect relationship with S&P500, however, according to the correlation test, it can be concluded they do have correlation with each other, which means it is effective to use twitter sentiment for stock prediction. Secondly, the effect of Twitter sentiment on stock market comes with the time lag effect because the correlation of longer lags tends to be stronger than that of shorter lags. Thirdly, the correlation of Twitter sentiment and stock market will be stronger for shorter period compared to longer period. The proportion of positive tweets is negatively correlated with stock price while the proportion of negative tweets is positively correlated with stock market. Fourthly, twitter sentiment can effectively help to predict S&P500 price. From the perspective of Granger Causality Test, the results show twitter volume has predictive information about the S&P500. As to the time series regression model, the prediction accuracies improve with the usage of twitter sentiment.

Recommendations

This study examines whether the COVID social media sentiment is related with stock market and whether it can contribute to S&P500 price prediction. While we focus on the sentiment in Twitter, we encourage future studies to explore additional social media platforms such as Facebook, Weibo and so on. Secondly, it might also be worthy to note different hashtags might lead to totally different results.

In our study, in order to mitigate the selection bias, we chose two different datasets, which scraped tweets with different keywords respectively, even we use the same dimension such as twitter volume, the results are different. Thirdly, the sentiment library can be better developed. The sentiment score highly depends on what NLP library is used. Our study uses the NLP library special for social media, however, the performance might be limited because the event our study focus on is the COVID. For a particular event, the word must have some patter, different from common comments in the social media. We believe a specialized library will improve the performance. Furthermore, as to the prediction part, our project only uses regression model with limited factors, it is worthy of noting there are many time series models such as ARIMA, XGBoost and so on. Therefore, future work can use more models to prove whether sentiment can predict S&P500 price.

Reference List

BBC NEWS. (2020) Earlier coronavirus lockdown 'could have saved 36,000 lives'. Available at: https://www.bbc.com/news/world-us-canada-52757150

BBC NEWS. (2020) Trump declares national emergency over coronavirus. Available at: https://www.bbc.com/news/world-us-canada-51882381

Bollen, J., Mao, H. & Zeng, X. (2010) Twitter mood predicts the stock market. *Journal of Computational Science*, **2**(1), 1-8. Available at: https://doi.org/10.1016/j.jocs.2010.12.007

Donnelly, B. & Abdel-Raouf, F. (2016) Statistics, Vol. 3: Dorling Kindersley Limited. Indianapolis.

Duan, Yuejiao and Liu, Lanbiao and Wang, Zhuo, COVID-19 Sentiment and Chinese Stock Market: Official Media News and Sina Weibo (June 6, 2020). Available at: https://ssrn.com/abstract=3639123

Lee, H.S. Exploring the Initial Impact of COVID-19 Sentiment on US Stock Market Using Big Data. Sustainability 2020, 12, 6648. Available at: https://doi.org/10.3390/su12166648

Lima, M.L.; Nascimento, T.P.; Labidi, S.; Timbo, N.S.; Batista, M.V.L.; Neto, G.N.; Costa, E.A.M.; Sousa, S.R.S. Using sentiment analysis for stock exchange prediction. Int. J. Artif. Intell. Appl. 2016, 7, 59–67. Available at: https://scholar.google.com.sg/scholar

Pagolu, V.S.; Reddy, K.N.; Panda, G.; Majhi, B. Sentiment analysis of twitter data for predicting stock market movements. In Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3–5 October 2016; pp. 1345–1350. Available at: https://arxiv.org/pdf/1610.09225.pdf

Sprenger, T.O.; Tumasjan, A.; Sandner, P.G.; Welpe, I.M. Tweets and trades: The information content of stock microblogs. Eur. Financ. Manag. 2014, 20, 926–957. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x

Zhang, L. Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation. Ph.D. Thesis, The University of Texas at Austin, Austin, TX, USA, 2013. Available at: https://repositories.lib.utexas.edu/handle/2152/20057

Appendix

1. Pearson Correlation Test

Table I: The coefficient of sentiment score and stock prices from Mar.23, 2020 to Jan.8, 2021

Sentiment Score	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag=0	0.263	0.210	0.272	0.153*	0.239	0.319	0.250	0.212	0.238	0.230	0.303	0.249
lag=1	0.259	0.201	0.256	0.134**	0.221	0.303	0.239	0.205	0.227	0.203	0.249	0.237
lag=2	0.270	0.202	0.255	0.120**	0.210	0.295	0.231	0.214	0.227	0.158	0.227	0.238
lag=5	0.296	0.225	0.255	0.167*	0.265	0.300	0.258	0.242	0.254	0.138	0.180	0.265

*: significant at 5% level
**: significant at 10% level

Table II: The coefficient of twitter volume and stock prices from Mar.23, 2020 to Jan.8, 2021

Twitter Count	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag=0	0.650	0.687	0.514	0.290	0.463	0.610	0.536	0.666	0.604	0.614	0.349	0.638
lag=1	0.646	0.689	0.510	0.267	0.464	0.605	0.541	0.666	0.611	0.608	0.336	0.638
lag=2	0.650	0.697	0.510	0.250	0.474	0.605	0.550	0.676	0.620	0.604	0.325	0.645
lag=5	0.642	0.701	0.503	0.167*	0.470	0.567	0.558	0.678	0.623	0.600	0.315	0.641

*: significant at 5% level

Table III: The coefficient of total twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

total count	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.54	0.66	0.35	-0.43	-0.13	0.53	0.23	0.66	0.53	-0.1*	-0.21	0.47
lag 1	0.55	0.7	0.37	-0.42	-0.11*	0.55	0.25	0.67	0.54	-0.084*	-0.177	0.49
lag 2	0.56	0.68	0.39	-0.4	-0.1*	0.6	0.27	0.68	0.55	-0.07*	-0.16	0.5
lag 5	0.61	0.71	0.44	-0.37	-0.03*	0.61	0.33	0.71	0.59	-0.013*	-0.1*	0.55

*: Not significant at 5% level

Table IV: The coefficient of positive twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

positive count	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.6	0.7	0.43	-0.44	-0.044*	0.56	0.33	0.71	0.6	-0.06*	-0.13*	0.53
lag 1	0.62	0.71	0.45	-0.43	-0.023*	0.58	0.35	0.72	0.61	-0.057*	-0.106*	0.55
lag 2	0.63	0.72	0.46	-0.41	-0.003*	0.59	0.37	0.73	0.62	-0.038*	-0.094*	0.56
lag 5	0.67	0.74	0.51	-0.38	0.05*	0.64	0.42	0.75	0.65	0.009*	-0.044*	0.6

*: Not significant at 5% level

Table V: The coefficient of negative twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

negative count	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.53	0.65	0.33	-0.42	-0.14*	0.51	0.22	0.65	0.51	-0.09*	-0.21	0.46
lag 1	0.54	0.66	0.35	-0.4	-0.115*	0.53	0.235	0.656	0.52	-0.076*	-0.178	0.47
lag 2	0.55	0.7	0.37	-0.39	-0.094*	0.55	0.26	0.67	0.54	-0.054*	-0.16	0.49
lag 5	0.59	0.7	0.43	-0.35	-0.03*	0.59	0.32	0.7	0.58	-0.0002*	-0.09*	0.54

*: Not significant at 5% level

Table VI: The coefficient of positive twitter percentage and stock prices from Feb.18, 2020 to Mar.23, 2020

Positive Percentage	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	-0.50	-0.53	-0.47	-0.52	-0.53	-0.49	-0.54	-0.47	-0.52	-0.56	-0.53	-0.51
lag 1	-0.50	-0.50	-0.40	-0.51	-0.51	-0.46	-0.51	-0.48	-0.48	-0.50	-0.46	-0.49
lag 2	-0.41	-0.43	-0.32*	-0.47	-0.46	-0.37	-0.45	-0.40	-0.42	-0.44	-0.39	-0.42
lag 5	-0.45	-0.43	-0.49	-0.49	-0.46	-0.49	-0.47	-0.48	-0.50	-0.47	-0.47	-0.48

*: Not significant at 10% level

Table VII: The coefficient of negative twitter percentage and stock prices from Feb.18, 2020 to Mar.23, 2020

Negative Percentage	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
lag 0	0.37	0.37	0.37*	0.34*	0.34	0.33*	0.35	0.34	0.35	0.35	0.34	0.35
lag 1	0.37	0.42	0.23*	0.38	0.35	0.32*	0.38	0.35	0.37	0.36	0.28*	0.36
lag 2	0.36	0.41	0.25*	0.40	0.36	0.35*	0.38	0.38	0.37	0.39	0.31*	0.37
lag 5	*	*	*	*	*	*	*	*	*	*	*	*

*: Not significant at 10% level

2. Granger Causality Test

Table VIII: The p-value of total twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

	total count											
lag	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
1	0.023	0.0167	0.0407	0.657	0.1811	0.0085	0.0406	0.0318	0.0266	0.7449	0.4108	0.0243
2	0.0363	0.0115	0.0625	0.6367	0.1716	0.0202	0.0357	0.0323	0.0243	0.6726	0.3591	0.0201
3	0.0908	0.0286	0.2012	0.7205	0.5138	0.0589	0.1228	0.0709	0.0731	0.9255	0.5104	0.0829

Table IX: The p-value of positive twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

positive count												
lag	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
1	0.0129	0.0258	0.0266	0.5895	0.1372	0.0047	0.0269	0.0257	0.0281	0.6943	0.4761	0.0182
2	0.0271	0.0391	0.0461	0.5774	0.121	0.012	0.0317	0.0395	0.0291	0.672	0.5033	0.019
3	0.0734	0.0546	0.1412	0.7225	0.4199	0.0034	0.1053	0.0589	0.0723	0.9276	0.6886	0.0695

Table X: The p-value of negative twitter volume and stock prices from Jan.22, 2020 to Dec.19, 2020

negative count												
lag	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	S&P 500
1	0.035	0.0238	0.0544	0.6587	0.1842	0.0128	0.042	0.0537	0.0254	0.7366	0.3978	0.0323
2	0.047	0.011	0.077	0.5542	0.1689	0.0252	0.0279	0.0447	0.0183	0.5534	0.4004	0.0218
3	0.0961	0.0253	0.2325	0.7238	0.513	0.0706	0.1011	0.0884	0.0547	0.8101	0.6259	0.0815

Time Series Regression Model

Table XI: The coefficient and p-value of sentiment score and twitter volume for predicting stock prices from Mar.23, 2020 to Jan.8, 2021

	lag1				lag2				lag3				R^2	Accuracy with sentiment	Accuracy without sentiment	
	Sentiment Score		Twitter count		Sentiment Score		Twitter count		Sentiment Score		Twitter count					
	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value				
Communication Services	5.73074614	0.783367375	-1.25785E-06	0.03053458*	14.9543724	0.493417665	5.34695E-07	0.4521015	-11.655288	0.56681019	7.92856E-07	0.18341292	0.97984924	98.59%	98.56%	0.03%
Consumer Discretionary	19.9672214	0.868210744	-5.72772E-06	0.09044352*	15.4631498	0.902983024	2.55631E-06	0.53525048	-107.37501	0.36068543	4.36716E-06	0.20866001	0.98790039	98.59%	98.57%	0.02%
Consumer Staples	-4.0876469	0.934818148	-1.48079E-06	0.28808919	-6.820969	0.896949774	3.82029E-07	0.82243468	-19.045884	0.69766438	1.21925E-06	0.38851105	0.97107178	98.98%	98.97%	0.01%
Energy	19.9654318	0.749887493	-1.00831E-07	0.95358118	-43.198914	0.514063864	1.17421E-06	0.58010096	13.2228791	0.82778542	-1.51118E-06	0.38646312	0.91580161	96.96%	96.94%	0.01%
Financials	34.17936	0.594425057	-9.0344E-07	0.60809013	-71.999652	0.286187231	1.47067E-06	0.4967194	-7.8464309	0.90045947	-1.86431E-07	0.91729285	0.87604372	97.96%	97.95%	0.01%
Health Care	53.1491496	0.649749973	-4.89833E-06	0.13322262	-107.06465	0.382045555	3.49643E-06	0.37659523	15.0953328	0.89588783	1.46414E-06	0.654703	0.95118362	98.75%	98.74%	0.01%
Industrials	75.9567293	0.357419372	5.49892E-07	0.80684911	-97.368844	0.259746127	-4.28576E-07	0.876379	-19.392712	0.8084993	9.38814E-07	0.68302511	0.98142679	98.31%	98.30%	0.01%
Information Technology	43.4538001	0.857277435	-1.49053E-05	0.02886863*	257.472362	0.312508316	1.0292E-05	0.21629694	-254.38703	0.28167042	7.18775E-06	0.30401998	0.98233167	98.35%	98.31%	0.04%
Materials	18.5650503	0.707832101	5.82993E-07	0.67093501	-21.176974	0.684345591	-3.85685E-07	0.81841747	-25.658861	0.59641495	2.4018E-07	0.86441775	0.98202721	98.31%	98.30%	0.01%
Real Estate	19.378941	0.496956827	-1.73748E-07	0.82729615	-25.533783	0.397655362	1.72459E-08	0.98583696	-29.092878	0.29965416	9.28476E-07	0.24835502	0.86929695	98.28%	98.25%	0.03%
Utilities	-30.786874	0.418745554	2.76335E-07	0.79341853	-1.7508039	0.965417834	-9.80231E-08	0.93919606	-19.955589	0.59568781	-5.23554E-08	0.96056444	0.88640155	98.37%	98.36%	0.01%
S&P 500	143.799032	0.660212382	-1.4389E-05	0.11342904	-76.121605	0.824547497	8.76183E-06	0.43078581	-238.3337	0.45609569	7.13443E-06	0.44388994	0.98090939	98.69%	98.68%	0.01%

Table XII: The accuracy of total twitter volume for predicting stock prices from Jan.22, 2020 to Dec.19, 2020

	accuracy without sentiment	accuracy with sentiment	
Communication Services	98.66%	98.68%	0.03%
Consumer Discretionary	98.77%	98.72%	-0.05%
Health Care	98.76%	98.78%	0.02%
industrials	98.49%	98.60%	0.11%
Information Technology	98.54%	98.49%	-0.04%
Materials	98.49%	98.52%	0.02%
S&P 500	98.88%	98.90%	0.02%