COVID-19 Sentiment and Daily Cases Analysis on Social Media

Jiawei Zheng, Zhou Li, Yunlin Tang

Abstract

With the unexpected impact of Covid-19, drastic changes were induced to people's health, lifestyle, and mentality. During the research last quarter, we noticed that the majority of posts in our Twitter dataset have strong emotions and sentiments. In this project, we trained our SVC tweet sentiment model using a dataset that contains 1.6 million data with text and sentiment labels from Kaggle. The trained model is used to predict sentiment scores on the daily tweets sampled from the Panacea Lab dataset. After that, we detrended the daily case data and performed multiple analyses including correlation, cointegration test, and Fourier transformation to study its relationship with the sentiment score.

1 Introduction

Covid-19 changed everyone, from the way we interact, to how we work, and our methods of communication, especially through social media. During this pandemic period, social media becomes a huge and important part of people's daily lives. It provides mobile users a convenient way to connect to each other around the world and acquire updated and trending information about the topic of Covid-19. Besides, people can also express their thoughts and feelings toward certain topics by posting on social media. Throughout the studying of this quarter, we noticed that there are a number of posts in our Twitter dataset that are related to the topic of Covid-19having some strong emotions and sentiments. In the meantime, a previous study[1] has shown that more people are experiencing negative emotions such as anxiety and panic under this pandemic period. Therefore, we are interested in analyzing the posts that are related to the topic of Covid-19 on social media and investigating the emotions of the results implied in these posts will lead to.

We start our investigation using the "Covid-19 tweets" dataset obtained from the Panacea Lab[3] by performing sentiment analysis on the tweet text. Sentiment analysis and opinion mining are useful in the sense that it contributes to the understanding of human emotions by observing people's engagement in social platforms. Using social media, we are able to monitor the user's feed with sentiment analysis. For the purpose of this project, we expect that the results can answer the potential investigating question: "How is the trend of daily sentiment related to the change in the number of daily Covid-19 cases?". The motivation behind this question is that Tweet sentiments can be analyzed in real-time with relatively minor effort, but Covid-19 case data requires huge amounts of human and economic resources to obtain. Studying the correlation between daily cases and sentiment features can provide useful insights into Covid-19's impact on social media.

2 Datasets

2.1 Data Collection

There are three datasets obtained for this project. First, we used the dataset which includes the daily Tweets IDs which can regenerate tweets about the Covid-19 from March 22 to November 30 (inclusive) in the year 2020, collected by the Panacea Lab at Georgia State. Then we sampled at a rate of one out of 360 tweet IDs per day for convenience purposes. On this subsampled dataset, we performed the Twitter collection process by using the Twitter API function "twarc" to rehydrate which requests the full tweet content based on the tweet IDs, and then got all the tweets about the Covid-19.

After obtaining all the tweets about Covid-19 in the subsampled dataset, a training dataset that includes similar tweet contents with sentiment labels will be required in order to build the prediction models for sentiment analysis in the Covid-19 tweets dataset. We then found a dataset that contains 1.6 million training data, and each row in the dataset contains the text of a tweet and a sentiment label, which the text variable can be extracted as the feature and sentiment label as the output result to make predictions to sentiment.

In order to observe and determine the significant correlation between the sentiment in the Covid-19 related social media posts and the numbers of disease daily cases, we obtained a dataset that contains daily new positive cases and death cases all over the world, which is from *Our World in Data*[4].

2.2 Data Processing and Cleaning

In the tweets content dataset which is about Covid-19, we extracted "tweet_id", "text", "location", "retweeted_status", "hashtag", "follower_count", "date", and "language" from the raw dataset, and we also adjusted all columns to a suitable format and saved them as csv files. In addition, in order to have a cleaner version of the tweet text, we have converted all the text into lowercase and removed all the punctuations, stopwords, and usernames contained in the tweets.

In the trained sentiment dataset, we only extracted "text" and "sentiment" from the original dataset, and we also made text lowercase, which can avoid standardizing the same words in different formats. Furthermore, we used "-1" to represent "negative" sentiment and "1" for "positive", which is easier for us to calculate the total sentiment score during the 14-day period. In the daily cases dataset, we extracted cases and dates and removed all other unrelated columns. In addition, in the Covid-19 daily cases dataset, since it is relatively difficult to specify the region that the tweet users come from, we then summed up all the numbers of new cases around the world for 200 countries per day.

	Number of Observations	Average Text Length (after cleaning)	Median Text Length (after cleaning)	Average Follower Counts
Stats	1,007,496	87.639	85.0	19146.348

Table 1: Statistics of Tweet Contents Dataset

	Number of Observations	Average Text Length (after cleaning)	Median Text Length (after cleaning)	Counts of Positive Sentiment
Stats	1,600,000	74.090	69.0	800,000

Table 2: Statistics of Trained Sentiment Dataset

	Number of Observations	Average Daily New Cases	Std of Daily New Cases	Median of Daily New Cases
Stats	254	496,645.024	322,896.980	457,871

Table 3: Statistics of Daily New Cases Dataset

3 Data Analysis

3.1 Text Analysis

After cleaning the text in the Twitter dataset, we have performed an exploratory data analysis on it. By calculating the term frequency and Tf-Idf throughout these Twitter posts, the tables of frequencies were acquired respectively. We noticed that these two vectorizers gave out similar results; for example, the three most frequent terms in both tables are “covid19”, “coronavirus”, and “trump”. In order to visually compare the results, a graph of the word cloud for both vectorizers was generated as shown in Figures 1 and 2 below.

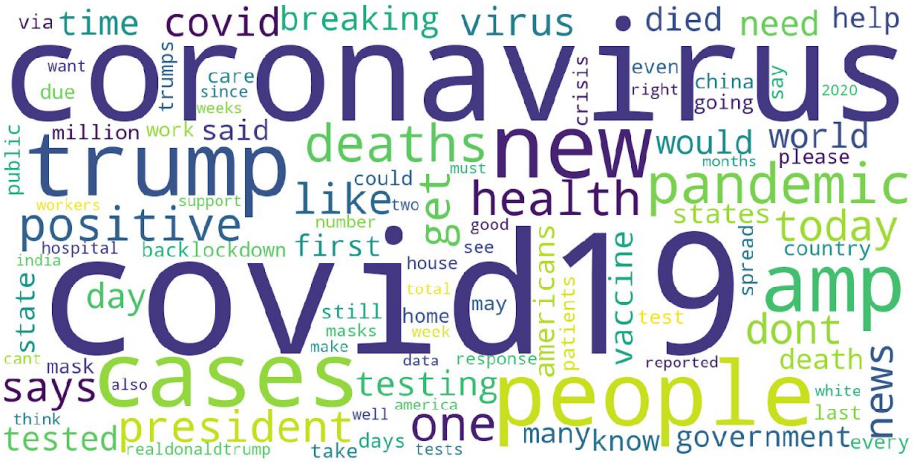

Figure 1: word cloud by using CountVectorizer

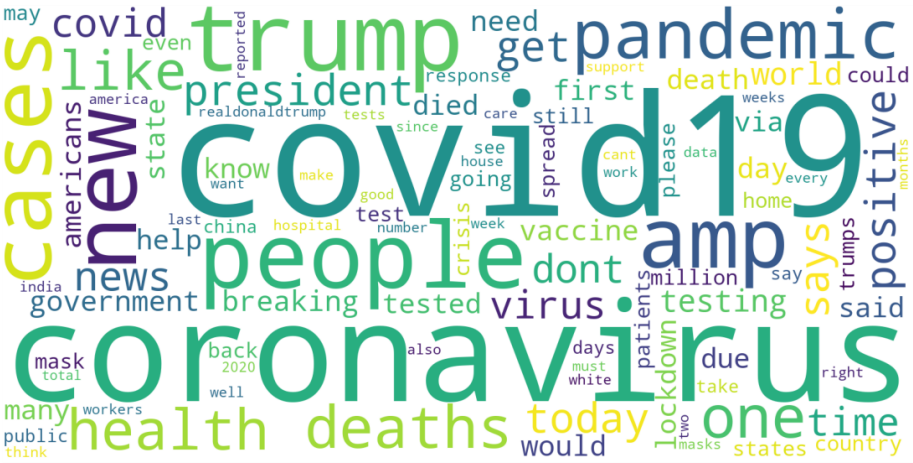

Figure 2: word cloud by using TfIdfVectorizer

To compare the daily term frequencies and the counts of daily Covid-19 cases, we tried to visualize the difference between trends by drawing the frequencies of specific terms by dates overlaid the plot of Covid-19 case numbers. After intuitively preselecting the words “great” and “sick” (two words that represent positive and negative sentiment), two graphs are generated by plotting the normalized counts of terms per day and daily case numbers from March 22 to November 30. As shown in Figures 3 and 4, we observed that there are no direct or obvious correlations between the trend of selected terms and the count of daily new cases.

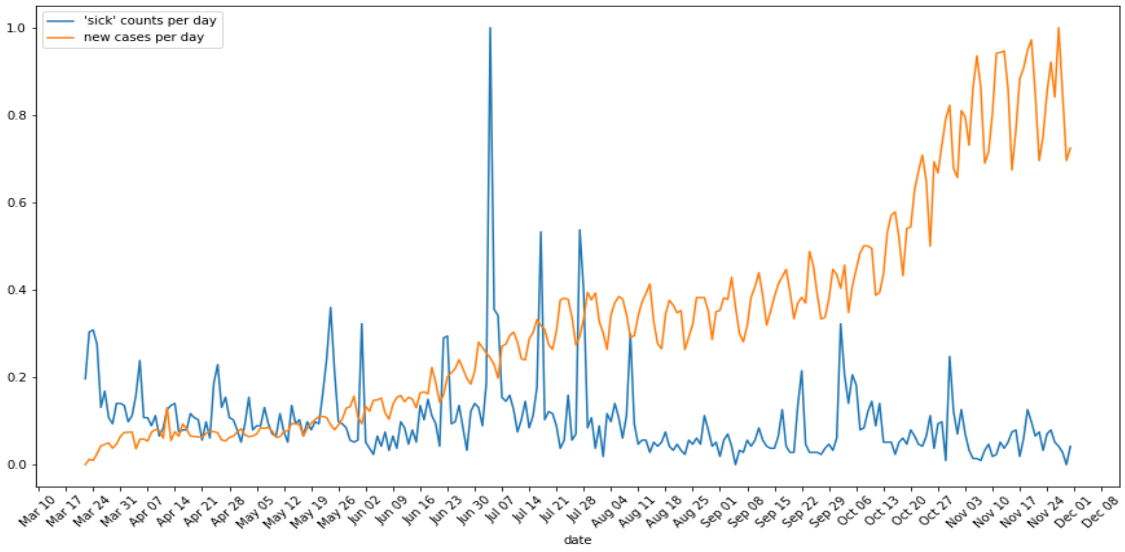

Figure 3: plot of term frequencies ("sick") overlaid by counts of new cases

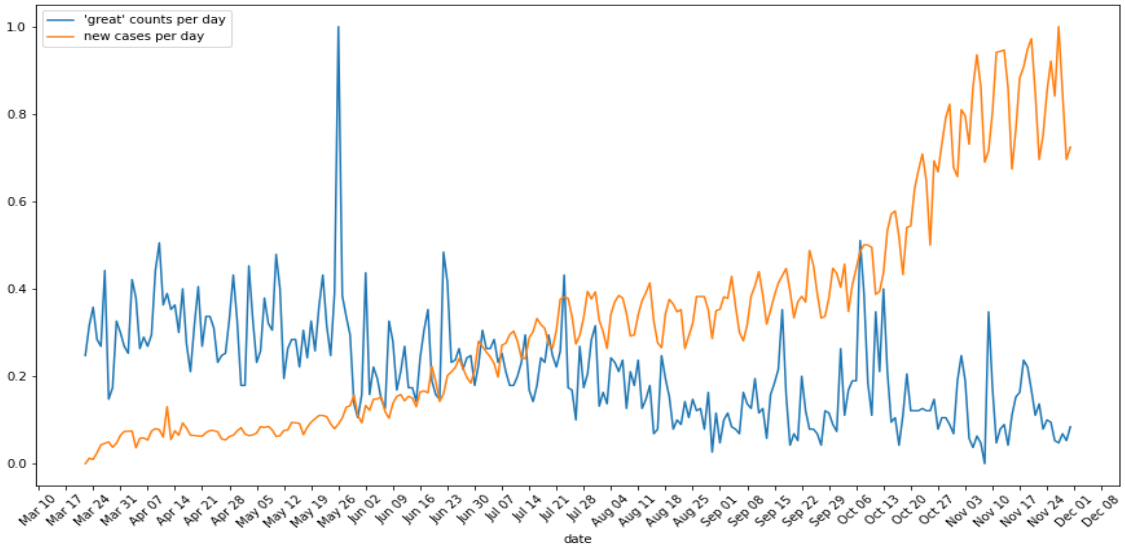

Figure 4: plot of term frequencies ("great") overlaid by counts of new cases

4 Methodology and Result

4.1 Baseline Model: BERT Tokenizer & Logistic regression

Before building models, we need to tokenize all text data into smaller units, so we are going to use BERT tokenizer to convert the whole text into numerical arrays.[1] By importing the transformers package, we used the "tokenizer" function to convert the text to arrays of numbers and then pad different arrays to the max length and convert it to the parse matrix. Then, we decided to use the logistic regression model as our baseline model to predict the sentiment of the

text in the Covid-19 tweets dataset, and we first got a base accuracy of 0.53, which is well-grounded.

4.2 Advanced Model: CountVectorizer and SVC

Although the accuracy of the baseline logistic regression model is acceptable, it is imprecise for us to predict sentiment for the Covid-19 tweets dataset. Therefore, we decided to build the SVC model as the advanced model. We are going to use the sklearn function "countvectorizer" to convert text to a vector of tokens, which can help us to use the resulting matrix as input to put it into the model. Then, when we trained an SVC model with default parameter values, we got an accuracy of 0.56, which hasn't reached our goal yet. Therefore, in order to improve the accuracy, we did some parameter tuning to the SVC model. First, we loop through different kernels which are linear, polynomial, and rbf. From figure 5, we find that the SVC model using linear hyperplane achieves over 75% accuracy, but SVC models using rbf and polynomial hyperplane's accuracy are only near to 0.5. Second, we also loop through different c values (the penalty parameter of the error term) which are 0.01, 0.1 and 1, and from Figure 6, we can find that when the c value equals to 0.1, the model reaches the highest accuracy which is 0.7782. Therefore, we decide to choose the SVC model with linear hyperplane and the c value equals 0.1 as the final advanced model which we will use to predict the sentiment for the Covid-19 tweets dataset.

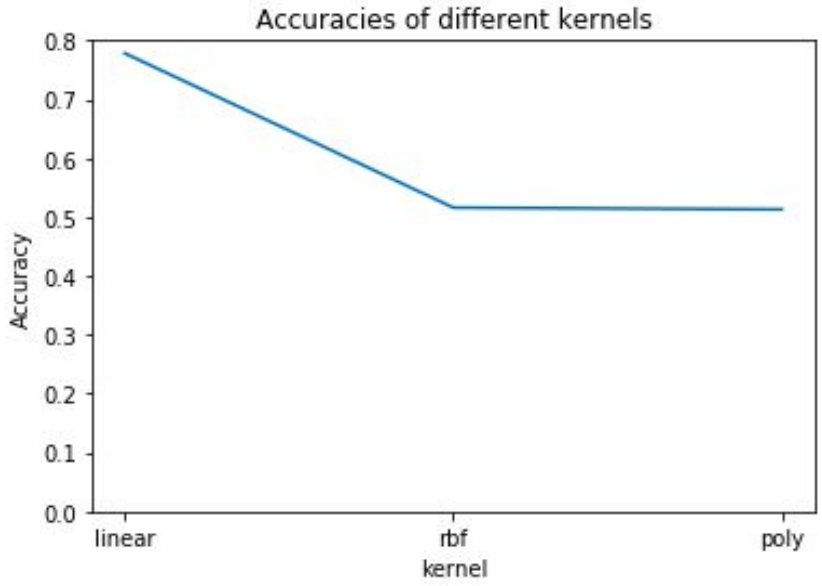

Figure 5: Accuracies of models using different kernels.

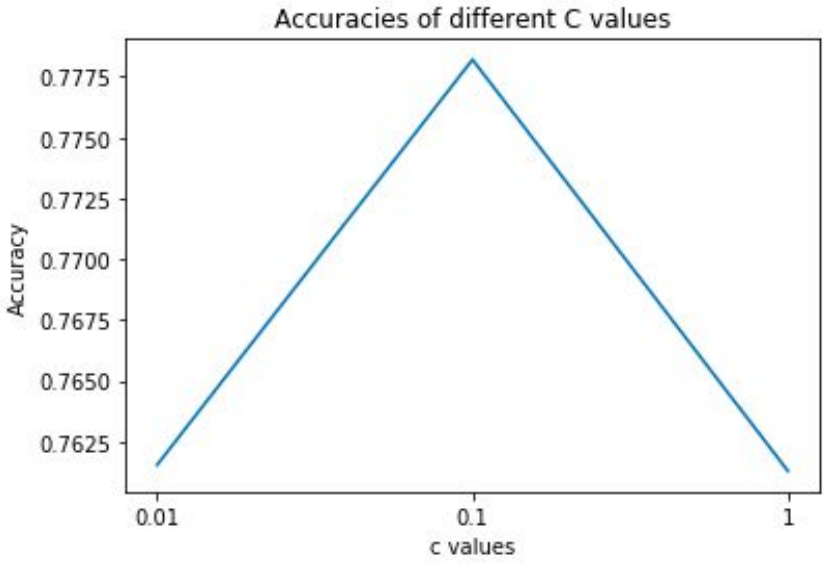

Figure 6: Accuracies of models with different c values

4.3 Analyzing Seasonality in Daily New Cases and Daily Tweet Sentiment

Our first step is to detrend the daily case data. As we can see in the graph, the daily case data has upward mobility which is the result of multiple factors such as exponential transmitting rate. Our sentiment score does not have a trend in the long run. However, both of the data have a seasonal component which could be correlated. To detrend the data, we used the seasonal decompose module to locate the trends and use regression of order 3 to fit the shape of the curve. We then subtracted the composition from the original data to obtain a flat version of daily cases with only the seasonality.

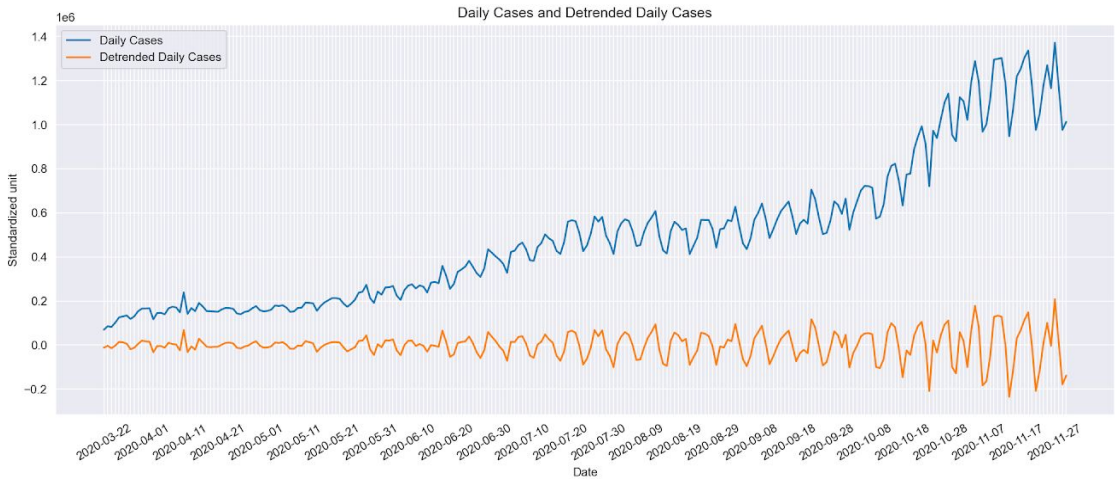

Figure 7: Daily Cases with Detrended Daily Cases

By plotting the sentiment data with the detrended data we can see that they do have similar fluctuations in the first three months, the crest and trough of the data roughly align with one another. Along the horizontal axis, we noticed that the time series match less and less. One possible reason for the irregularities in later periods is that our detrended cases daily did not take into account how the upward trend affect the magnitude of the fluctuations. With the increase in cases per day, the scale of fluctuation also increases. On the other hand, the magnitude of the seasonality in sentiment does not vary significantly[4].

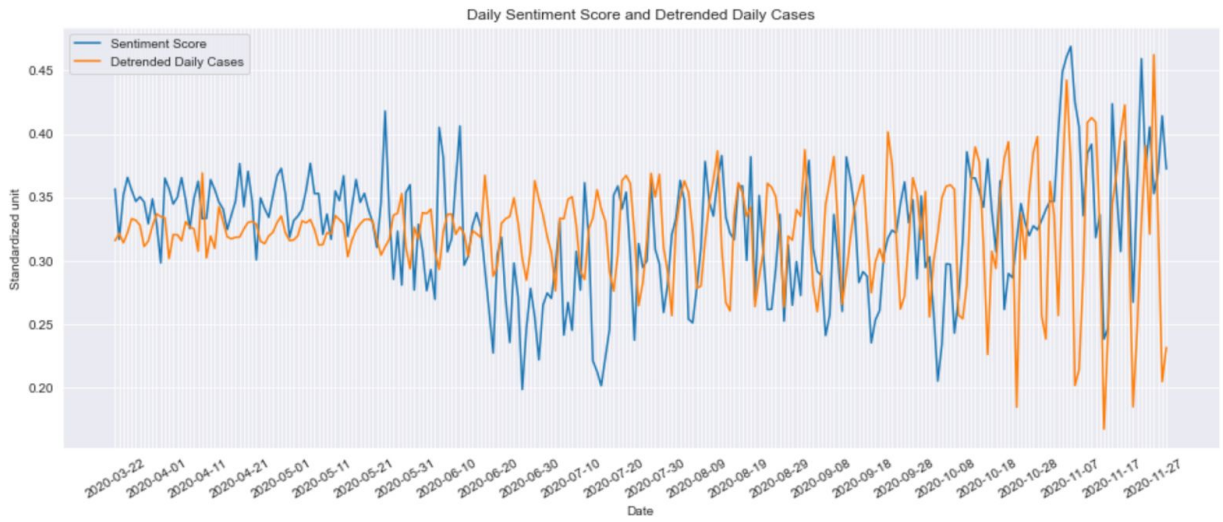

Figure 8: Sentiment Score with Detrended Daily Cases

Due to the nature of time-series data, we are not exploring the causality of these two variables since they can easily be correlated to the same set of exogenous factors which results in omitted variable problems. Instead, we study the correlation of them to gain insight into how people's moods are represented by social media affected by the daily COVID cases. We first calculated the Pearson correlation of the time series data and got a result of 0.073. The cointegration test is a statistical technique that examines if two time series are integrated together at a specified degree[5]. To perform this test, we first tested the stationarity of sentiment score and detrended cases using the Augmented Dickey-Fuller unit root test. Both of them produce a p-value of almost 0.

4.4 Frequency Decomposition

In addition, we analyzed the seasonal frequency of the time series. Fourier transformation decomposes a function of time into temporal frequencies. We performed the transformation and plotted the frequencies for two separate time series. Comparing the two plots, we found that Daily cases have a dominant frequency at around 0.14. Converting to days, 0.14 represents approximately 7 days which suggests daily cases mainly oscillate weekly.

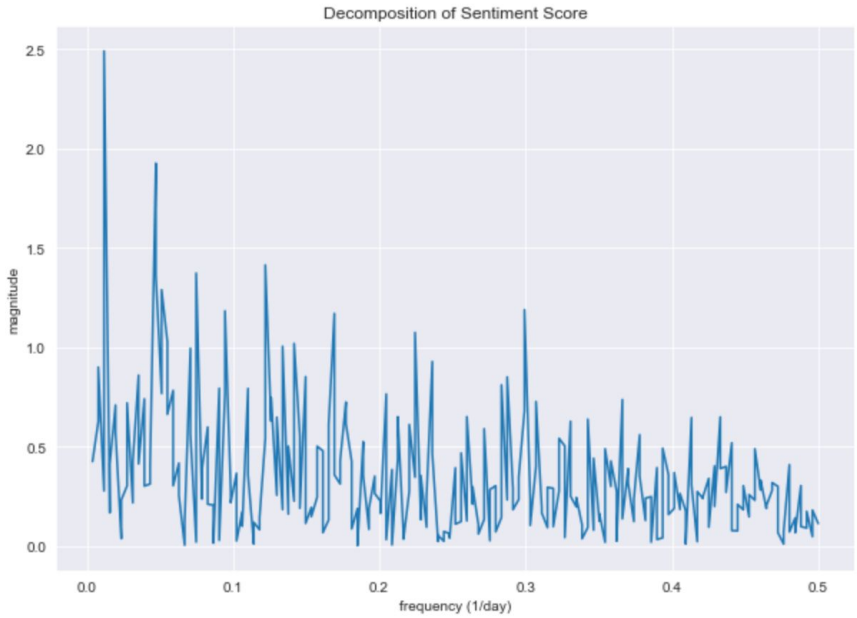

Figure 9: Sentiment Score Decomposition

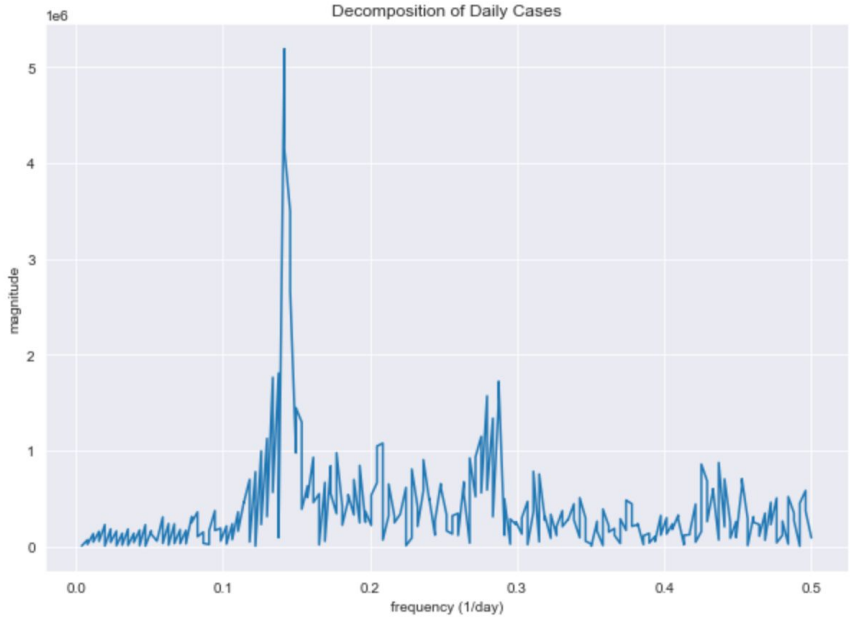

Figure 10: Daily Cases Decomposition

5 Result Summary and Discussion

After model evaluation and parameter tuning, we adopted the SVC model(c = 0.1, kernel = 'linear') with an accuracy of 0.7782 and used it to predict the daily sentiment score on the Panacea dataset.

To compare the time series of daily sentiment scores and detrended daily cases, we plotted them on the same graph. Although we observe some initial correspondence between the two variables, in order to make sure they are statistically significant, we conducted multiple tests including Pearson Correlation, Augmented Dickey-Fuller unit root test, and Fourier transform. We used Pearson Correlation to get an approximate idea of the overall correspondence between sentiment and daily cases since the test is subject to the effect of noise in time series. A score of 0.073 would suggest a weakly positive relationship in an independent context. In this case, both of our data are dependent on time which makes the interpretation subtle. We can only deduce that there is no strong linear relationship between Sentiment Score and Detrended Daily Cases. After that, we performed cointegration tests which check whether two time series are integrated in a way that does not change in the long run. The Augmented Dickey-Fuller unit root test gives both of the time series a near 0 p-value which suggests that we should reject the null hypothesis that there is non-stationarity in the data. We accept the alternative hypothesis that our data is stationary. This effectively states that there is no cointegration between the two since both of them are stationary time series while cointegration only exists on non-stationary data.

Since the previous two tests ruled out observable relationships between our variables. We studied features that differentiate them by using Fourier Transformation which decomposes temporal frequencies out of time series. By plotting the frequencies, we noticed that sentiment score has no dominant frequency while daily cases have one at 0.14 which corresponds to a period of roughly 7 days. This could explain why the first two tests do not find a strong relationship between them. If the frequency of the two does not coincide with each other, these two time series are constantly out of phase which results in low statistical significance in correlation. The weekly period of the daily cases is likely connected to the reporting mechanism. For example, some facilities may choose to report their weekly cases on Monday instead of reporting daily which may result in a high volume of cases at the start of the week. The sentiment score is calculated on a continuous daily basis, the discrepancy resulted in different oscillating frequencies of the seasonality in daily cases and sentiment.

6 Conclusion

Covid-19, as one of the biggest problems facing humans in this century, has affected all aspects of people's lives, from the way people interact, people's lifestyles, to social media. In this project, we investigated the relationship between tweet sentiment and daily cases. Although there is an observable correspondence between sentiment and daily cases during the initial phase of the pandemic. Using statistics tools, we found no testable statistically significant correlation between sentiment score and daily cases. Further analysis suggests that different oscillating frequency in the seasonality between the two is one of the reasons for low correlation.

7 Appendix

[A] Project Proposal:
https://docs.google.com/document/d/1eTF2AxABvALoJzeXtc8iFgzcj7FD4qFlAFDSwLtwMPQ/edit?usp=sharing

8 Reference

[1] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, Aboul Ella Hassanien, Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media, Applied Soft Computing, Volume 97, Part A, 2020, 106754, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.106754.

[2] Michela Del Vicarioa, Alessandro Bessib, Fabiana Zolloa, Fabio Petronic, Antonio Scalaa,d, Guido Caldarellia,d, H. Eugene Stanleye, and Walter Quattrociocchia,1 (2015) The spreading of misinformation online.

[3] Jmbanda, covid19_twitter(2020), GitHub repository, https://github.com/thepanacealab/covid19_twitter/tree/master/dailies

[4] edomt, Data on COVID-19 (coronavirus) by *Our World in Data*, GitHub repository, https://github.com/owid/covid-19-data/tree/master/public/data

[5] Corrius, Jesus. "Simple Stationarity Tests on Time Series." *Medium*, Bluekiri, 9 Oct. 2018, medium.com/bluekiri/simple-stationarity-tests-on-time-series-ad227e2e6d48.

[6] Rybnik, Rafal. "Introduction to Fourier Analysis of Time Series." *Medium*, Towards Data Science, 28 Jan. 2021, towardsdatascience.com/introduction-to-fourier-analysis-of-timE-series-42151703524a.

Yunlin Tang: Data Preprocessing, EDA, Text Analysis

Zhou Li: Introduction, Text Analysis, Sentiment Training

Jiawei Zheng: Sentiment Training, Time Series model building