Discovering the Impact of Social Media Sentiment on the Covid-19 Infection Rate Using Machine Learning and Shifts in Data

Kyler Kang

August 2, 2020

Abstract

Since the start of 2020, the global pandemic that is Covid-19 has spread at an unimaginable rate, causing panic among the public in ways never seen before. In this journal, we illustrate an approach to evaluate the impact of user sentiment generated over social media on the recent increase in the spread of Covid-19. We explore posts and comments on a well-renowned social media known as Twitter, and we incorporate the associated sentiments of the tweets. From this research, we derive a strong, positive correlation between sentiment, specifically negative, and the daily number of Covid-19 cases. We also prove the existence of a delay between the day when a tweet was posted and the day in which the number of cases increases. We interpret these datasets using different regression techniques for estimating the number of Covid-19 cases based on user sentiment. Our predictive models are evaluated using statistical scores and graphical methods that have proven effective in prediction and forecasting problems. The study of user sentiment on social media can inform us on whether or not our words are contributing to the spread of Covid-19 in a costly manner.

1 Introduction

Over 650,000 people worldwide have fallen due to the recent global pandemic, Covid-19 [1]. According to Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases, the United States may see up to 100,000 Covid-19 cases a day in future months to come [2]. Many of these cases arise from a failure to take safety precautions and a lack of discipline, which may be linked to sentiment, both in public and on social media. The field of sentiment analysis focuses on deriving meaning and emotions from people's words and phrases [3]. As a result, the public's understanding on a certain topic can be detected through a study on their sentiment, which can be especially useful to identify when the public is misled, either by each other or by the media. Our field of research delves deeper into social media sentiment surrounding Covid-19, and how this has affected the spread of it. The spread of deadly diseases like Covid-19 can be controlled if the correct measures are taken, which may save us from another deadly outbreak of Covid-19 or any other disease that may peak in the future.

A critical challenge to preventing the spread of Covid-19 arises due to ignorance and misunderstanding. Hence, those who are uneducated or unaware are more prone to the unintentional spreading of Covid-19. Furthermore, many scientists are pursuing studies on the most important factors regarding the rapid spread of Covid-19 and diseases in general. The use of social media sentiment to study Covid-19 can only go so far due to a computer's ability to understand a human-written text. Some literary techniques, including sarcasm, irony, and rhetorical devices can only be accurately interpreted by humans [4]. Therefore, public sentiment is generally a difficult task to measure.

To prove the existence of a delay between the day of a tweet and the day the case is confirmed, we experimented by shifting the Covid-19 cases dataset back. Next, we utilized Pearson's r coefficient to determine if there was a high correlation between the number of tweets and the number of cases, from a shift of 0 all the way to a shift of 21. To our knowledge, the correlation between tweets and Covid-19 infection rate has not been studied yet with our method of obtaining the peak in correlation after a certain shift in days past the initial tweet. With the delay, we were able to create a best fit model for the relationship between tweets for each sentiment and daily Covid-19 cases. Our method effectively discovers the delay in days in which the correlation between tweets and infection rate is highest and provides a different outlook on the impact of social media on disease spread.

Again, it is vital to understand the public opinion surrounding Covid-19 because credible sources and accurate information can help reduce confusion around the pandemic and educate those who are unaware of the dangers of Covid-19. Ultimately, our research aims to discover if sentiment has an impact on Covid-19 infection rate, so necessary measures can be taken to stop the spread rather than fuel it. Boosting public morale can go a long way in preventing even more deadly outbreaks of Covid-19. If indeed sentiment plays a major role in the spread of diseases, the correct actions can be taken to mitigate the problem before its too late.

2 Literature Review

Many researchers and businesses have started to utilize sentiment analysis and Natural Language Proccessing (NLP) for many purposes. Companies and businesses have implemented sentiment analysis to better understand consumer trends and customer reviews. Researchers have begun to form predictions on certain trends with the versatility of sentiment analysis models [5].

One example of researchers utilizing sentiment analysis to their benefit is to predict stock market trends. In an article, two authors pursue the effects of public opinion on stock market prices. The main objective of the project for the researchers was to create an accurate predictive model for stocks, through the use of sentiment analysis on Twitter. They discovered that the Self Organizing Fuzzy Neural Network (SOFNN) produced the highest percentage of accuracy and correlation when predicting prices based on Twitter sentiment. Classification techniques like logistic regression performed poorly on the data, while linear regression performed decently. Ultimately, they identified that their research failed to accurately classify public sentiment, which may have caused some errors [6].

Another research study done by a group of researchers demonstrates an algorithmic approach to sentiment analysis. The authors derived a method to measure sentiment in certain texts, utilizing a dictionary with a large database of words, all marked with a value of -1, 0, 1, signifying whether a word is positive, negative, or neutral. From the words in the text, they were able to calculate the sentiment score for a statement based on the amount of negative and positive words in a statement. However, the paper aims to reveal the challenges and limitations of their algorithm. Specifically, their algorithm has trouble detecting spam messages, identifying fake reviews, and judging context [7].

In the field of sentiment analysis as a whole, some words and phrases can only be accurately understood by humans themselves rather than computers. While these occasions are rare, human language features like sarcasm can be misunderstood by computers, and therefore, lead to inaccurate results in a study. Nowadays, researchers utilize unique approaches to analyze sentiment in text. A sentiment analysis process explained by Tom Yuz from the Medium journal uses text tokenization and filtering to identify the sentiment (Figure 1).

In our research, we do not expect to build a breakthrough sentiment analyzer that addresses all these flaws. Instead, we aim to transform and normalize our data of sentiment scores to reduce fluctuations and inaccuracy within the dataset. Therefore, we can use user sentiment to discover its impact on Covid-19 infection rate. That being said, we hypothesize that negative user sentiment can contribute to an increase in daily Covid-19 cases, while positive user sentiment can help lower infection rate.

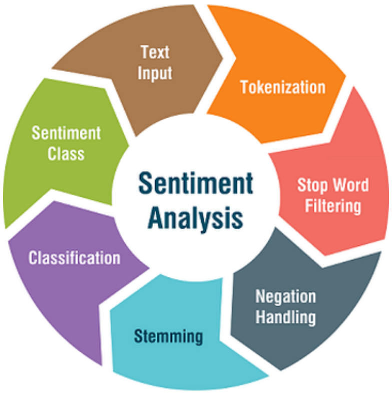

Figure 1: A diagram representing a modern approach to sentiment analysis. The process itself starts with "Text Input" and ends at "Sentiment Class" [8].

3 Purpose

1. Discover possible trends between negative, positive, neutral tweets and Covid-19 infection rate after a specific shift in data.
2. Inform the public on the possibility that sentiment on social media can impact the spread of a disease.
3. Accurately portray user sentiment on social media using data transformation and normalization techniques.

4 Methods

Discovering the effect of Twitter sentiment around Covid-19 on infection rate involves five main steps including data gathering, data normalization, data transformations, graphing, and regression. This is displayed in Figure 2.

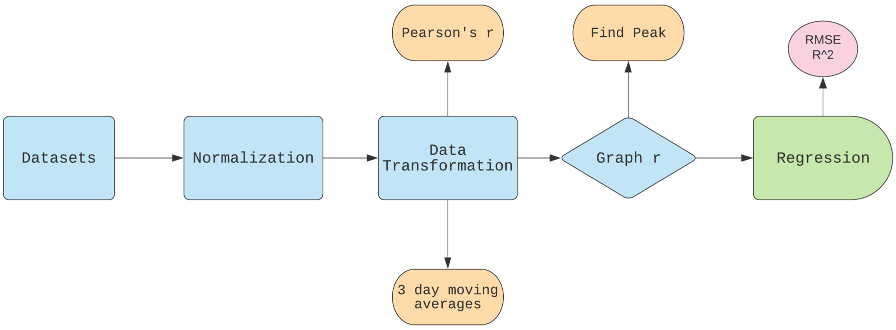

Figure 2: A diagram representing the process used to conduct our research.

4.1 Data Gathering

To measure Twitter sentiment around Covid-19, we utilized a group of datasets of all the tweets (English only) from each day starting on March 20, 2020 up until July 1, 2020. The datasets are from IEEE, and they contains millions of tweet IDs for every day and their respective sentiments [9]. The tweets contain specific keywords and hashtags related to Covid-19, including “corona”, “coronavirus”, “covid”, and more. Furthermore, each tweet in these datasets comes attached with sentiment scores ranging from -1 to 1. These scores, attained by the authors of the datasets, were derived from a pre-trained sentiment analysis model called Textblob. Textblob receives lines of text and produces a raw sentiment score based on the amount of positive and negative words in the statement [10]. While all sentiment analysis models have their flaws and limitations, we used these sentiment scores because Textblob is a reliable, effective, and convenient sentiment analyzer.

Additionally, we used a dataset from Our World in Data (OWID) of all the daily confirmed Covid-19 cases in the United States from March 20, 2020 to July 22, 2020 [11]. The number of daily Covid-19 cases will also be referred to as the infection rate in this journal. To process and perform all our methods on the data, we utilized many Python packages including numpy, pandas, matplotlib, and scikit.

4.2 Normalization

Because the sentiment scores of the Twitter dataset are normally distributed (Figure 3), we were able to normalize the data. Before we normalized the data, we took random samples of 250,000 tweets out of each day in our specific time period and combined them into one list. We did this so that we could have a constant number of tweets for each day. We set the random state to 25 when selecting random samples of tweets. While we initially used raw sentiment score means to represent the sentiment for each day, we felt that normalizing the data was much more accurate, since the mean value does not reflect the amount of negativity and positivity in one day on Twitter. We then calculated the mean and standard deviation of all the sentiment scores for all the tweets. For each of these tweets for every day, we assigned any tweet within 1 standard deviation from the mean with a value of 0, meaning neutral. Anything to the right and left of the 1 standard deviation would be assigned with a value of 1 and -1 respectively. After this process, a count of all the negative (-1), neutral (0), and positive (1) tweets from each dataset were gathered and appended to specific lists for each day. To clarify, the three representative sentiments used in this journal are negative, neutral, and positive, and they will be referenced throughout.

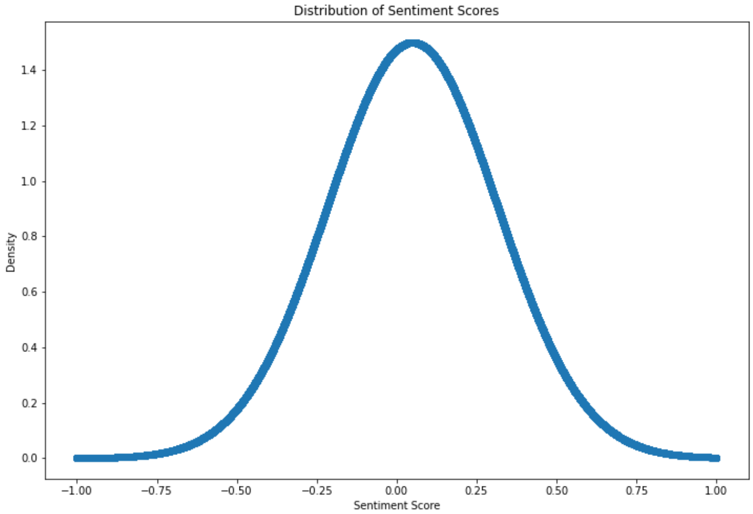

Figure 3: This figure shows that the Twitter sentiment score data is normally distributed.

4.3 Data Transformations

Instead of using the raw values for both the daily number of Covid-19 cases and the number of tweets of each sentiment, we used three day moving average values. We made the decision to use three day moving averages because these values reduce unnecessary noise and short-term fluctuations in the graphs and data. While we considered the use of seven day moving averages, we believed that this would average the data out too much which would remove some useful insights that can be gained from the data.

To best measure correlation between tweets and cases, we utilized Pearson's R, which is best described as, $r_xy = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$. Relating this to our project, x represents the number of tweets for a specific sentiment and y represents the daily number of cases. A variable with a bar on top signifies the mean of that series. First, we calculated the correlation between the number of tweets for each sentiment and the daily number of Covid-19 cases. We then shifted the Covid-19 cases dataset back one day and calculated the correlation again with the tweets. We shifted the dataset back all the way up until 21 days after the inital tweet. In total, we calculated one correlation coefficient for each delay of days up to 21. For example, a shift of 21 days would mean that a dataset with the number of negative tweets starting on March 20, 2020 would be paired with the daily number of Covid-19 cases starting on April 10, 2020, and a correlation would be calculated. We gathered a total of 62 Pearson correlation coefficients, 21 for each sentiment.

4.4 Graphing Pearson's R

To visualize peaks in correlation between tweets and infection rate, we graphed the change in correlation over the number of days offset, from 0 to 21 days, for each sentiment. We then found the absolute maximum of each graph and located the X value of that peak. For graphs with mostly negative correlation, we found the absolute minimum for these graphs instead. This is because we were looking for the highest absolute value in correlation. This X value would be our official delay in days which would be used for the final step in the project. We had a total of three X values for delays, one for each graph. The results can be found in section 5.1.

4.5 Graphing Tweets vs. Cases

Our next step includes graphing the number of tweets for each sentiment with daily Covid-19 cases in order to create a best fit model, in which we could predict cases based off the number of tweets. We graphed the number of negative, positive, and neutral tweets in each day (1 graph for each sentiment) and compared that to the daily number of Covid-19 cases, delayed by the X value found in the step above. This X value would be our official delay. Therefore, if the delay was set to 10 days for negative tweets, then we would pair negative tweets on March 20, 2020 with the newly confirmed Covid-19 cases of March 30, 2020, and so on. Again we used the three day moving average values for the number of tweets and cases instead of the raw values when creating our final results.

4.6 Regression

After plotting three graphs of tweets of each sentiment versus cases after a specific delay, our last step was to find a best fit model for all three graphs. In order to do this, we created a function to test which linear / polynomial model fit the data the best. Our function designed a polynomial model of every degree up to 50 and fit that to the data. Specifically, we utilized a python package known as scikit to fit our model and predict values. We calculated the Root Mean Squared Error (RMSE) and the coefficient of determination (R-squared) for every model of each degree to determine which model had the best fit. The RMSE gives an accurate representation of the difference between predicted values and actual values, which is perfect in our case when trying to find a good fit. The R-squared coefficient also produces a value to determine whether the model fits the data. We then charted the change in these values as the degree of polynomial increased for each sentiment and chose the most optimal degree for each. These results can be found in 5.2. In this instance, the models with the lowest RMSE and the highest R-squared value would be considered optimal. We also used LASSO regression for all three graphs, using the same package as mentioned above. When experimenting with LASSO, we used all values of alpha from 0 to 1 incremented by 0.1.

After identifying the best fit model for each graph, we used the weights and bias values for our final equations. The final graphs with the best fit model can be found in 5.3.

4.7 Sub-experiments

Before we performed the methods listed above that produced our final results, we approached this research problem with three main steps. First, we gathered all the datasets mentioned previously for all the tweets IDs that had sentiment scores correlated with them. With these scores, we calculated the sentiment score mean of each day and compared the scores with the daily number of Covid-19 cases, deaths, recovered cases, and public mobility percentages. We experimented with different days of offset to best represent the situation. For example, we graphed sentiment score mean versus the daily number of Covid-19 cases offset by 14 days. We experimented more by comparing different datasets with each other. Afterwards, we performed regression on these graphs. To do this, we utilized python packages, specifically scikit, to make a model and attempt to produce a linear or polynomial fit. However, before we went deeper into this experiment, we decided that a raw sentiment score mean for every day was not an ideal way of representing user sentiment. Instead, we used the methods described in 4.1 - 4.6 for our final results.

Furthermore, we implemented another pre-trained sentiment analysis model from NLTK to produce our own sentiment scores with samples of tweets. We used a geo-tagged dataset with Tweet IDs and ran the tweets through the sentiment analysis model after hydrating it with an application called Hydrator. Hydrator takes the tweet ID and produces the actual text of the tweet and other important imformation, including location. The NLTK Sentiment Analyzer produces four sentiment scores: compound, negative positive, and neutral. We calculated the sentiment score means for every dataset. The purpose of this step was to find the public sentiment of each state, however, we did not have enough time to further explore this field.

5 Results and Discussion

5.1 Correlation Graphs

The correlation between negative tweets and Covid-19 cases has the highest performance at around 15 days past the original tweet day (Figure 4). The correlation between the number of negative tweets and new confirmed cases was around 0.481 at that point, which represents a strong linear relationship.

The correlation between positive tweets and Covid-19 cases was mostly negative after starting out positive for the first few days. The highest performance in correlation between positive tweets occurred with an offset of 21 days, with a value of around -0.335 (Figure 5). While this value is the lowest rather than the highest, we were looking for the highest performance, which is the greatest absolute value in correlation.

The correlation coefficient for neutral tweets and Covid-19 cases also starts positive but then turns negative after 5 days. We find that the highest performance in the correlation between neutral tweets and confirmed cases occurs at a delay of around 14 days, with a value of -0.223. (Figure 6) Again, we took the lowest value since it gave us the greatest absolute value in performance.

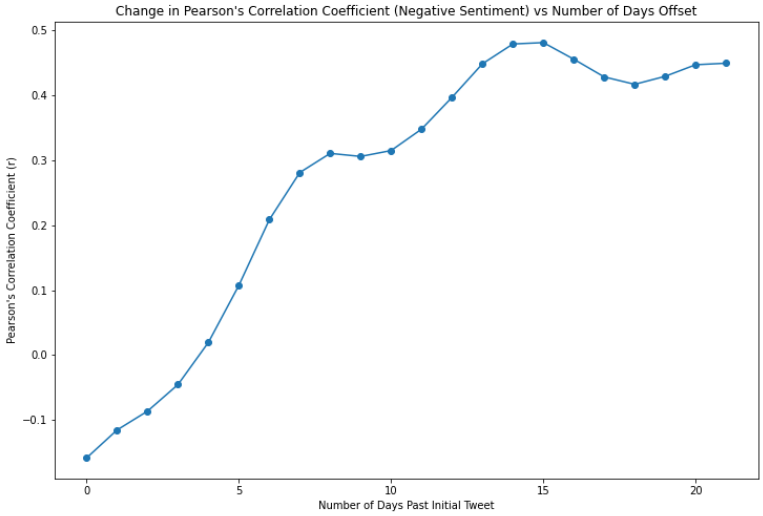

Figure 4: This graph represents the change in correlation between negative tweets and new Covid-19 cases as the number of days the Covid-19 dataset is shifted back increases. These values are all calculated using three day moving averages. There is an obvious peak in correlation at 15 days past the original tweet.

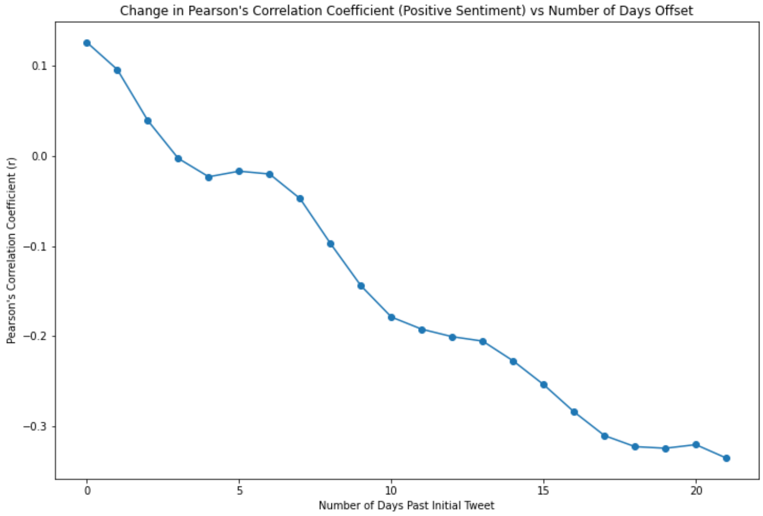

Figure 5: This graph represents the change in correlation between positive tweets and new Covid-19 cases as the number of days the Covid-19 dataset is shifted back increases. These values are all calculated using three day moving averages. After 21 days, the absolute value of the correlation coefficient is the highest.

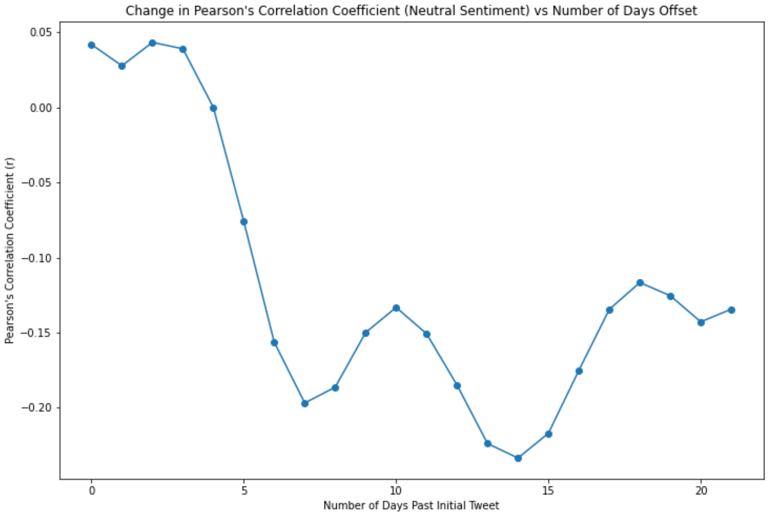

Figure 6: This graph represents the change in correlation between neutral tweets and new Covid-19 cases as the number of days the Covid-19 dataset is shifted back increases. These values are all calculated using three day moving averages. After 14 days, the absolute value of the correlation coefficient is the highest.

5.2 Regression Results

When testing for the highest coefficient of determination (R^2), we saw that there was a peak in R^2 at a degree of around 3 for all three sentiment graphs (Figure 7). For the negative graph (blue line), the coefficient of determination was higher at a degree of 15. However, a high degree usually leads to overfitting, and the difference in R^2 is not significant. When testing for the lowest root mean squared error, there was an absolute minimum at around degree 3 again for all three graphs (Figure 8). Same with that of the R^2 value, the negative graph had a lower RMSE at degree 15, however, using the same logic, a degree of 3 is optimal in this situation since a polynomial with degree 15 has a high chance of overfitting. Therefore, by testing for the best fit for every polynomial model from degree 1 (linear regression) to degree 50, we discovered the optimal polynomial model for all three graphs was at a degree of 3.

When utilizing LASSO regression (not shown in the graphs), we used an alpha value of 1.0 since it produced the lowest RMSE value for all three graphs compared to other values of alpha. However, the LASSO regression had a very high RMSE compared to linear and polynomial regressions. The model had a minimum RMSE of 10260 for the negative graph, 13435 for the positive graph, and 10620 for the neutral graph. Therefore, we stuck with the 3rd degree polynomial for all 3 graphs as our line of best fit.

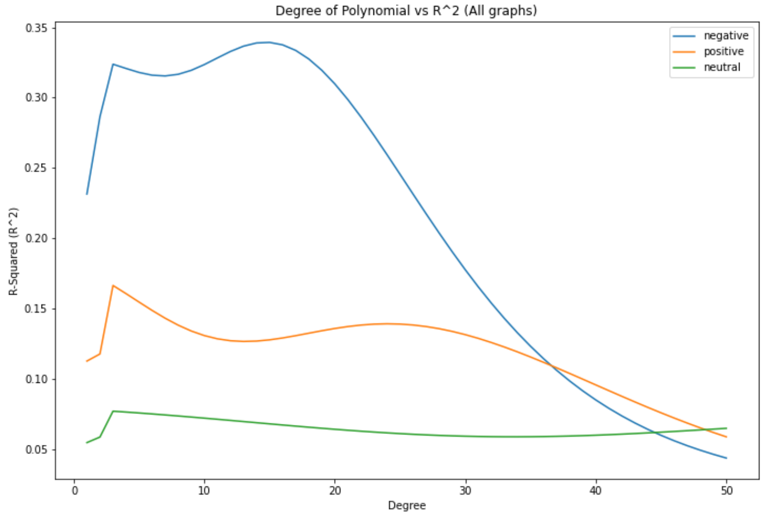

Figure 7: This graph represents the coefficient of determination (R^2), displayed on the Y axis, for each polynomial model from a degree of 1 to a degree of 50 as portrayed in the X axis. Each line represents a different sentiment. A higher R^2 value is optimal.

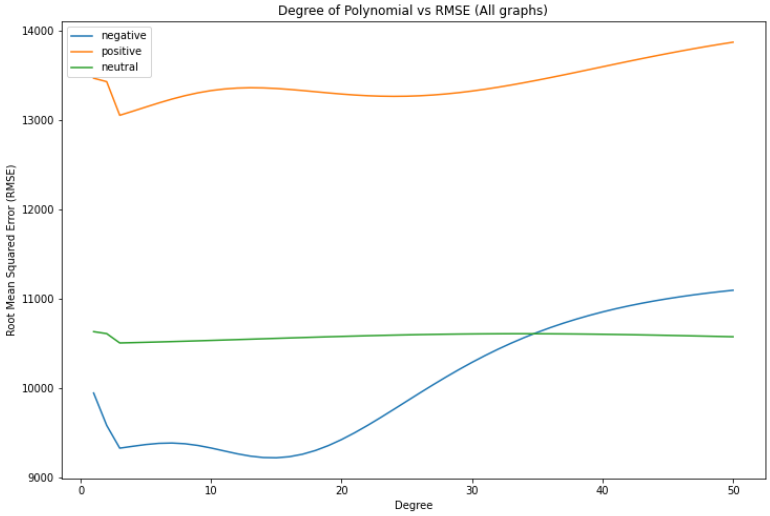

Figure 8: This graph represents the Root Mean Squared Error (RMSE), displayed on the Y axis, for each polynomial model from a degree of 1 to a degree of 50 as portrayed in the X axis. Each line represents a different sentiment. A lower RMSE value is optimal.

5.3 Tweets vs. Infection Rate

When experimenting with different regression techniques, we found that a 3rd degree polynomial outperformed the LASSO regression and the linear regression for all three graphs, as shown in the results above. When graphing the number of negative, positive, and neutral tweets with the daily number of Covid-19 cases, we delayed the Covid-19 cases dataset by a certain number of days. By delay, we mean that we compare the number of tweets for each day, starting on March 20, 2020, with the newly confirmed Covid-19 cases X number of days after the original day, with X being the official delay found in 5.1. This means that if the delay was 10, then the Covid-19 cases dataset would start on March 30, 2020.

When graphing negative tweets with infection rate, we found a strong, positive correlation between the amount of negative tweets in one day and the daily number of Covid-19 cases 15 days after the initial tweet (Figure 9). The final polynomial that we used to best model the data is also displayed in Figure 9 with red dots, and it is a 3rd degree model with equation: $\hat{Y} = -8.15 * 10^1 X_1 + 3.44 * 10^-3 X_1^2 - 4.65 * 10^-8 X_1^3 + 648557.34$ (Figure 9). This polynomial model holds a RMSE of 9331 and a R^2 of 0.324. All values used in these graphs used three day moving averages as usual.

Unlike the negative sentiment graph, the correlation between the number of positive tweets and infection rate with a delay of 21 days was relatively negative. Again, a 3rd degree polynomial was used to best predict the data with an equation of: $\hat{Y} = 3.32 * 10^2 X_1 - 9.96 * 10^-3 X_1^2 - 9.83 * 10^-8 X_1^3 - 3619132.96$ (Figure 10). This model carries a RMSE of 13055 and a R^2 of 0.17.

The correlation between the number of neutral tweets and infection rate, delayed by 14 days, has less of a strong trend compared to the other graphs. The best fit model is a 3rd degree polynomial with the equation: $\hat{Y} = 1.22 * 10^3 X_1 - 6.36 * 10^-3 X_1^2 + 1.10 * 10^-8 X_1^3 - 78198896.52$ (Figure 11).This model has a RMSE of 10508 and a R^2 of 0.077.

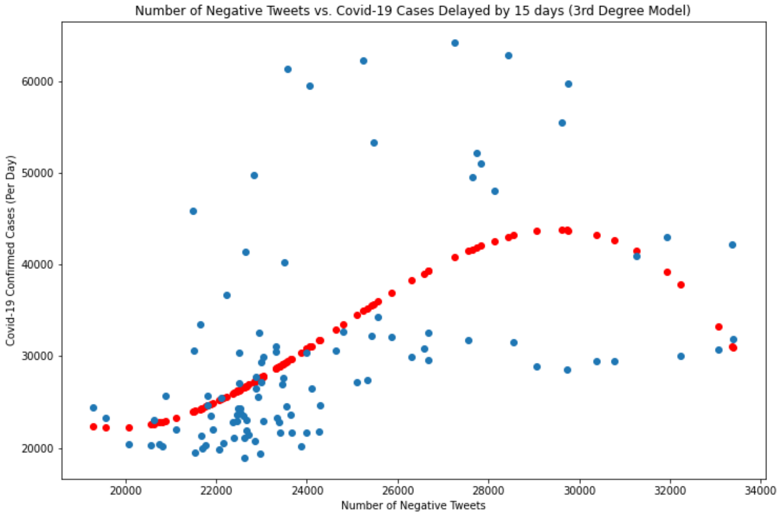

Figure 9: This scatter plot reflects the number of negative tweets in a day compared with the number of daily Covid-19 cases delayed by 15 days. We obtained the delay of 15 days from the previous results (change in correlation vs number of days offset). These values are all calculated using three day moving averages. The red dots denote the line of best fit for the graph which is a 3rd degree polynomial.

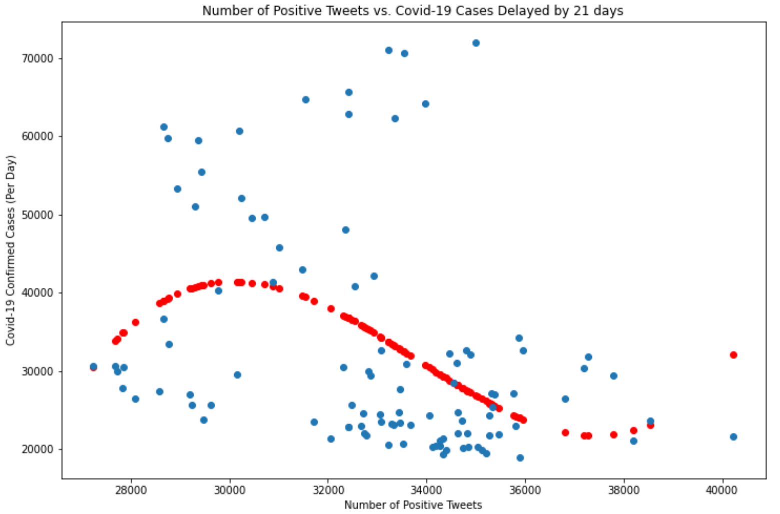

Figure 10: This scatter plot reflects the number of negative tweets in a day compared with the number of daily Covid-19 cases delayed by 21 days. We obtained the delay of 21 days from the previous results (change in correlation vs number of days). These values are all calculated using three day moving averages. The red dots denote the line of best fit for the graph which is a 3rd degree polynomial.

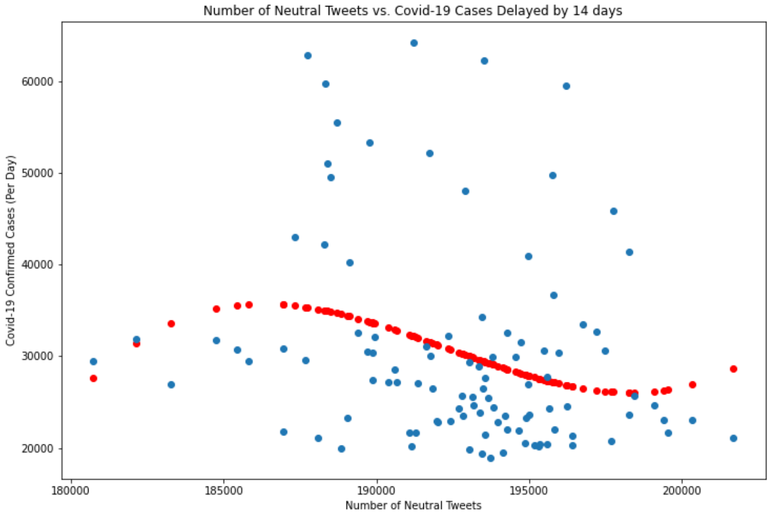

Figure 11: This scatter plot reflects the number of neutral tweets in a day compared with the number of daily Covid-19 cases delayed by 14 days. We obtained the delay of 14 days from the previous results (change in correlation vs number of days). These values are all calculated using three day moving averages. The red dots denote the line of best fit for the graph which is a 3rd degree polynomial.

5.4 Discussion

The most useful way to measure the impact of user sentiment on Covid-19 infection rate is by using the correlation between negative sentiment and Covid-19 infection rate. Out of the three categories of sentiment, the correlation between negative tweets and daily Covid-19 cases was the only one that had a positive correlation at its peak. The infection rate was highest 15 days after the initial tweet, with a correlation of 0.481. The correlation reveals a strong, positive relationship between negative tweets and infection rate, further supporting our original hypothesis. Our research demonstrates that someone who tweets something that causes them to be exposed to Covid-19 will take on average, 15 days to test Covid-19 positive, since the correlation is highest at that point. This happens perhaps because negative sentiment reflects risky actions, therefore leading to a higher rate of Covid-19 cases. However, the chances of that exact situation happening is low, hence why there is not a 100 percent correlation because not everyone who tweets negatively will be exposed to Covid-19. Nonetheless, this discovery makes sense because previous public health research states that a person can feel symptoms of Covid-19 2-14 days after they were initially exposed to the disease [12]. The trend in our results aligns with this statement, as it usually takes a few extra days after feeling symptoms to get tested and confirmed as an official Covid-19 case.

Our findings lead to a call to action for society to take safer precautions in order to stay unharmed. It is best that our words do not lead to unnecessary and critical actions that may contribute to the spread of Covid-19 and jeopardize others. The results from finding the correlation between negativity and infection rate can be beneficial to the study of public health and diseases like Covid-19. Along with finding a cure, the public can help spread hope among social media and increase positive sentiment so people are encouraged rather than fearful.

From the polynomial models gathered in the results, we can predict Covid-19 cases based on tweets. However, due to the high RMSEs and low R^2 values, it is hard to use these models to accurately reflect Covid-19 cases in the future. Essentially, our best fit models limit our ability to predict cases based on tweets of a specific sentiment; however, more accurate predictive models can be created in the future to more confidently predict Covid-19 cases based on user sentiment. Still, we can see that as the number of negative tweets increase, the number of Covid-19 cases 15 days later also increases exponentially. However, this trend halts after the amount of negative tweets passes 30,000 per 250,000 tweets. The positive trend between negative user sentiment on Twitter and Covid-19 infection rate for the first 30,000 tweets is significant because this indicates that negativity may be a contributing factor to spikes in infection rate.

With this information, it is worth it to spread positivity over negativity on social media. While this will not stop the infectious spread of Covid-19, it cannot hurt to educate and calm the public in times of despair, which may help people make smarter decisions regarding Covid-19. Informing the public on the impacts of negative user sentiment on social media can be extremely beneficial and help remove Covid-19 cases that are caused solely from negative tweets.

The trend between positive tweets and infection rate is also significant. Again, because there is a high RMSE with the polynomial model, we cannot predict cases based on the number of positive tweets in a day. However, it is interesting that as the amount of positive tweets increase past a certain point, the infection rate decreases. This logic follows the trend of our previous result since cases decrease as positivity increases. The correlation found in the negative and positive graphs demonstrates the impact of positivity and negativity on the Covid-19 infection rate.

Lastly, our results demonstrate that there is a very weak trend with neutral tweets and infection rate. It makes sense that neutral tweets on social media have no real impact on people's thoughts and actions. In the dataset we used, neutral tweets include retweets of articles and news headlines and general statements with little bias and opinion. Therefore, it is hard for us to relate neutral tweets with a spike or decrease in Covid-19 cases.

6 Conclusion

In our research, we used an approach of finding the highest correlation between tweets and cases after the cases are shifted back a certain number of days. By doing this, we could effectively evaluate the impact of tweets on the Covid-19 infection rate. While we did not directly address the limitations brought upon by sentiment analyzers, our method of data transformation and normalization effectively reduces fluctuations and errors that may be caused by inaccurate sentiment readings from computer machines.

Our research demonstrates that negative tweets have the largest correlation with Covid-19 infection rate compared to other sentiments after a certain delay.

When calculating the delay, we used Pearson's correlation coefficient to compare the number of tweets in a sentiment and the confirmed Covid-19 cases for that day. We then shifted the Covid-19 cases dataset back, from 1 day to 21 days, and calculated the correlation for each value of offset. Therefore, we gathered 21 total correlation coefficients for each sentiment.

For the negative tweets graph, we discovered that there was an obvious peak at 15 days past the original tweet. More specifically, one who tweeted and got exposed to Covid-19 has the highest chance of testing positive 15 days after their original tweet. Additionally, positive and neutral tweets had the most negative correlation after 21 and 14 days respectively, which connects with the logic that negativity increase infection rate.

After testing for the best regression model, we discovered that a 3rd degree polynomial was most optimal for every graph. From these prediction models, we can conclude that negative tweets have the best ability to predict Covid-19 cases in the future. Obviously, this should be taken lightly due to the high RMSE and low R^2 of the polynomials. It is also difficult to evidently state that there is a cause and effect relationship between negative tweets and infection rate, but our study focuses on the correlations between the two, which is evidently strong.

Our results demonstrate that an increase in negative tweets strongly correlates with an increase in Covid-19 cases. Our motivation for this research was to discover the weight of words on Twitter, and whether these words actually impact the spread of Covid-19. While we cannot definitively say that negative tweets caused a spike in Covid-19 infection rate, we successfully found that negative user sentiment on social media does, in fact, positively correlate with daily Covid-19 cases, so therefore, certain behaviors on social media regarding Covid-19 should be mitigated for the benefit of the world. Any decrease in Covid-19 cases is significant for the world, and an increase in positive sentiment could help this cause in the long run. While increasing positivity and awareness may not single-handedly stop the spread of Covid-19, it will do more good than harm for society, and simultaneously it may help mitigate the spread of Covid-19.

In general, it is hard to perfectly understand public opinion surrounding Covid-19 through sentiment analysis because it is difficult for a computer to comprehend certain phrases and rhetorical devices. Many times, computers can take certain sentences out of context. Future work on sentiment analysis models can vastly improve the field and help generate precise representations of public sentiment. Further research incorporating location of tweets would be beneficial for understanding the impact of sentiment on Covid-19 cases in certain counties. Furthermore, these results could be used to predict cases in certain counties, especially in counties that have just started testing. With this information, the pandemic that is Covid-19 can be mitigated if the right precautions are taken in time. Lastly, additional regression techniques can be implemented to improve our study and our predictive model. In the future, we will attempt to use Vector AutoRegression (VAR) to accurately predict Covid-19 cases based on tweets.

References

[1] Coronavirus Death Toll and Trends - Worldometer. Library Catalog: www.worldometers.info.

[2] U.S. could see 100,000 new Covid-19 cases per day, Fauci says, June 2020. Library Catalog: www.statnews.com Section: Health.

[3] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, May 2012. Publisher: Morgan & Claypool Publishers.

[4] The benefits (and limitations) of online sentiment analysis tools. Library Catalog: typely.com.

[5] Introduction to sentiment analysis: What is sentiment analysis?, March 2018. Library Catalog: algorithmia.com Section: Machine learning.

[6] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[7] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. Opinion mining and sentiment analysis. pages 452–455. IEEE, 2016.

[8] Tom Yuz. A Sentiment Analysis Approach to Predicting Stock Returns, April 2018. Library Catalog: medium.com.

[9] Rabindra Lamsal. Coronavirus (COVID-19) Tweets Dataset. March 2020. Publisher: IEEE type: dataset.

[10] TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation.

[11] Coronavirus Source Data. Library Catalog: ourworldindata.org.

[12] How long does it take for COVID-19 (coronavirus) symptoms to appear?, June 2020. Library Catalog: www.medicalnewstoday.com.