Social Media Analytics and Digital Marketing

Prof. Xin Wang

Final Project

Coronavirus Subreddit and Recommendations for Government

June 28, 2020

Owen Thurston, Dhyey Patel, Noah Hendriks, Sam Kahn

Executive Summary

This report highlights the steps taken to analyze and draw conclusions from a data set of Reddit posts relating to the current COVID-19 pandemic. The data is analyzed through the lens of government bodies who can leverage sentiment analysis to guide the timing and magnitude of their actions in times of crisis.

The programming language, R, is used to analyze the data set through descriptive and inferential statistics, correlation and regression analyses, and multiple tests of difference. Excel is used for basic computation as well.

Analysis begins with cleaning the data, identifying word frequency, and reviewing overall sentiment across all 20 days. A sentiment shift becomes apparent on March 10th, 2020; therefore, the posts are then broken down into two groups at this dividing point. Analysis is then conducted on each individual grouping of words with the goal being to deduce reasoning behind the sentiment shift. It becomes evident that certain keywords are driving this shift through their change in frequency over time. Given this finding, it is hypothesized that a model could be developed to test the predictive power of keywords on future sentiment.

The terms proven to hold significant predictive power are used to make recommendations to governments on how they could have used sentiment analysis on the given data to take action in maintaining high levels of public sentiment between March 1st and March 20th. Recommendations are also made as to how governments can use real-time data to produce the same effect going forward in any time of crisis.

Introduction

Background on Data

The data set analyzed consisted of content from 5,000 online posts under the subreddit, "coronavirus", on the popular forum hosting website Reddit. These posts spanned a 20-day period beginning March 1st, 2020 and ending March 20th, 2020. Each day consists of hundreds of posts containing links, pictures, and personal opinions on the overarching Coronavirus topic. A key date that was immediately recognized within the data was March 11th, the day that COVID-19 was officially deemed a global pandemic by the World Health Organization[1]. In subsequent analysis, posts made before and after this day will be looked at separately to recognize the changes in sentiment that this announcement brought upon the general public.

Data Familiarization

After an initial scan, it was determined the data is predominantly headlines of different COVID-related occurrences across North America with some reactionary comments and anecdotes from users. The time range is when the pandemic began to hit the inflection point of new daily cases and restrictions from government began specifically for North America and some of Europe. The connection to helping government develop a response plan for subsequent waves was then developed and feasible with the given data set.

There consists of many different variables for the large dataset, so early on it was decided the most important variables were only the date and the content because, for a government organization, individual level data was not of major concern relative to the content and timing of the post.

Data Preparation

Prior to analyzing the data provided, it was first essential to clean it to a standard format for effective use in R. The process of cleaning the data can be broken down into six steps. It began with following standard data cleaning procedures: isolating words, removing punctuation and numbers, removing stop words and stemming words down to their roots. The fifth step involved joining words that are often used in tandem. This included combinations like 'United States' or 'cruise ship'. By joining these words, we were able to analyze their impact as a bi-code instead, like they were meant to be used, rather than as two separate words. The final step in preparing the data is to remove the terms 'COVID' and 'coronavirus' from the data set as it is seemingly redundant. Since the subreddit topic is coronavirus, the content will include coronavirus matter, there is no need to specify on an individual post level. It was also clear that leaving these words in the data set would add insignificant insight to our sentiment analysis going forward.

---

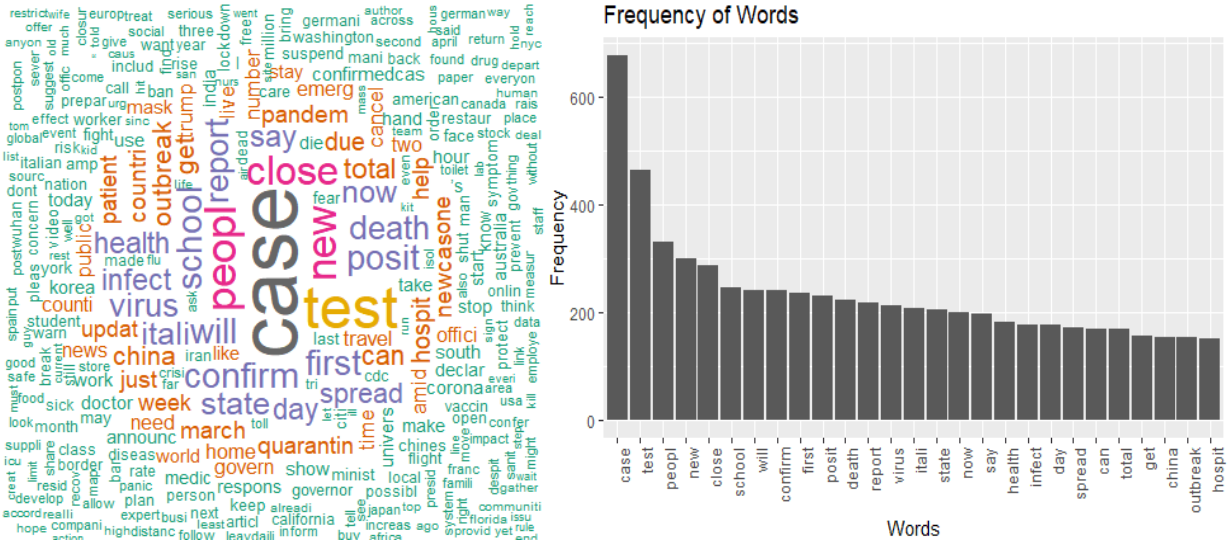
[1] https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19--11-march-2020

Analysis

Initial Observations

After the data was cleaned, the first step is to determine the frequency of each word in the 5,000 titles. The frequency was visually presented through a bar graph and a word cloud (see Exhibit 1), both showing the frequency of most of the words is similar with 'case' and 'test' being slightly higher.

Exhibit 1 – High-Level Analysis



Sentiment Analysis

A new variable called sentiment was created by using the Sentiment Analysis package in R. This would be a representation of the sentimental value of each title. From the summary statistics of the sentiment, the range is from -1 to 1 with a mean of 0.027. The mean represents the fact that the overall sentiment of all 5000 titles is slightly positive.

To take a further look at the sentiment of the posts, NRC Sentiment was calculated which includes the 8 emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust along with the negative and positive sentiments. The mean of each emotion was calculated (see Exhibit 2), and fear and trust were the most predominant emotions throughout the posts.

Exhibit 2 – Mean Emotion of Each Post

```
colMeans(nrcSent)
     anger anticipation      disgust         fear
    0.1678        0.3150       0.1264       0.4352
       joy       sadness     surprise        trust
    0.1486        0.3052       0.1546       0.4478
```

To check if the emotions changed before and after the announcement, a graph was plotted (see Exhibit 3). This graph shows that the changes in emotions before and after the announcement were not drastic, but the overall positive sentiment and positive emotions like joy went up, while negative sentiment and negative emotions like fear went down. Thereafter to determine if the emotions before and after the announcement are statistically different, 8 t-tests were conducted (see Exhibit 4). The results show that surprise is the only emotion for which the data before and after the announcement is statistically different.

Exhibit 3 – Sentiment Comparison

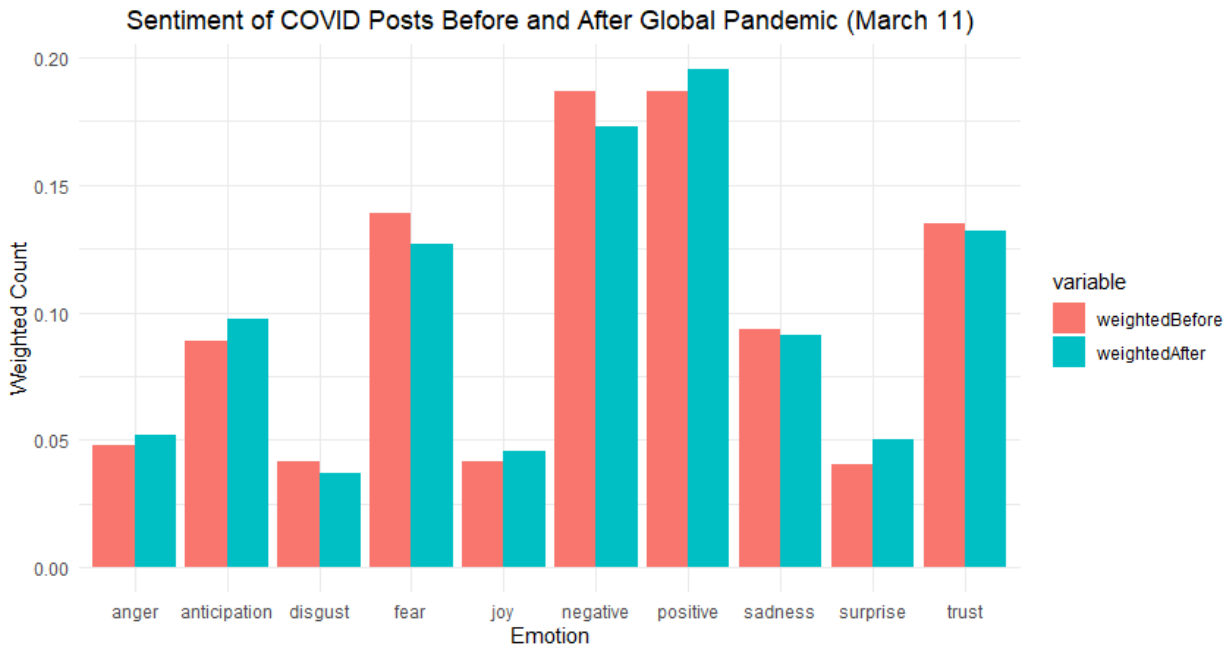


Exhibit 4 – T-Test Results

Anger

```
data:  sent2$anger and sent3$anger
t = -1.3148, df = 2969.5, p-value = 0.1887
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.044946584  0.008864539
sample estimates:
mean of x mean of y
0.1583893 0.1764303
```

Anticipation

```
data:  sent2$anticipation and sent3$anticipation
t = -1.9358, df = 3094.2, p-value = 0.05299
alternative hypothesis: true difference in means
 is not equal to 0
95 percent confidence interval:
 -0.073738425  0.000472551
sample estimates:
mean of x mean of y
0.2939597 0.3305927
```

Disgust

```
data:  sent2$disgust and sent3$disgust
t = 1.0362, df = 2882.6, p-value = 0.3002
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.01120233  0.03631252
sample estimates:
mean of x mean of y
0.1369128 0.1243577
```

Joy

```
data:  sent2$joy and sent3$joy
t = -1.3311, df = 3225.5, p-value = 0.1833
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.040998528  0.007841544
sample estimates:
mean of x mean of y
0.1375839 0.1541624
```

Sadness

```
data:  sent2$sadness and sent3$sadness
t = 0.021425, df = 3168.5, p-value = 0.9829
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.03494282  0.03571490
sample estimates:
mean of x mean of y
0.3093960 0.3090099
```

Trust

```
data:  sent2$trust and sent3$trust
t = -0.0040316, df = 2992.3, p-value = 0.9968
alternative hypothesis: true difference in me
ans is not equal to 0
95 percent confidence interval:
 -0.04437197  0.04418987
sample estimates:
mean of x mean of y
0.4469799 0.4470709
```

Fear

```
data:  sent2$fear and sent3$fear
t = 1.3324, df = 2929.9, p-value = 0.1828
alternative hypothesis: true difference in
 means is not equal to 0
95 percent confidence interval:
 -0.01388791  0.07278231
sample estimates:
mean of x mean of y
0.4597315 0.4302843
```

Surprise

```
data:  sent2$surprise and sent3$surprise
t = -2.9726, df = 3432.7, p-value = 0.002973
alternative hypothesis: true difference in means
 is not equal to 0
95 percent confidence interval:
 -0.06032587 -0.01237440
sample estimates:
mean of x mean of y
0.1328859 0.1692360
```

To better analyze the NRC sentiment, a line graph was used to plot the daily value of each emotion (see Exhibit 5). The graph shows that the fluctuations and frequency of emotions increased drastically after the announcement. The cause of this was likely the volume of posts increasing after the announcement. To dive deeper into the fluctuations, a bar plot was created (Exhibit 6) which tracked daily change in emotion. When looking at the change in emotion it is noticeable in many of the cases, the emotions are all going up together or they are going down together. More analysis was conducted to determine the cause of all the emotions, but no concrete reason was found.

Exhibit 5 – Daily Emotion Value

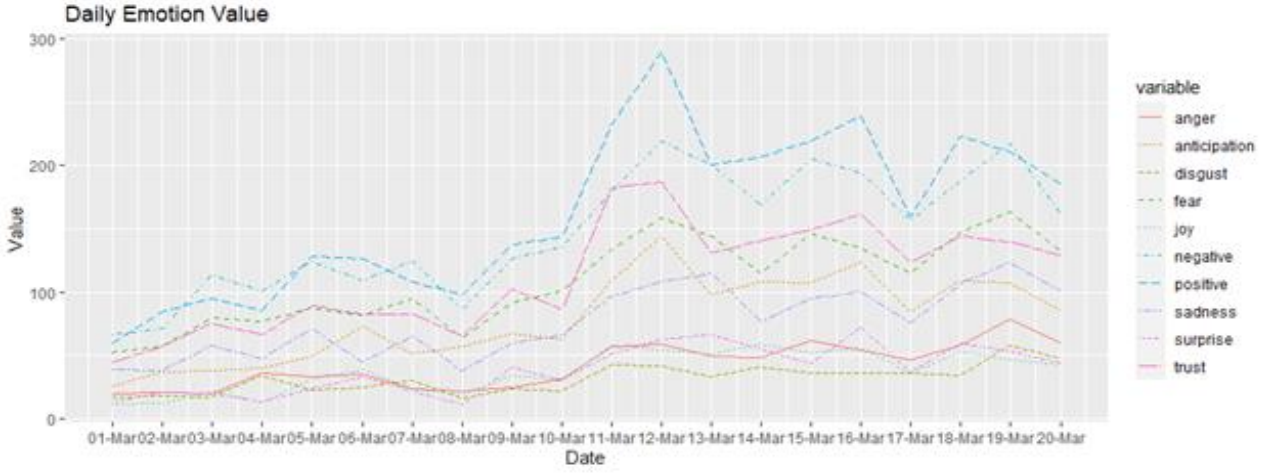
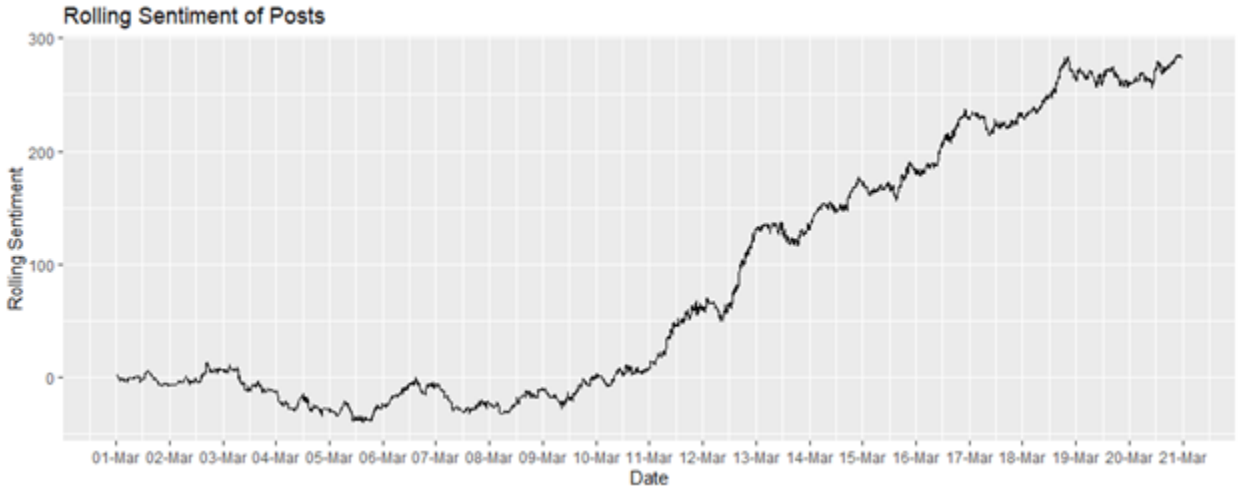


Exhibit 6 – Daily Change in Emotion

Relationship Analysis

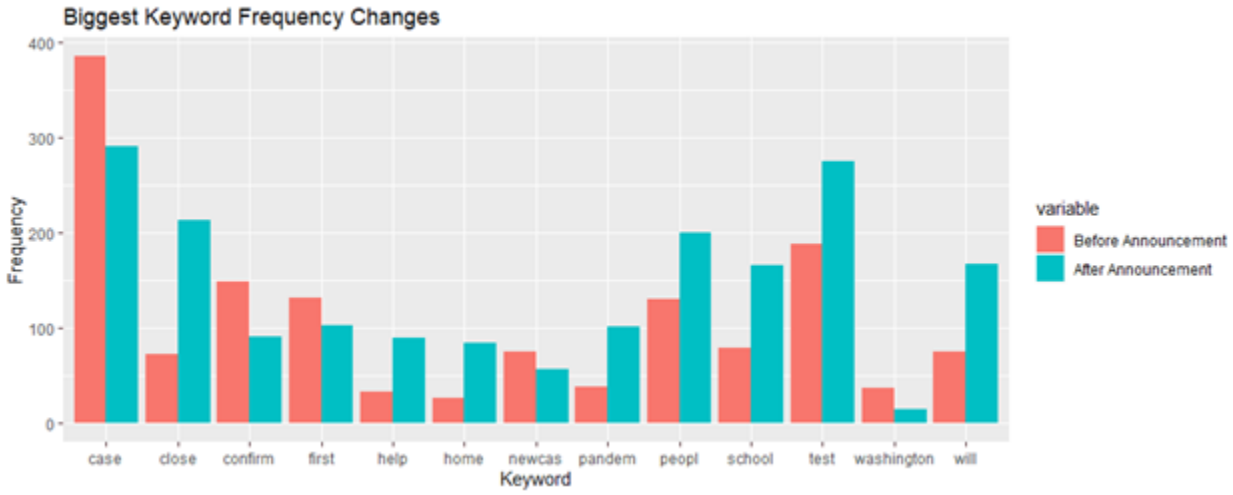
As a result, the focus was then shifted to looking at positive and negative sentiment and analyzing the rolling sum across each day (see Exhibit 7). This graph portrays the rolling sentiment that increased drastically after March 10th, 2020. To narrow in on a cause as to what changes caused the rolling sentiment to increase drastically, a bar graph was created to compare word frequencies between the two periods (see Exhibit 8). The words which increased in frequency the most and decreased in frequency the most were selected as part of the bar graph. The words with the largest decreases in frequency included case, confirm and first. This displays that respondents began focusing less on the initial increases in new cases confirmed, even though cases kept rising at an accelerating rate[2] all the way thru to March 20th.

Exhibit 7 – Rolling Sentiment



[2] https://coronavirus.jhu.edu/map.html

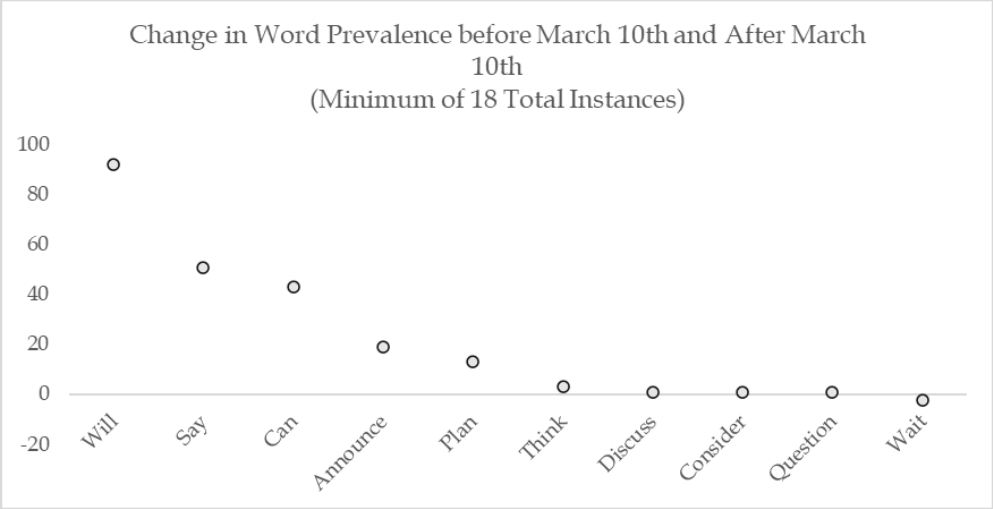
Exhibit 8 – Biggest Keyword Frequency Changes



The public's focus, instead, began drifting to the pool of words that had increased most drastically over the first half to the second half. Words like 'close', 'school', 'will' and 'home' all increased by over 100% from the first half to the next. It was hypothesized that when the government, or people in general, began to take proactive actions against the virus, the overall sentiment goes up even though the virus itself was getting *worse*. The public could not handle policymakers 'wait-and-see' approach as they felt helpless as the virus continued to spread through Italy and expected it to continue to do so through Spring Break.

To further test the hypothesis, verbs were selected from the same table of the biggest movers (see Exhibit 9). The results supported the hypothesis as words like will, say, can, plan and announce all increased in the higher sentiment period whereas words like think, discuss, consider, question, and wait all decreased or increased less. This shows the public appreciated when governments began to take the virus seriously and act instead of avoiding the virus and waiting for it to make the first move.

Exhibit 9 – Increases and Decreases in Action Words



Overall, the analysis of relationships between sentiment and specific words at this point was inconclusive as there was no direct link between the action words and restriction words and the sentiment increase. Therefore, a predictive model is employed to bridge the gap between word and subsequent emotion.

Predictive Model

Based on the words found to be most popular during the latter 10 days and least popular during the former 10 days of data, we narrowed these down to an inexhaustive word bank of 22 words that all could relate back to specific actions by the government. See below for the 22 words used:

Case, will, school, close, pandemic, confirm, lockdown, quarantine, trump, infect, say, vaccine, spread, travel, panic, death, restaurant, question, test, people, work, home

Individually each word can be used as a reactive word to describe actions taken by policymakers. For example, the word 'travel' is used when describing actions taken, or not taken, by the government to restrict travel at the initial outbreak. The word 'will' is used when actions are taken the same way the word 'say' is used when employing methods of communication, both of these action words were deemed to hold a level of positive sentiment by the public in the previous study.

The initial predictive model employed is a multiple variable regression analysis. The frequency of each of the 22 keywords is used as independent variables with a forward-looking sentiment indicator as a dependent variable. Words like "death" directly influence sentiment of posts they are within as the sentiment algorithm reads it as both 'fearful' and 'negative'. Therefore, a method to read the response of this word usage in subsequent posts was developed to analyze the lasting impact and reaction to the use of words like 'death' that directly impact the sentiment and skew results. This method is through the forward-looking sentiment indicator which analyzes at the next 10 hours of sentiment after one of the independent variables was used. The variable takes the positive sentiment

of hour one and subtracts the negative sentiment in hour one and then sums this with the next 9 hours of posts. For example, f a post containing the word 'vaccine' is published at 10am, the sum of sentiment from 11am to 9pm is pulled for that day as the responding variable to 'vaccine'

The goal of the model is to predict which words will have the most incremental benefit to the general public's perception of COVID-19. One post with one word will not influence subsequent posts every time, but the model is used to find words that produce a trend in order to deliver recommendations for the government and their policy decisions. The group does not expect the model to be bulletproof and will be satisfied with 90% confidence intervals.

The first step in every regression is to run a correlation analysis to test covariance. In this case, 'confirm' is correlated with 'case' so 'confirm' was removed and 'close' is correlated with 'school' so it was also removed.

After running the regression and removing the insignificant variables, we were left with seven significant words (see Exhibit 10). Each word has a positive coefficient so they all influence sentiment in a positive way after they are used. 'School' and 'lockdown' were both the most impactful with an increase of greater than 4 to sentiment when used. This shows that the public reaction was most positive when lockdowns were enacted, and schools were likely closed. 'School' has the lowest p-value in the group; therefore, it is successful and consistent in driving sentiment of subsequent posts. When schools were announced to be closed in mid-March, the reaction was positive as the general population felt as if it increased their personal safety having their kids be home from a hotbed like school.

Exhibit 10 – Sentiment Regression Results

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2279     0.8140  -0.280  0.77960
will          1.5257     0.7318   2.085  0.03762 *
school        4.2639     0.6020   7.083 5.17e-12 ***
lockdown      4.0132     1.4583   2.752  0.00616 **
infect        1.9130     0.8533   2.242  0.02543 *
say           2.6838     0.8416   3.189  0.00152 **
vaccine       3.1287     1.6638   1.880  0.06066 .
travel        2.1662     1.0749   2.015  0.04445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.51 on 470 degrees of freedom
Multiple R-squared:  0.2285,    Adjusted R-squared:  0.217
F-statistic: 19.89 on 7 and 470 DF,  p-value: < 2.2e-16
```

The R-squared and Adjusted R-squared values are quite low and that is a limitation of the model as most of the changes cannot be described using these seven words. More words were not added to the word bank however as words like "Italy" for example are not relevant to our recommendation.

The government cannot enact change by implementing policy with Italy for COVID in the long-term as the country likely will not be an epicenter again. The model was only relevant when the inputs were related to actions the government could take again in the future.

Trust Model

In times of crisis, like with COVID, the government's goal is to be trusted by their people so that citizens listen to their government's recommendations. In extreme situations society could fall into anarchy without influence from strong leadership. That is why a second regression was employed to analyze which words, from the same word bank, were related to increases in "trust". The only difference is that since trust is only given as a positive value instead of a balanced pos/neg, the rolling mean value of trust over the subsequent 10 hours is used instead of a rolling sum for 10 hours.

The significant words in the trust model are similar to the first, but with more emphasis on words related to changes in levels of restrictions (see Exhibit 11). For example, on top of 'school', this model includes 'work', 'travel', and 'restaurants'. During this period, most of these things/actions were becoming limited by the government. This shows that the general population agreed with the government's actions to limit working in office, eating out at restaurants and barring international travel. The public trusted the government more through the strict actions taken. Overall, the word with the highest coefficient in this model is vaccine showing that when the topic of vaccine is brought up, people trust their actions to stay home and wait for the vaccine to come.

Exhibit 11 – Trust Regression Results

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.55422    0.13130  27.069  < 2e-16 ***
school       0.39000    0.09866   3.953 8.91e-05 ***
pandemic     0.29908    0.13524   2.211 0.027492 *
lockdown     0.38669    0.23156   1.670 0.095606 .
say          0.46330    0.13377   3.464 0.000582 ***
vaccine      0.92731    0.26622   3.483 0.000542 ***
travel       0.28796    0.17267   1.668 0.096040 .
restaurant   0.77513    0.26726   2.900 0.003904 **
people       0.29379    0.10190   2.883 0.004120 **
work         0.59801    0.21262   2.813 0.005122 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.007 on 468 degrees of freedom
Multiple R-squared:  0.2346,    Adjusted R-squared:  0.2199
F-statistic: 15.94 on 9 and 468 DF,  p-value: < 2.2e-16
```

Conclusion

Recommendation

Based on the analysis of the data and the current situation with the COVID-19 pandemic, it is recommended the government focus their actions to three specific areas, public spaces, policy, and healthcare. Our model suggests that by taking action in these areas, it will generate a rise in sentiment among the general public.

Public Spaces

Actions that affect public spaces should be prompt in closing and then slowly reopening these spaces, all while being in accordance with guidelines provided by WHO and other medical institutions. More specifically, the government should focus on releasing announcements that relate to schools, restaurants, work and travel. When the media responds to announcements that relate to these topics, there tends to be a rise in sentiment. In future waves, there is no assumption that there will be a positive reaction to closing more schools or more workplaces, but it does mean the public will focus on these keywords as they create the most interest over trust and sentiment.

Furthermore, these announcements should have clear communication as to why precautions are taking place, as the public reacts more positively to this. This was determined through the words 'say' and 'will' being positively correlated with reactionary sentiment; the public appreciates knowing who is taking action and how through clear transparency.

Policy

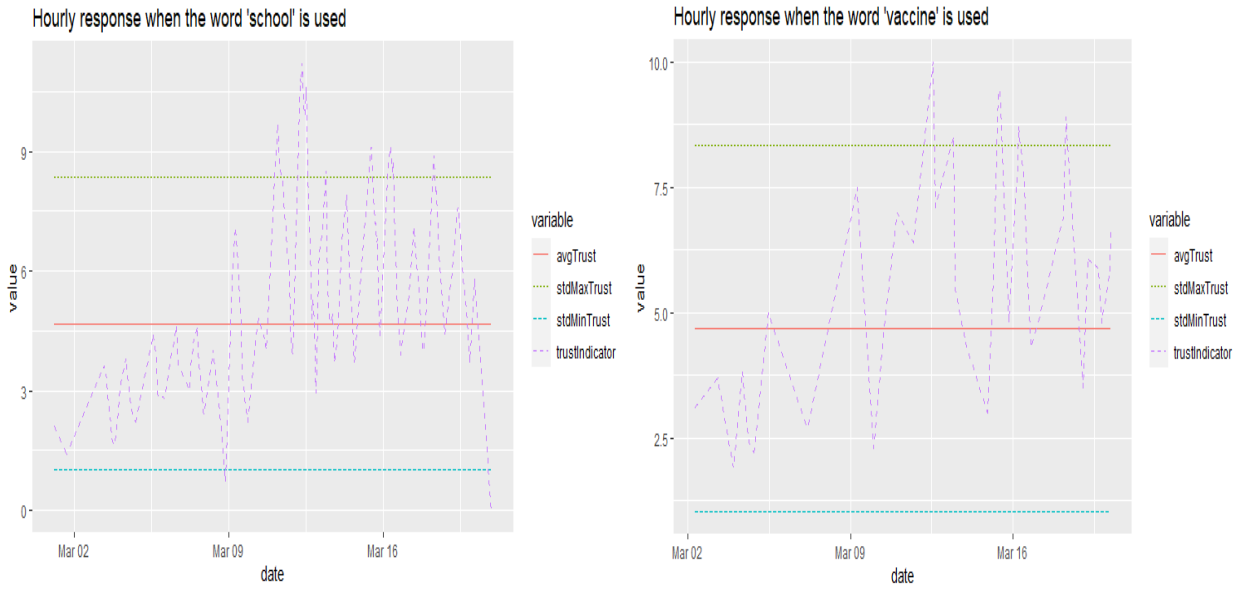
In terms of policy, releasing consistent updates on lockdown measures with explanations for how this prevents the spread of COVID-19 is particularly effective in generating an increase in sentiment. The analysis also showed taking immediate action on implementing lockdown procedures generates a positive sentiment, so if/when another pandemic occurs, or if there is a second wave of COVID-19, this is something that the government should consider.

Healthcare

Finally, with regards to healthcare, the government should release funding budgets that relate to how much will be spent on developing and eventually distributing a vaccine at the earliest time possible. The reasoning is the public has a positive trust and sentiment response to being informed on the status of a vaccine, as it allows them to feel more informed. Since the vaccine dialogue was in the very early stages from March 1st to 20th, it is assumed that few monumental developments were made to cause a significant increase in vaccine trust sentiment (see Exhibit 12 & 13). Relative to

'school', where there are significant spikes on sentiment change above one standard deviation, 'vaccine' is within one standard deviation of max and min trust values in all but 5 instances.

Exhibit 12 and 13: Hourly Responses to 'Vaccine' and 'School'



It is also recommended that the government should be transparent with the number of cases and provide a trajectory of where that may go, based on COVID-19 test data. This is because the public prefers to be informed on the developments of the situation, rather than hearing nothing. Additionally, guidance should be provided for how to prevent infection and the actions that individuals can take, as this makes the public feel safer, which increases sentiment.

The purpose of these recommendations is to increase sentiment; however, the model can also be used to minimize the decrease in sentiment in the case that the current situation worsens. This means that, for example, if schools will have to be moved to smaller class sizes or online schooling, the public wants to be informed on this, rather than not knowing what is going to happen. With the model, the government is able to expect shifts in public sentiment. If the government has to release bad news about case numbers, they could counteract the expected negativity with positive rhetoric about vaccine data. As the situation evolves, constant updates on how this affects things such as schools should be talked about.

Risks and Mitigations

A key limitation of the model that was developed is that it is based on 20 days of data, which means that if the pandemic continues for an extended period of time, it is difficult to know how accurate the current model will be. To combat this, one of the benefits of the current model is that it is dynamic, meaning that it can be fed new data and be updated, which may actually increase its overall effectiveness. The next steps that should be taken would be to update the model with the most

recent 100 days of data and see how the coefficients on the key words selected within the regression have changed. Public sentiment around 'lockdown' and 'school' may not be as high as it was in the period from March 10th to March 20th, so it would be beneficial to see the reaction of the public surrounding more recent government actions.

Lastly, it is important to note that this model and these recommendations should not be used with the sole purpose of increasing sentiment among the general public. The most important thing is public safety. When bad news arises, as there will likely be more of, if it is released poorly, there will likely be a fall in public sentiment, as it could cause panic and a lack of trust from the public towards the government in the long term. By following the model that has been built and the recommendations that have been made, it is clear that the most effective way to present the message is directly and transparently, as this will minimize the decrease in sentiment. This is because it will not cause people to panic and people will still trust the actions that the government is taking. Overall, by monitoring sentiment of select words through the entirety of the COVID-19 Pandemic, governments will be able to influence public sentiment without withholding information from their people. Thus, in future waves and pandemics, the model can continuously improve with more data to a point where there is a specific playbook that governments learn to use through predictive sentiment.