A Survey of the Effect of COVID-19 on Social and Economic Communication

Zion Steiner, in collaboration with Chetan Birthare and Oliver Quinton

INTRODUCTION

COVID-19 has changed the daily lives of billions of people. Financial, medical, political, and social institutions are still adapting to the effects of pandemic. As people and organizations navigate this situation, they will include COVID-related discussion in their communications. These come in the form of medical documents discussing the virus, corporate response briefings, an opinion shared via Twitter, and more. The goal of this project is to explore how COVID-19 is affecting a few specific varieties of text communication. To this end, we gathered corporate earnings reports, Tweets, and Reddit comments to analyze. The body of this report details the dataset collection process, methods of analysis, and insight drawn.

PART I - EARNINGS FILINGS

The SEC requires that most publicly traded companies release quarterly earnings reports called 10-Q filings. In the documents, companies discuss their performance over the past quarter. They will address whether or not the company hit their price guidance or not, and what factors led to that outcome. Stock prices often react to insights in these documents. Professional traders read these filings and trade accordingly as soon as the reports are released. If NLP can be used to automate the summarization of key points as input for a market direction classifier (up/down), trades can be entered early and take advantage of any reactionary market movements the document release causes.

In this section, we analyze these documents in the context of COVID-19. We identified the industrial sector distribution of COVID-related terms being mentioned in 10-Qs. We also attempted to measure the sentiment of 10-Q filings mentioning these terms in comparison with the sentiment of all filings. We then test the significance of computer filing sentiment as a predictor for the next-day change in stock price.

DATASET

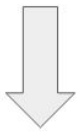
The SEC filing dataset was collected from two sources: sec.gov for the 10-Q filings, and Polygon for extraneous company info, including pricing information. After collecting all filings, filings from companies with a market cap under $1 billion were removed from the dataset. This was to reduce the amount of text processing, so any information presented here cannot be

extended to companies of a smaller size. The dataset entries from large companies (over $1 billion) were then joined with extraneous info retrieved from the Polygon API. This includes the company sector and closing stock prices ranging from Jan 1, 2019 to April 20, 2020.

Next, the filings were cleaned. Because the filings are stored in XML format, most of the cleaning steps involved removing XML tags, links, and other markdown-specific text. The cleaned filings were then "diffed". This idea was taken from this project. Given two documents, diffing is the process of filtering out content that appears in both documents. This distills the new information introduced in a document and removes all repeated boilerplate language. This example shows the result this has on 10-Q filings:

CAUTIONARY STATEMENT

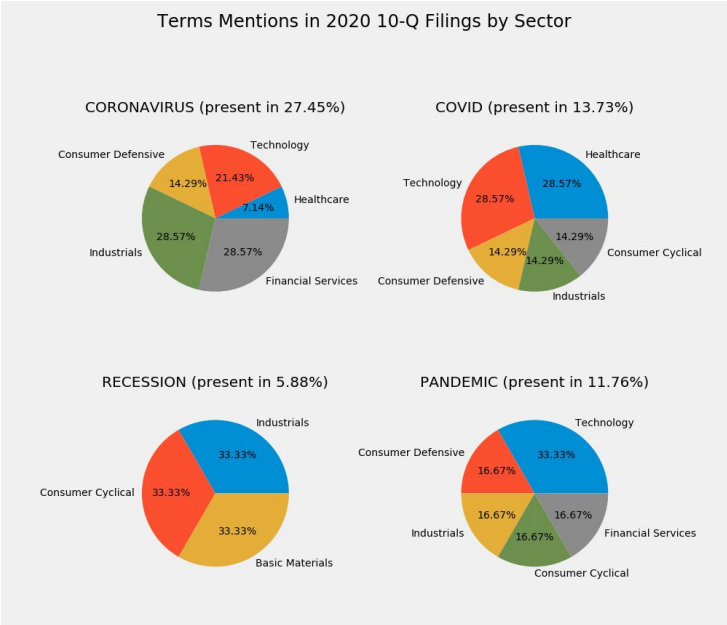
"This Report on Form 10-Q contains, or incorporates by reference, certain forward-looking statements within the meaning of the Private Securities Litigation Reform Act of… This quarter's earnings were 2% higher than expected."



"This quarter's earnings were 2% higher than expected."

ANALYSIS

We started analyzing this dataset by searching for four COVID-related terms within the filings corpus. The terms used are *covid, coronavirus, pandemic, and recession*. If a filing mentioned one of these terms, the filing was labeled as having used that term. The pie chart below shows the term-mentions by sector.



The most common term mentioned is *coronavirus*, which is present in 27% of all 2020 10-Q filings. Among these, the industrials and financial services sectors mentioned *coronavirus* the most frequently. However, absolute sector mention frequency may not be meaningful if some sectors have more companies than others. For example, if there are 10x as many tech companies as healthcare, and healthcare companies mentions a term more frequently than the tech sector does, then the healthcare sector is mentioning that term over proportionately.

To check this, we can adjust for how many companies from each sector have filed yet in 2020. We do this by dividing the number of companies in each sector that mention a term by the total number of companies from that sector that have filed already. For *coronavirus*, we have the following values:

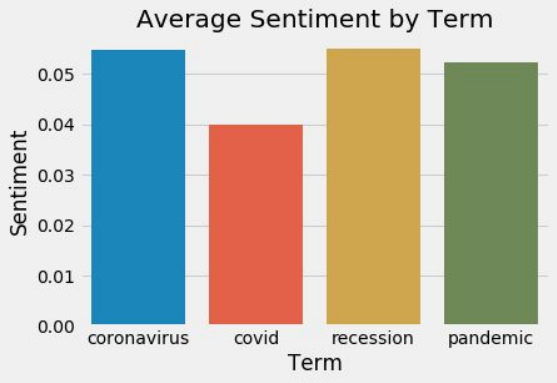
Sector	**% of sector mentioning CORONAVIRUS**
Tech	0.273
Industrials	0.364
Healthcare	0.200
Consumer Defense	0.500
Financial Services	0.444
Consumer Cyclical	0.000
Basic Materials	0.000
Communication Services	0.000

This means that although the same number of industrials and financial companies mentioned *coronavirus* in 2020 (28.57%), financial companies mention *coronavirus* more frequently than industrials do on average. Take note that these are the value counts for the sectors of filings published in 2020, so the sample size here is too small to make any concrete claims. The fact that this is a developing situation and that more 10-Q filings are being published every week are also important to consider.

Sector	**# 2020 Filings**

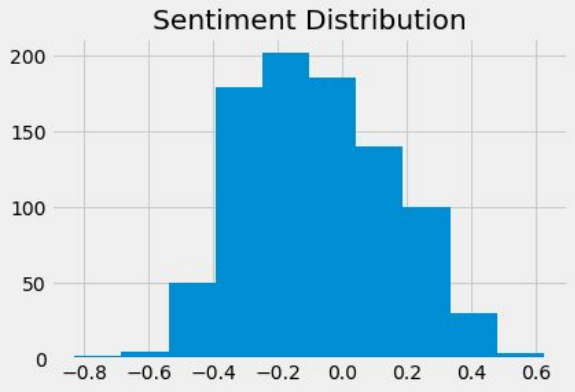
Tech	11
Industrials	11
Financial Services	9
Consumer Cyclical	6
Healthcare	5
Consumer Defense	4
Basic Materials	3
Communication Services	2

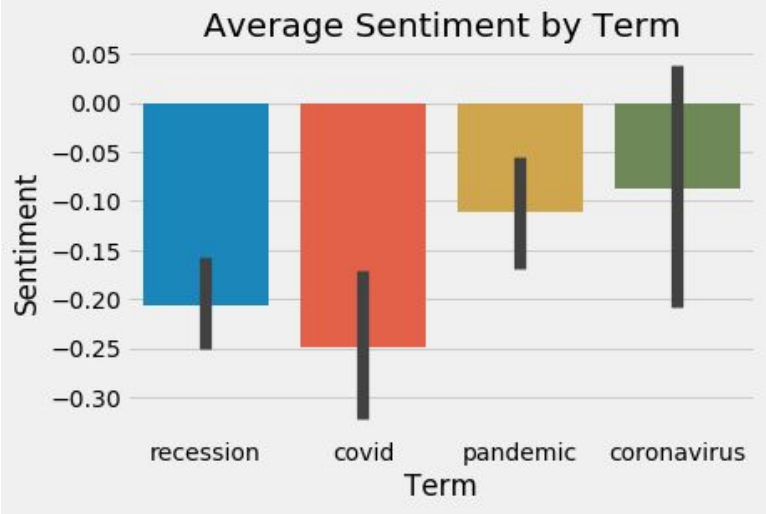
We also wanted to compare the average sentiments of 2020 filings mentioning one of these terms. We started by using the TextBlob sentiment model, but this didn't work very well. On the 10-Q filings, it gave neutral scores (around 0) for many known negative statements. This is probably because the filings use language that TextBlob's naive bayes classifier didn't see often enough to be able to work well.



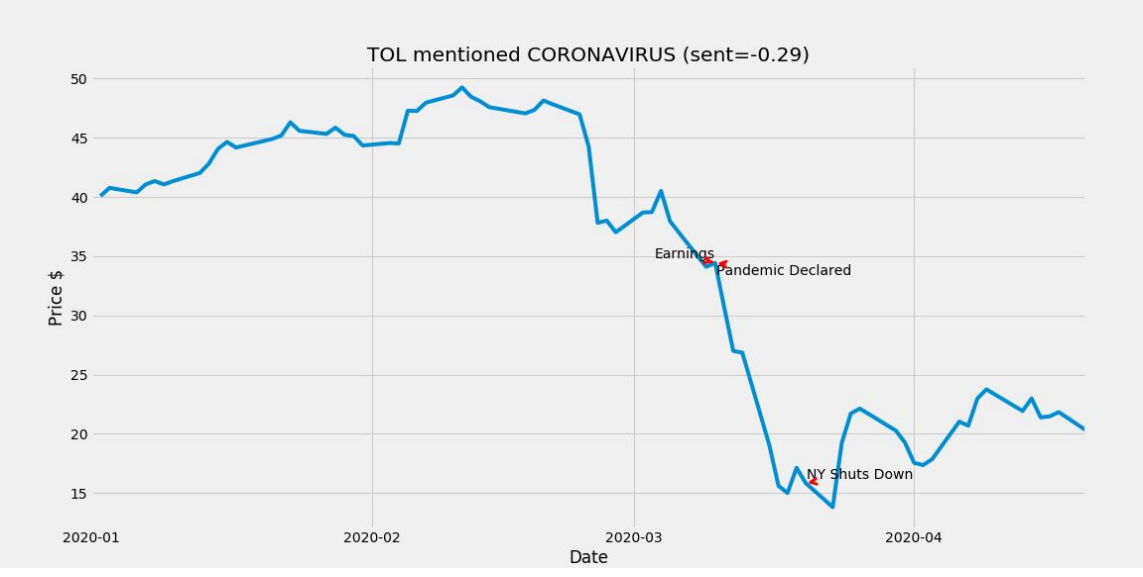
To account for this, we found a dictionary of positive and negative connotation words found often in 10Q filings. This dictionary was put together by Loughran and McDonald in 2011, and was built with 10-Ks in mind, which are very similar to 10-Qs. The one problem with this dictionary is that it has 6.65x more negative words than positive ones. I used sklearn's countvectorizer to count how many times each term appears in each filing, then calculated sentiment as [EQUATION HERE].

After doing this, the sentiment histogram was still centered further to the left than would be expected. The overall high sentiment was 0.38 vs a low negative of -0.97, which doesn't make sense considering the SPY index grew 31.22% in 2019 (sentiment should be great). This is most likely due to there being many more negative words than positive in the sentiment dictionary. To fix this, we added a weight of 2.0 to each positive word counted. This centered the sentiment distribution and gave more sensical results. Interestingly, although the term *coronavirus* was the most frequently mentioned among the four terms, it has the largest variance in sentiment. This is seen in the black confidence interval overlaying the green bar in the figure below.





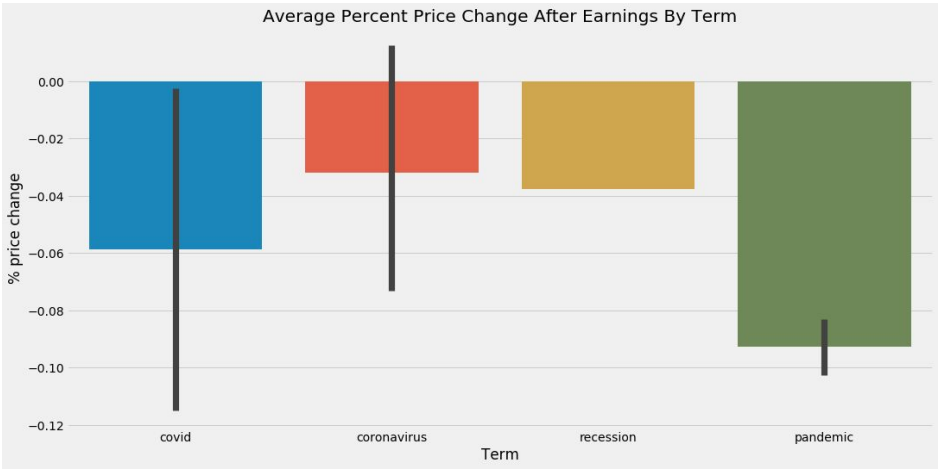
The next step was to see if the relationship between 10-Q sentiment and the next-day change in stock price is significant. If it is, then this method of sentiment analysis has promise for developing an automated trading strategy. If not, then more work needs to be done to engineer a sentiment feature that contains useful information about price direction. The picture below illustrates the effect specific events can have on a stock's price.



It's not straightforward how to quantify the relationship of each of these events with the stock price on the following day. Even if the prices for all companies mentioning *coronavirus* drop after their earnings report, we can't know if that drop is because of an already present bearish-trend in the stock, or if it really was the corona-mentioning 10Q filing. With this in mind, here are the average price changes for filings mentioning each term, found by using this to measure price change:

[EQUATION]

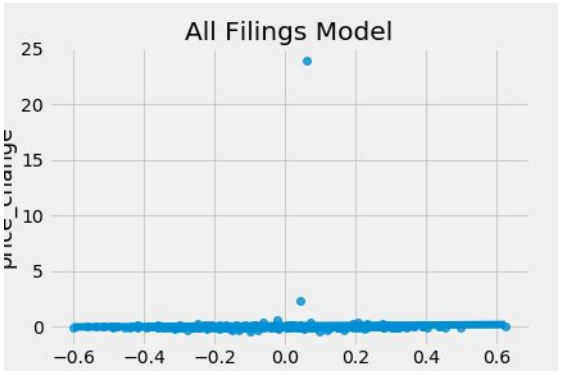
The barchart below shows the results of this test using the 2020 term-mentioning 10-Qs. Once again, the data associated with the term *coronavirus* has the most variance, despite it having the largest sample size. The only terms whose confidence intervals didn't include positive results are pandemic and recession. This means that we cannot say for certain that the inclusion of the term *covid* or *coronavirus* in a 10-Q report doesn't cause an increase in stock price, on average. While this is not likely the case, it would be interesting to identify which companies are benefiting from this situation. Considering Netflix recently announced 16 million new subscribers in their Q1 2020 10-Q, it's definitely possible.

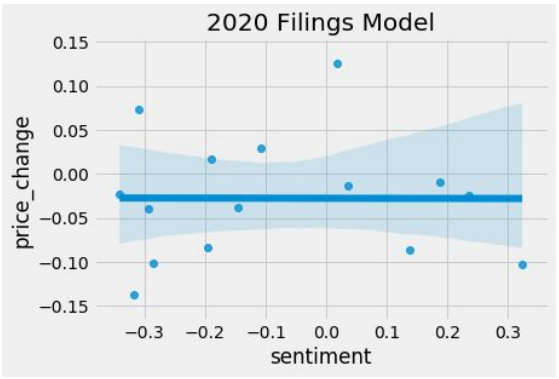


Before concluding that there is no meaningful relationship between calculated 10-Q sentiment and price change, we constructed an ordinary least squares model to do inference on the relationship.

[EQUATION]

The overall correlation between sentiment and price change is -0.002922, with a p-value of 0.529 on the B1.





Given the results of this research, there are two routes going forward. Either more work can be put into engineering a domain-specific sentiment metric for 10-Q filings, or we can test our luck at throwing a neural network at the problem.

PART II - SOCIAL MEDIA

We were also interested in evaluating the impact of COVID on social media communication. Specifically, we analyze Tweets and Reddit comments.

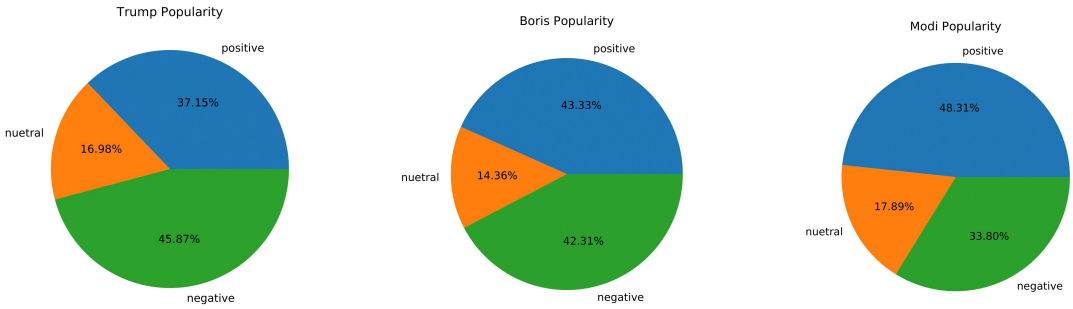
DATASET

1. Collected tweets for keyword “economy” from Illinois, Florida, New York, and California from April 11th to 20th
2. Collected data for USA, UK and India for search keyword as their respective leader’s name (Donald Trump, Boris Johnson, and Narendra Modi) from 13th to 20th april
3. Collected data for covid cases for the countries and states mentioned above

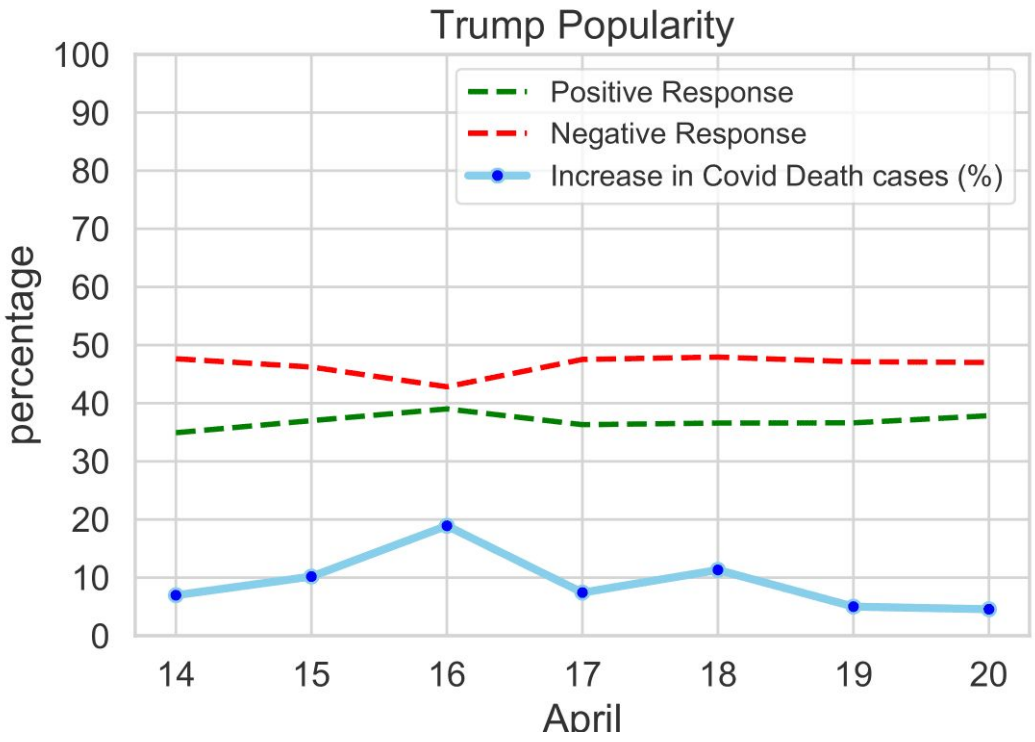
Each tweet was cleaned using NLTK and lemmatized with Spacy.

ANALYSIS

Several NLP methods were used to analyze the tweets. First, Vader was used to estimate the sentiment of tweets involving country leaders. This info can be helpful for gauging public opinion on a leader’s response to the pandemic. The results:



Of the three leaders tested, Modi was looked upon most favorably by the netizens of Twitter, while Donald Trump was mentioned in the most negative terms. We were also interested in measuring sentiments over time. The figure below shows the ratio of positive/negative tweets mentioning Donald Trump against the US COVID death count. While it does appear that the slight majority of responses are negative, the ratio does not change much over time. The ratio also appears to have no relationship with the increase in COVID death rate.

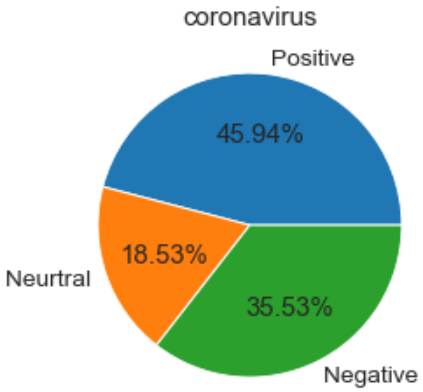


Next, the Tweets were used to train an updated Word2Vec, followed by a reduction to 2-dim using PCA. The results from the *Trump* label tweets were graphed. In the figure below, several term groupings can be identified. Terms whose encodings are located nearby tend to have similar meanings.

I'll leave it up to the reader to interpret the significance of these groupings.

REDDIT DATA

The results from sentiment analysis on Reddit comments did not work out very well. Many comments related to negative-connotation term (ex. *coronavirus*) had overall positive sentiments, which did not make sense. We could not identify why Vader was working this way. Because of this, the data collection and analysis sections will be omitted. However, they were very similar to what was done for the Twitter analysis.



CONCLUSION

This project was exploratory in nature. Our main goal was to learn to use NLP tools to automate the digestion of the massive amounts of new information being published about recent events that the world is being flooded with. This experience has familiarized our team with some of the challenges of working with text data and the difficulty in telling a computer how to read human language. Overall, we can conclude that sentiment analysis is a hard task, and requires specialized models to work well.

https://github.com/zionsteiner/CS5830-Final-Project