CMPT 459 - Course Project

A Data Mining Study on COVID-19 Pandemic Growth & Related Social Media Dynamics

2020/08/10

Author	Ya Qi (Jerry) Liu
Student ID	301255583
Contact	liuyal@sfu.ca
Code Repo	GitHub Link
Issue Date	August 10, 2020

Table of Contents

Table of Tables

Table of Figures

1. Motivation & Background

Figure 1: COVID-19 and Twitter (GETTY IMAGES)

Coronavirus disease 2019 or COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. First identified back in December 2019 in Wuhan province, China, the COVID-19 virus has resulted in nearly 20 million confirmed cases globally (as the writing of this report) [2]. Amidst the COVID-19 crisis, social media usage on platforms such as Facebook, WhatsApp, Twitter, and etc. has surged significantly [3]. As the general population relies heavily on various social media platforms to gather the latest information in regards to the pandemic, resulting in an unprecedented amount of content and information.

An interesting data mining topic to focus on for the COVID-19 pandemic is to determine the relationship between COVID-19 related trending topics and sentiments on the social media platform Twitter, with the number of reported confirmed cases for a given country over a period of time. Topic modeling and Tweet sentiments classification could be a useful measurement for determining the general attitude expressed for a given population in regards to COVID-19. Since the virus originated from China, and only after a three-month period where it quickly spread across North America resulting in to more than 20 million confirmed cases globally [2]. The changes in daily trending topic of interests and tweet sentiments due to the influx of confirmed cases as COVID-19 begins to spread in a particular country or region would be very interesting to discover.

The measurement of relationship between the growth of the pandemic and social media topics and sentiment can help determine how a general population express their opinions, concerns, and general awareness throughout such a global event. Sentiment information and topic modeling can be used by government bodies around the world to determine the general population’s level of attitude towards the pandemic as it grows, and how to issue proper procedures and restrictions with appropriate messaging and policy decisions in times of crisis for future global events and pandemics.

2. Problem Statement

Social media platforms such as Twitter provide access to unprecedented amount of content and information. Information spread on such platforms can strongly influence a population’s behaviour and general attitude towards global level events, such as the COVID-19 pandemic. The country of focus for this data mining project will be the United States, since the US has the highest number of confirmed COVID-19 cases globally, along with a fairly large frequent Twitter user population. This data mining project will also focus on English language tweets as English language tweets are better suited for the natural language processing methods developed for topic modeling and sentiment analysis.

Using various data mining techniques, the relationship of two particular COVID-19 related datasets can be extracted, processed, and analyzed. The first dataset is the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [2]. The CSSE dataset contains daily confirmed COVID-19 cases from all countries and is an excellent aggregated data source. The second dataset is the COVID -19 Tweet ID dataset gathered by University of Southern California Ph.D. student Emily Chen [4]. This Tweet ID dataset contains COVID-19 related tweet IDs collected as early as January, 2020, and is updated frequently. Sampling and Hydration methods were performed on the IDs to extract a limited number of the original tweet text for topic modeling and sentiment analysis.

In order to quantify a linkage between the on-going COVID-19 pandemic and daily trending Twitter topics and tweet sentiments, a measurement must be defined to represent this relationship. Using NLTK topic modeling techniques, COVID-19 related tweet topic evolution can be identified and modeled. And for sentiment analysis, a Naïve Bayes Classifier was trained with the Twitter for Sentiment Analysis (T4SA) dataset [5], which is an extensive set of tweet sentiment training data. The change in COVID-19 related topic and semantics can be classified into positive, neutral, and negative tweets, and be compared with the daily increase of number of confirmed cases in the US to derive a conclusion and would be a suitable quantification for this data mining task.

3. Datasets

Two COVID-19 related datasets are selected for this data mining project. The first dataset is the COVID-19 Data Repository (**Link**) provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [2]. The dataset includes information relating to number of confirmed cases, deaths, and recovered cases which are organized by country/region and by date. Reported cases are presented at the province level in China, county level in the US, state level in Australia and Canada, and at the country level otherwise. The CSSE dataset is composed of a variety of aggregated data sources such the DXY - an online platform run by members of the Chinese medical community, World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDC), and many other data sources around the world. A snippet of the dataset can be found in Table 1 below.

FIPS	Province State	Country Region	Confirmed	Deaths	Recovered
45001	South Carolina	US	88	0	0
08003	Colorado	US	7	2	0
22001	Louisiana	US	654	32	0
51001	Virginia	US	1031	14	0
16001	Idaho	US	1166	22	0

Table 1: CSSE Dataset Snippet

The second dataset is the COVID-19 Twitter ID dataset (**Link**) gathered by Emily Chen from the University of Southern California, which compose of tweet IDs acquired from Twitter stream which are related to COVID-19 chatter. The dataset was collected beginning from January, 2020. Tweet IDs are collected specifically using a selection of keywords related to COVID-19. Each tweet ID can be hydrated with the Tweepy API to obtain the original text for processing. The top daily word patterns (terms, bigrams, and trigrams) can be derived from the hydrated tweets to show frequency of occurrence of the patterns and for topic modeling and sentiment classification. The data collected captures all languages with higher prevalence for English language tweets (65.63% of total tweets). A snippet of the hydrated and filtered tweets can be found in the table below.

Index	ID	Create At	Raw Text	User
2	1242138595693070000	Mon Mar 23 17:18:13 +0000 2020	@Angrysausagetv @SkyNews I am not an expert BUT most experts predicted a 2nd wave in China when they reopened the qâ€¦ https://t.co/12kkrHuXBN	vinceandpen
4	1242138641725410000	Mon Mar 23 17:18:24 +0000 2020	RT @narendramodi: One thing I specially requested all media houses to do is to keep reiterating the importance of social distancing and beiâ€¦ One thing I specially requested all media houses to do is to keep reiterating the importance of social distancing aâ€¦ https://t.co/PH94Lc83Vu	rohitsambhar
6	1242139505617960000	Mon Mar 23 17:21:50 +0000 2020	@SpeakerPelosi @HouseDemocrats https://t.co/11TICX2HWS Pelosi Schumer and Democrats tanking the economy https://t.co/C81pbjsHG9	sharonaquino10
30	1255182956546660000	Tue Apr 28 17:11:51 +0000 2020	RT @business: BREAKING: Trump plans to order U.S. meat plants to stay open amid the coronavirus pandemic https://t.co/DzgdIBTI36 https://t.â€¦ BREAKING: Trump plans to order U.S. meat plants to stay open amid the coronavirus pandemic https://t.co/DzgdIBTI36	lavendar_l

Table 2: Filtered & Hydrated Tweets Dataset Snippet

4. Architecture & Pipeline

The data mining architecture follows the basic Knowledge Discovery and Data Mining (KDDM) process, where the entire process is split into five phases as shown in Figure 2. Staring with data collection for obtaining the raw datasets, then data preprocessing to prepare the datasets for transformation. Data mining methods are then performed on the transformed dataset and the results are visualized for knowledge comprehension. A detailed diagram of the data mining pipeline can be found in Figure 3.

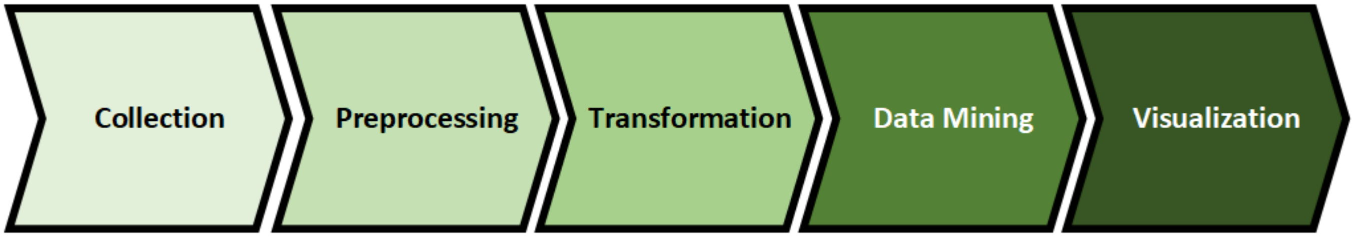

Figure 2: KDDM Process Pipeline

During the data collection phase datasets from the CSSE COVID-19 Git repository and the Daily COVID-19 Twitter ID GitHub repository are queried using HTTP requests to populate a local database. The preprocessing of the collected raw data prepares the data for transformation by data filtering and cleaning of missing entries. CSSE confirmed cases dataset was cleaned up by removing missing data entries and filtering out non-US related records. Since there is an extreme amount of tweets within the data repository and in order to make the data mining process simpler, the ID's are sampled randomly by selecting 1000 tweet ID from each date. Then the original texts are obtained using tweet hydration with the Tweepy API, with all non-English language tweets filtered out.

Data transformation was performed by either data aggregation or text tokenization to prepare the datasets for modeling and visualization. In the final data mining process, the transformed data is input through trained models to extract useful information such as patterns, trends, and classifications. Lastly, for data visualization the final processed datasets were integrated together and visualized using various tools such as Python Matplotlib library and Excel. The visualization process can help interpretive and evaluate useful and knowledgeable information about the datasets.

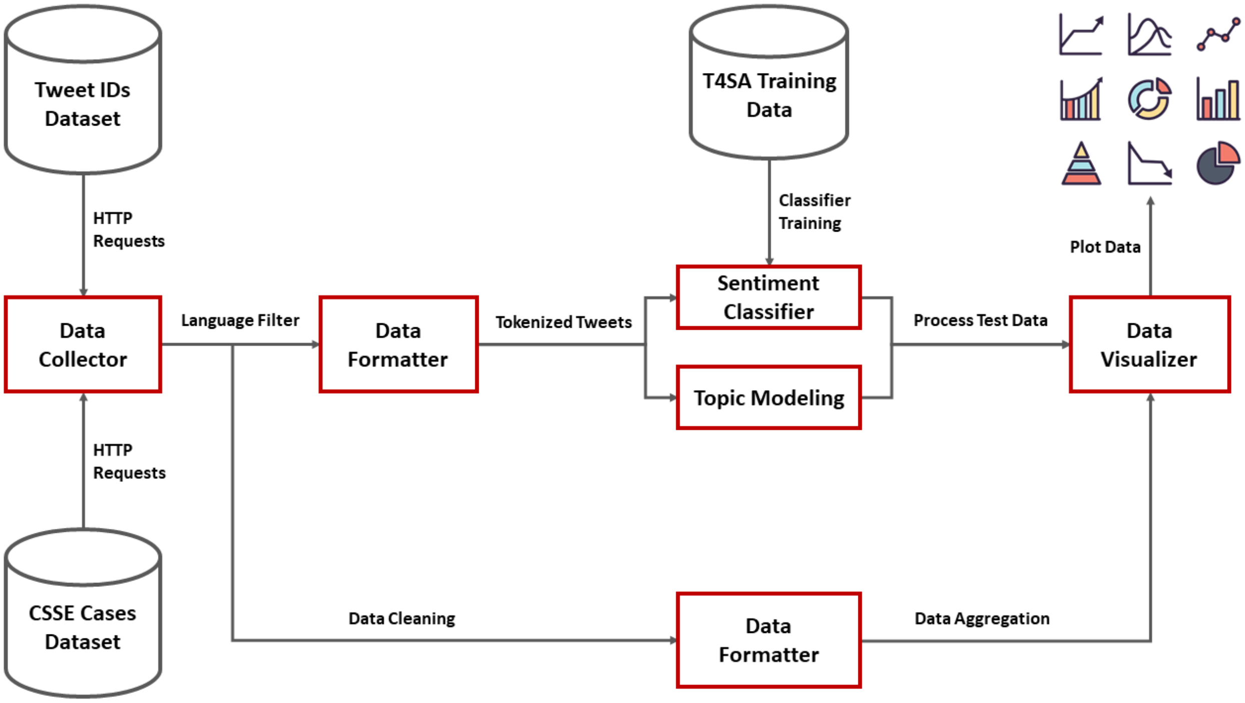

Figure 3: Data Mining Pipeline

5. Methodology

5.1 Location Data Processing

The CSSE daily confirmed cases dataset collected by John Hopkins University provides an aggregated data source for number of confirmed cases for all affected countries. This dataset is queried directly from the GitHub repository using HTTP requests, where it is then stored locally into a database. During processing, the data is first converted into data frames using Pandas - a Python Data Analysis Library, before further data aggregation is performed. Since the country of focus is the United States, and daily cases are the subject of focus, the data frame is aggregated using the SUM function on the number of confirmed cases, and grouped by the reported date. Which results in a transformed dataset containing only daily confirmed cases for the United States. This dataset will be used to gauge the growth of the COVID-19 pandemic in the US throughout the time period.

5.2 Twitter Data Preprocessing

The Daily Tweet IDs data repository contains tweet IDs relating to COIVD-19 chatter collected starting from January, 2020. After tweet texts are hydrated using the tweet IDs and the Tweepy API, all non-English tweets were filtered out and removed. The resulting raw tweet text are then formatted before inputting through topic modeling and sentiment classification functions. Raw tweets are first tokenized by spiting the text into sentences and the sentences into words, and lowercasing each word while removing all punctuations, emoji, and URLs. For each word in the set of tokens, words with less than 3 characters are removed along with stop words. Words are then lemmatized to change from third person to first person and verbs in past and future tenses are stemmed to reduced to their root form. An example of this process can be seen in Figure 4 below. Once the tweets are tokenized, they are ready to be put through the sentiment classifier and topic modeling algorithms.

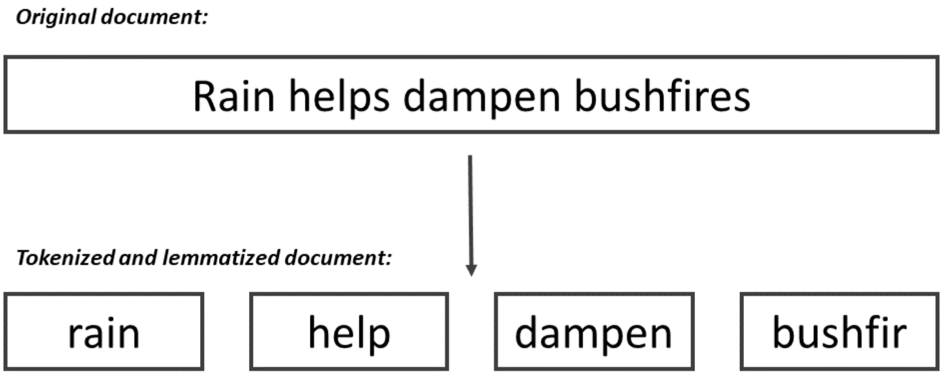
Original document:
Rain helps dampen bushfires
Tokenized and lemmatized document:
rain
help
dampen
bushfir

Figure 4: Tokenization Example

5.2.1 Tweet Topic modeling

Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents [7]. A Latent Dirichlet Allocation (LDA) model was build using a dictionary of words constructed from the tokenization process for auto detecting and interpreting topics within the tweets. Using the Gensim (Generate Similar NLP package) LDA model, and a dictionary of the tokenized tweets, a corpus was rebuilt. This corpus contains a word count vector for each tweet, which consists of the frequencies of all the words in the vocabulary for that particular tweet. The Gensim LDA model was able to generate topics for each collection of dates in the dataset. Each topic for each date is a set of keywords, each contributing a certain weight or importance to the topic. An example of the results for January 21, 2020 can be seen below. Using a performance evaluation by sample document using LDA TF-IDF (Term frequency to Inverse Document Frequency) model a dominant score can be determined for a particular topic. For this particular topic example below, the score was determined to be 0.9527 or 95.27% probability of occurrence.

$$0.149 * "case" + 0.126 * "confirmed" + 0.065 * "disease" + 0.065 * "control" + 0.053 * "new"$$

5.2.2 Tweet Sentiment Classifier Modeling

Sentiment analysis is a common Natural Language Processing (NLP) task, which involves classifying texts or parts of texts into a pre-defined sentiment. In order to create a classifier that can detect tweet sentiments to predict whether a tweet is negative, neural, or positive, a Naive Bayes classifier was built to perform text sentiment classification. The Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The core of the classifier is based on the Bayes theorem [6]. The Twitter for Sentiment Analysis (T4SA) training dataset was used as the training dataset for the Naive Bayes classifier, which contained 371,341 positive tweets, 629,566 neutral tweets, and 179,050 negative tweets (T4SA sentiment distribution shown in Figure 5). The T4SA dataset were tokenized and split into positive, neutral, and negative categories each with a distinct label (POS, NEG, NEU). The final training tweet sentiment detection model contained a 3-way sentiment polarity detection, which was able to classify the daily COVID-19 related tweets into positive, neutral, and negative tweets.

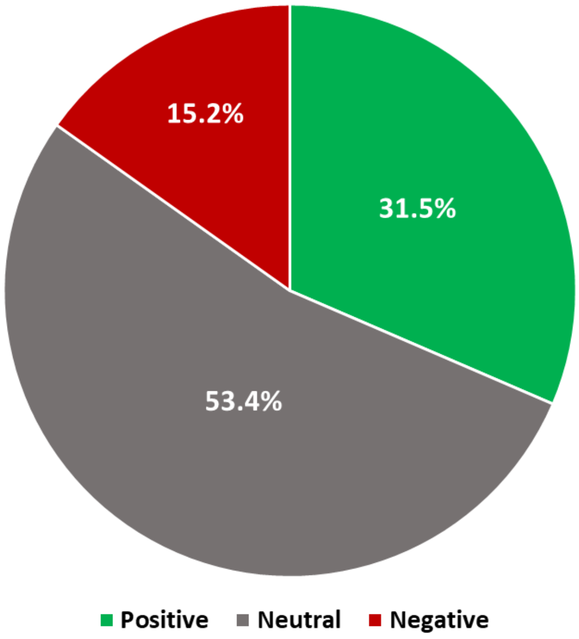

Figure 5: T4SA Training Data Sentiment Distribution

6. Results

Worldwide dissemination of the COVID-19 Virus has been extremely rapid, at the time of issue of this document a total of 18,079,136 confirmed cases and 689,347 deaths had been reported across 215 countries. After simple data processing of the datasets, a rapid exponential growth of daily confirmed cases for the United states and globally can be visualized over the last couple of month along with the daily number of collect COVID related tweets. As shown in the figure below, this provides a clear picture for the severity and contagious effect of the COVID-19 pandemic in relation to the amount of COVID-19 related tweets generated.

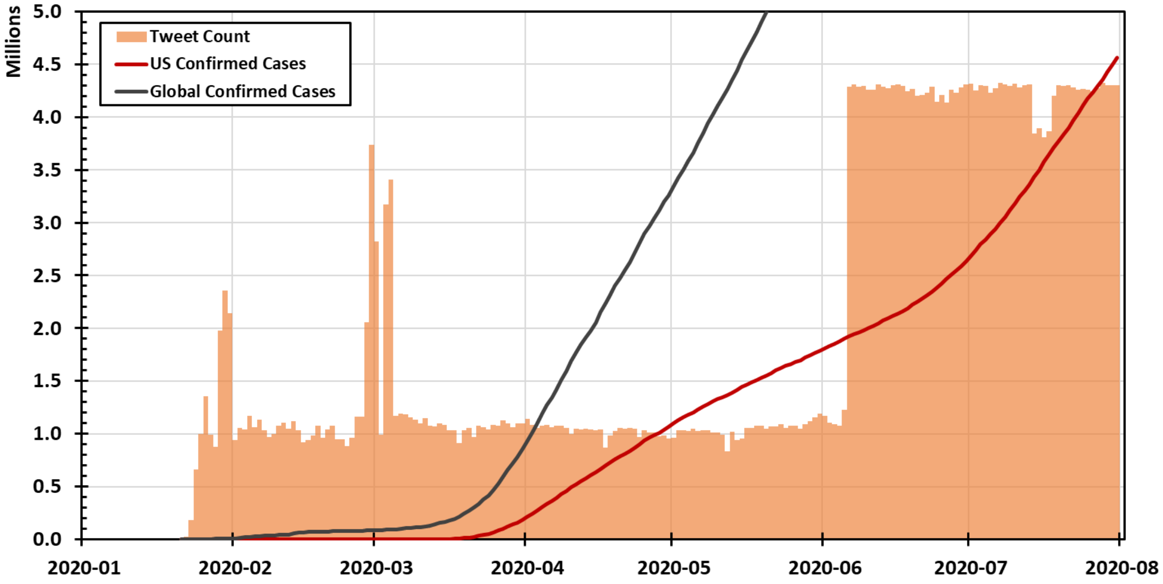

Figure 6: Tweet Count & Daily Confirmed Cases [9]

6.1 Tweet Sentiment Analysis

In total, over 193 days, 126,838 Tweets were sampled, collected, and analyzed. Using the trained tweet sentiment classification model, the daily COVID-19 related tweets were classified into three classes: negative, neutral, and positive. For each date, the frequency of occurrence of classified tweets for each class was plotted in Figure 7 below. Overall as depicted from graph, negative tweets have the highest daily frequency count, followed by neutral tweets and positive tweets. Trend for the positive and neutral tweets has stayed relatively the same throughout the timeline, but trend for the negative tweets were shown to spiked during specific periods in time.

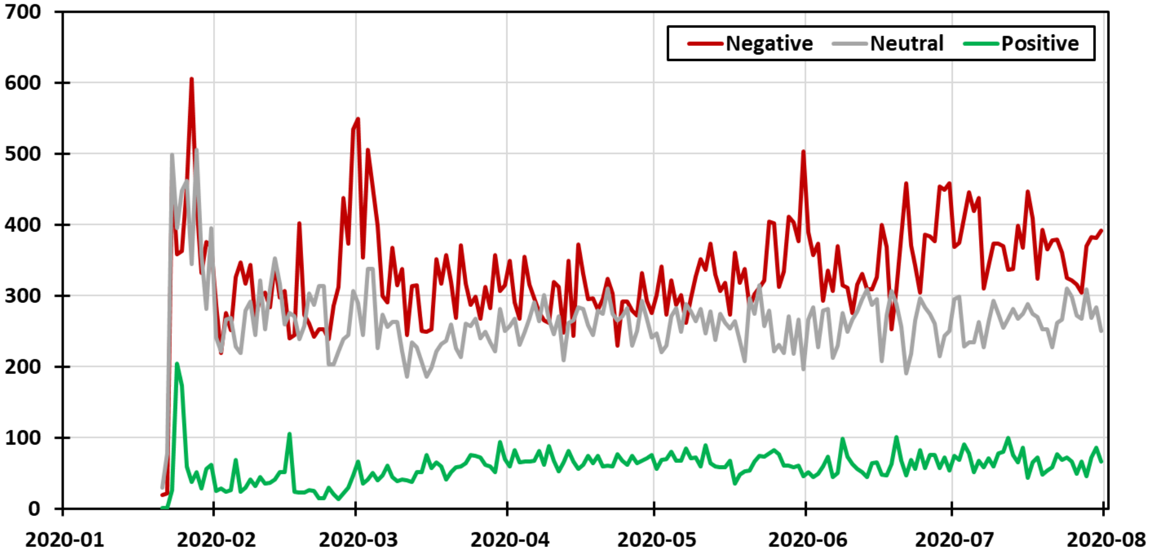

Figure 7: Daily Tweet Sentiment Count Over Time

Since the general sentiment for COVID-19 topics were considered to be negative, it was expected that the overall sentiment would favour negative/neutral classes. From the class distribution chart in Figure 8, it is shown that 50.6% of all collected COVID-19 tweets were classified as negative, 40.3% were classified as neutral and only 9.0% were classified as positive. Which validates the initial assumption about COVID-19 related tweets favouring negative sentiments.

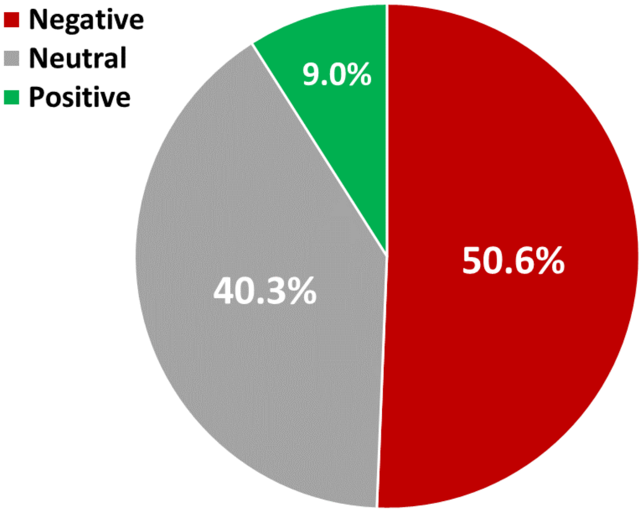

Figure 8: Tweet Sentiment Detection Distribution

After integrating the daily number of confirmed cases with daily tweet sentiment count, the trend of the two datasets could be better observed. As COVID-19 progressed and its seriousness became evident, the growth of negative sentiment tweets was also apparent. This relationship could be observed during two different time frames: the first near the end of the month of February, and the second from mid-May onwards. From the graph shown in Figure 9, the first spike in negative sentiment tweets was near the end of February where the COVID-19 growth starts to manifest rapidly in areas outside of China, according to the COVID-19 timeline [8]. The second increase in negative sentiment tweets occurs from mid-May onwards, which can be correlated to the explosive growth within the United States. As the number of confirmed cases increased from around 1.5 million to 4.5 million in the span of two months, so did the frequency of negative sentiment tweets. The day to day increase in number of negative sentiment tweets can be observed as the severity and number of confirmed cases increased due to COVID-19 pandemic.

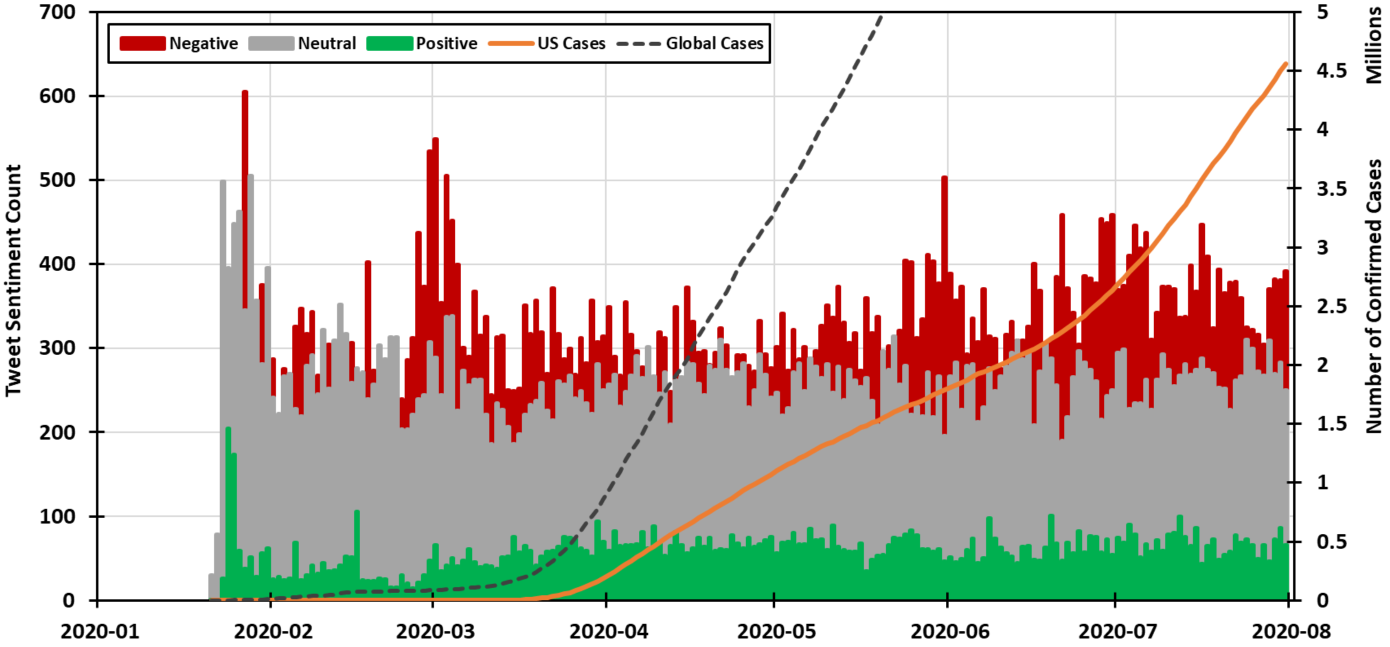

Figure 9: COVID-19 Growth and Sentiment Frequency

6.2 Tweet Topic Modeling Analysis

To visualize the top 10 most common words of all tweets from the daily tweet dataset, a word count distribution plot was created (Figure 10). Overall, there were 17,108 unique words in the collected set of tweets, with the word **covid** having the highest frequency of occurrence. The top 10 words of the tweet dataset could be interpreted as a summery of the general topic for the collective COVID-19 related tweets. Words with higher frequency of occurrence have higher weight in the overall topic model.

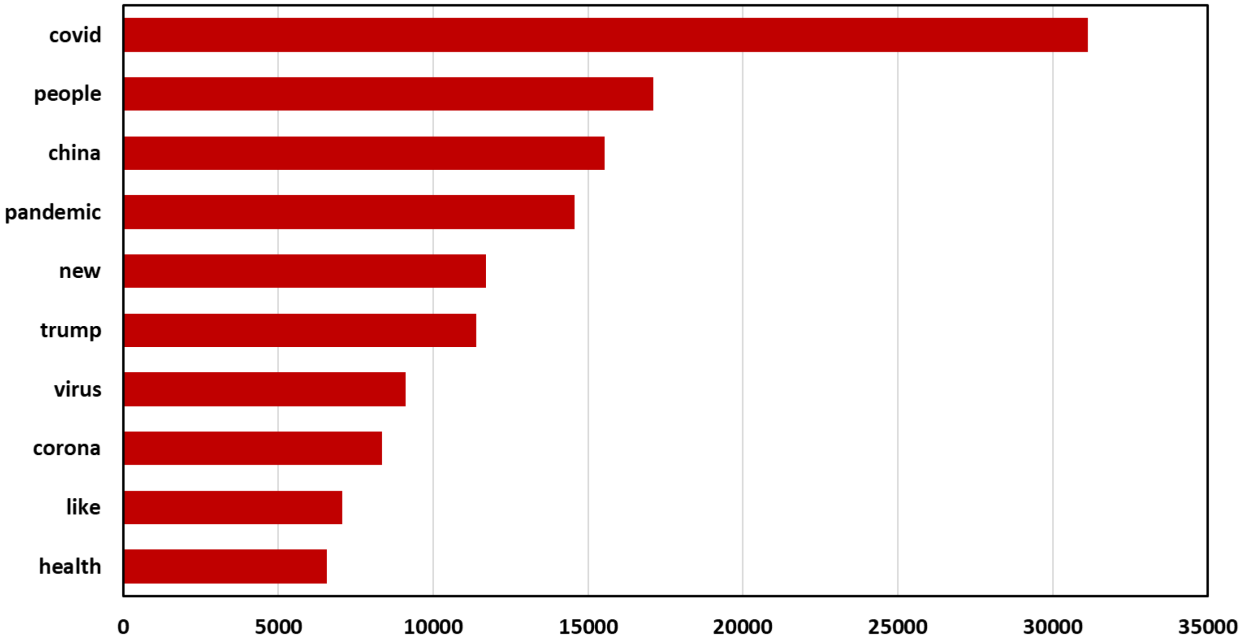

Figure 10: Total Word Count Distribution

A word cloud was also generated to further visualize the frequency of occurrence of the top most common words as shown in figure below. Words with high frequency are shown to be brighter and larger while words with lower frequencies are shown to be darker and smaller. From the word cloud, the words **covid**, **people**, **china**, **pandemic**, and **trump** were shown to have the highest frequency of occurrence from the collected tweets.

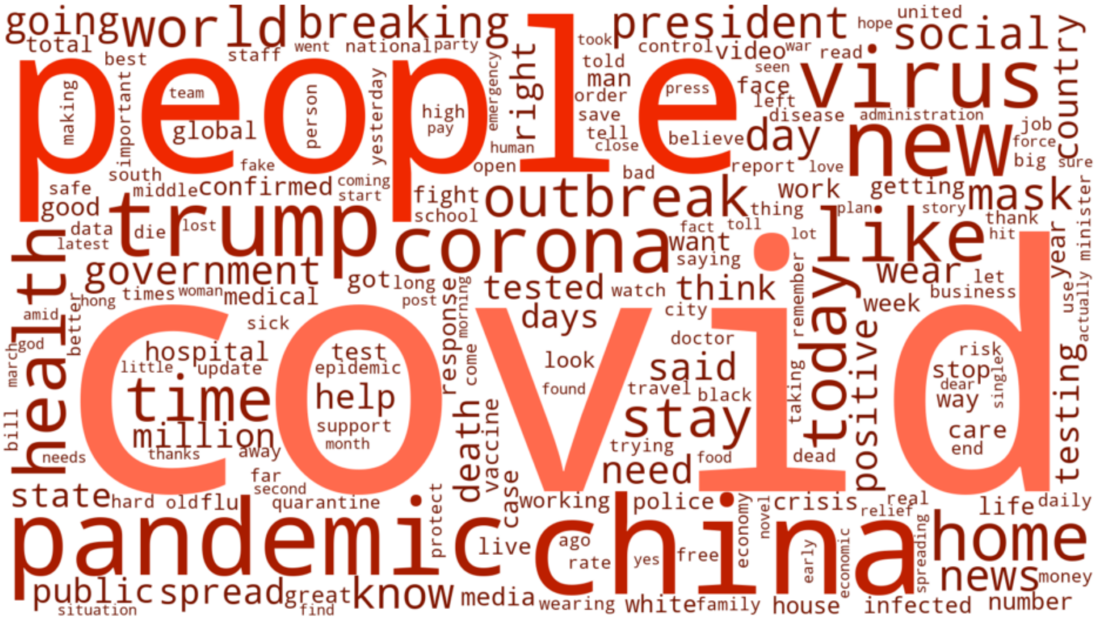

Figure 11: Word Cloud for Tweet Dataset

In order to determine the change in daily topics of the tweet dataset throughout the COIVD-19 pandemic, tweets from each date was grouped together and tokenized. An example of the top 10 topic models from January 25th, 2020 can be seen in Table 3 below. During the initial couple of months of 2020 between January and March, the topics mainly focused on the Chinese coverage of the outbreak as the keywords relating to the early China COVID-19 outbreak were found to be more frequent in the topic models, along with a higher weight. Words such as **China**, **virus**, **spread**, and **outbreak** can be seen to have a high occurrence within each topic.

Topic #	Topic Model
0	0.118*health + 0.073*virus + 0.044*ministry + 0.034*china + 0.030*linked + 0.029*corona + 0.020*old + 0.017*people + 0.017*falling + 0.016*today
1	0.051*doctor + 0.046*hospital + 0.032*video + 0.030*government + 0.030*days + 0.026*think + 0.026*local + 0.023*spreading + 0.023*quit + 0.023*revoke
2	0.136*china + 0.130*death + 0.129*toll + 0.084*people + 0.073*breaking + 0.043*news + 0.028*pushing + 0.016*flu + 0.014*senator + 0.013*story
3	0.120*confirmed + 0.104*breaking + 0.082*china + 0.074*new + 0.059*strand + 0.044*people + 0.034*patient + 0.029*tested + 0.027*said + 0.023*travel
4	0.068*new + 0.043*medical + 0.031*time + 0.025*city + 0.023*streets + 0.021*lunar + 0.020*staff + 0.020*shanghai + 0.018*maybe + 0.017*according
5	0.109*china + 0.104*contain + 0.099*hard + 0.097*united + 0.095*working + 0.094*greatly + 0.052*spread + 0.052*stay + 0.043*safe + 0.028*share
6	0.091*new + 0.066*people + 0.059*infected + 0.055*million + 0.047*population + 0.046*china + 0.045*strand + 0.045*combined + 0.044*deadly + 0.024*day
7	0.108*outbreak + 0.050*china + 0.047*year + 0.046*novel + 0.036*update + 0.021*epidemic + 0.020*media + 0.018*virus + 0.016*wildlife + 0.015*pneumonia
8	0.091*china + 0.068*case + 0.044*confirmed + 0.038*like + 0.031*state + 0.028*lab + 0.027*second + 0.026*virus + 0.019*world + 0.018*built
9	0.195*life + 0.192*save + 0.084*know + 0.046*need + 0.032*study + 0.022*family + 0.016*global + 0.014*threat + 0.014*taking + 0.012*humanity

Table 3: Topic Model for 2020-01-25

However, as COVID-19 progressed and the explosive spread was observed within the United states, the focus of the topics can be seen to shift to relate more to the coverage of the COVID-19 pandemic within the US. Keywords such as **mask**, **president**, **covid**, and **trump** could be seen to have a higher occurrence within each topic.

Topic #	Topic Model
0	0.084*covid + 0.031*got + 0.029*school + 0.028*data + 0.024*testing + 0.022*work + 0.022*parson + 0.022*mike + 0.019*world + 0.019*month
1	0.107*mask + 0.062*covid + 0.057*wear + 0.026*know + 0.024*working + 0.018*social + 0.015*nail + 0.015*salon + 0.015*dad + 0.015*enforce
2	0.062*vaccine + 0.055*covid + 0.054*today + 0.036*new + 0.033*time + 0.032*immune + 0.030*safe + 0.025*corona + 0.023*study + 0.019*media
3	0.039*covid + 0.035*said + 0.030*high + 0.029*going + 0.027*let + 0.021*help + 0.019*starting + 0.018*rate + 0.017*dead + 0.016*president
4	0.089*people + 0.053*pandemic + 0.028*covid + 0.028*second + 0.023*country + 0.022*matter + 0.018*million + 0.018*week + 0.016*case + 0.015*breaking
5	0.052*year + 0.048*old + 0.039*wore + 0.037*complain + 0.036*straight + 0.035*home + 0.020*way + 0.018*public + 0.017*thinking + 0.016*mass
6	0.061*virus + 0.048*china + 0.033*think + 0.032*right + 0.021*days + 0.016*crap + 0.016*find + 0.015*hell + 0.014*buy + 0.014*morning
7	0.038*health + 0.024*worry + 0.024*continue + 0.018*positive + 0.017*disappear + 0.017*wearing + 0.016*death + 0.016*exactly + 0.015*flu + 0.015*slow
8	0.090*trump + 0.045*early + 0.037*daily + 0.032*future + 0.028*white + 0.027*pandemic + 0.026*stop + 0.025*getting + 0.019*house + 0.019*climate
9	0.031*trying + 0.027*chief + 0.019*air + 0.017*nursing + 0.016*free + 0.015*test + 0.014*friend + 0.014*conspiracy + 0.014*officer + 0.013*level

Table 4: Topic Model for 2020-07-15

These small snippets of the topic modeling data captures the changes in trending topics as the pandemic progresses. Initially, the outbreak happened in China which outside of the United States, as such the tweet topic of focus will contain more keywords relating to the Chinese coverage of the outbreak. However, as the COVID-19 pandemic reached the US, the shift in topics can be observed as the trending Twitter topics changes to focus more on US media related coverage. The complete list of dominant daily topics can be found in this **GitHub Link**, and the daily top 10 topic models can also be found in this **GitHub Link.**

7. Evaluation

Overall, the data mining results are promising. However, it should be noted that the classification accuracy of the Naive Bayes Classification model was found to be quite low when used on a set of labeled sample data. The positive tweet detection accuracy was determined to be around 60.9%, and negative tweet detection accuracy was around 45.24%. Neutral detection accuracy tests were not conducted as the sample test dataset only had a 2-way polarity (positive, negative).

There are a number of improvements to be considered for future iteration of this data mining project. First of all, the number tweet sampled each day was around 1000, and after filtering non-English Tweets the number of tweets for each day was reduced to around 500 to 700 tweets. Since the number of sampled tweets for each date is different, comparison of daily sentiment count would be some what inaccurate as the daily sample sizes are different. A Quick solution would be to ensure that the filtered number of tweets for each date are the same before classification. Otherwise, the process can use a larger sample of data or use the complete set of tweet data. However, if a larger or the complete set of tweet data is used, additional hardware support would be needed as there are around 360 Million number of tweets within the repository. Hydration and processing of the Tweets would also take up a significant amount of hardware resources and time.

Another approach is to try different sentiment classification models and sentiment APIs to achieve a higher accuracy for tweet sentiment detection. Classification Algorithms such as Linear Regression, Support Vector Machines, Deep Learning, and hybrid approaches can be considered. There are also many commercial sentiment detection APIs such as the Monkey Learn Text Analysis Platform, Cloud Natural Language by Google Cloud, and LEXALYTICS which are readily available.

8. Summary

The World Health Organization (WHO) defined the SARS-CoV-2 virus outbreak as a severe global threat, since 18,079,136 confirmed cases and 689,347 deaths had been reported across 215 countries. Throughout the events of this pandemic, social media platforms such as Twitter provided access to unprecedented amount of COVID-19 related content and information. An data mining analysis of social media dynamics was conducted on COVID-19 related tweets in relation to daily confirmed cases growth within the US throughout a given time period to measure the change in daily and overall tweet sentiments, and trending topics.

Two datasets were selected for this data mining task; the first was the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, containing an aggregated data source for daily confirmed COVID-19 cases globally. Since the country of focus was the United States, and daily cases were the subject of focus, data was aggregated and formatted to extract information of interests. The second dataset was the COVID -19 Tweet ID dataset gathered by University of Southern California Ph.D. student Emily Chen. The Tweet ID dataset contains COVID-19 related tweet IDs collected beginning from January, 2020. Sampling and Hydration methods were performed on the tweet IDs to extract a limited number of the original tweet texts for topic modeling and sentiment analysis.

From the data mining results, it was concluded as the number of confirmed cases increased, so did the frequency of negative sentiment tweets. A day to day increase in negative sentiment tweet count was observed from the processed data as the severity and number of confirmed COVID-19 cases increased within the United States. Which indicates a growing negative attitude on the Twitter platform as the pandemic worsens. Furthermore, the change in trends for Tweet topics was also observed as the spread of the COVID-19 pandemic developed. Since the outbreak initially occurred in China and later spread throughout the United States, the change in topic models reflected this process. At early stages, the topic of focus contained more keywords relating to the Chinese coverage of the outbreak, and as COVID-19 reached and spread rapidly within the US, trending topics shifted to focus more on US media related coverage.

As the COVID-19 pandemic develops, trends on social media platforms such as Twitter can reflect the growing changes and the general population’s behavior and attitude. Sentiment information and topic modeling derived from data mining of tweeter data can be used by government bodies around the world to determine the general population’s level of attitude towards the pandemic as it progresses and grows. Which allows for proper issue of emergency procedures and restrictions with appropriate messaging and policy decisions in times of crisis for future global events and pandemics.

9. Reference

[1] M. Clinic, “Coronavirus disease 2019 (COVID-19),” Mayo Clinic, 16-Jun-2020. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963. [Accessed: 25-Jun-2020].

[2] L. Gardner, “Mapping COVID-19,” JHU CSSE. [Online]. Available: https://systems.jhu.edu/research/public-health/ncov/. [Accessed: 25-Jun-2020].

[3] R. Holmes, “Is COVID-19 Social Media's Levelling Up Moment?” *Forbes*, 24-Apr-2020. [Online]. Available: https://www.forbes.com/sites/ryanholmes/2020/04/24/is-covid-19-social-medias-levelling-up-moment/#93725e96c606. [Accessed: 25-Jun-2020].

[4] E. Chen, K. Lerman, E. Ferrara, I. S. Institute, C. A. C. C. A. E. Ferrara, C. Author, C. C. A. E. Ferrara, Close, and L. authors..., “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set,” JMIR Public Health and Surveillance. [Online]. Available: https://publichealth.jmir.org/2020/2/e19273/. [Accessed: 25-Jun-2020].

[5] L. Vadicamo, “Cross-Media Learning for Image Sentiment Analysis in the Wild Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, Maurizio Tesconi,” T4SA Dataset - Twitter for Sentiment Analysis Dataset. [Online]. Available: http://www.t4sa.it/#dataset. [Accessed: 25-Jun-2020].

[6] S. M. Piryonesi and T. E. El-Diraby, “Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems,” Journal of Transportation Engineering, Part B: Pavements, vol. 146, no. 2, p. 04020022, 2020.

[7] S. Li, “Topic Modeling and Latent Dirichlet Allocation (LDA) in Python,” Medium, 01-Jun-2018. [Online]. Available: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24. [Accessed: 25-Jun-2020].

[8] C. Kantis, “UPDATED: Timeline of the Coronavirus: Think Global Health,” Council on Foreign Relations. [Online]. Available: https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus. [Accessed: 26-Jun-2020].

[9] J. Nimchuk, “Confirmed Cases & Tweet Count Contingency Table”, Statistical Analysis Proposition. 1st ED. S. CITY, 2020