How People React to COVID 19: An Exploration to March Tweets with NLP Techniques

Sunny Li, MPP Student, Harris School of Public Policy, University of Chicago

Yilun Xu, MACSS Student, Division of the Social Sciences, University of Chicago

Introduction

From the end of 2019, COVID 19 gradually spread and became a terrible epidemic that has swept the world. This disease poses a serious threat to the lives of many people and greatly affects everyone's life. The new coronavirus (COVID-19) pandemic has put tremendous pressure on the citizens, resources, and economy of many countries around the world. Social alienation measures travel bans, self-isolation, and business failures are changing the structure of the global society. Worldwide, the measures taken by different regions and countries to respond to COVID 19 are also different. We hope to understand how people respond to this public health crisis by understanding people’s views on COVID 19. The purpose of this study is not only to inspire citizens to do a good job in epidemic prevention but also to provide a reference for government departments to make relevant decisions in response to COVID 19.

Twitter is a well-known social platform in the United States, and many citizens use Twitter. People can express their opinions on Twitter while interacting with others. Twitter provides a platform for communication and cooperation. On Twitter, information exchange, dialogue, and feedback are all direct and rapid. Twitter is a two-way communication platform and participating in the discussion of topics can produce many substantive effects. At present, many studies have been carried out using Tweets as data, and the research scope is mainly based on the range, frequency, timing, and content of tweets (Jansen et al., 2009). This article will draw on these studies and use tweets as a data carrier to represent people’s thoughts.

Data

This article will use Tweets in and before March 2020[1]. The entire data set contains tweets sent in many different languages. In this study, we only considered tweets sent in English. Besides, due to the limitation of computing power, we only extracted 1.25% of tweets for analysis.

[1] Coronavirus (covid19) Tweets. (2020) https://www.kaggle.com/smid80/coronavirus-covid19-tweets

Methods

We will divide the Tweets into three data sets according to the time they were generated: early, middle, and late. The early tweets included all tweets before March 12. The reason we chose this date as the time node is that this day is the day when Covid19 was determined to be a global pandemic. Mid-term tweets include all data from March 12 to March 20, while late tweets include all data from March 21 to March 31. In different analysis sections, we will use Early, Middle, and Late to refer to the three datasets.

We aim to learn from the Tweets about how people's attitudes towards COVID 19 have changed. Specifically, we will separately analyze the emotional characteristics (positive or negative) of people in these three stages through the Tweets, and study which topics are reflected in the three stages. We will analyze these tweets from micro and macro perspectives. From the micro perspective, we will explore how different words are connected in tweets. When given some words, do we get some information from the tweet and speculate other information in the tweet? The methods involved in this part mainly include communication networks and word embedding and text generation. From a macro perspective, we want to understand which words can express people's views on COVID 19, and how people's attitudes towards COVID 19 in different periods Changed. The methods involved in this part mainly include classification, sentiment analysis, Gensim model, and dynamic modeling.

Communication Networks

A social network is a social structure composed of many nodes. Nodes usually refer to individuals or organizations, and social networks represent various social relationships. With text data, we can understand the connection between different words through social networks. For a specific word, we can explore other words and their connection. The connection between words helps us understand the semantic structure. Many concepts of the social network model come from graph theory because the social network model is essentially a graph composed of a node (word) and edge (social relationship in the text) (Monge et al., 2003). In this part, we mainly study the following questions for tweets in three different stages: Which words are the core words in these tweets? Which words are closely related to the COVID 19 incident? Therefore, we will first construct social networks for the three data sets. Secondly, we will filter out the words that are highly relevant to the COVID 19 event and build ego-networks with nouns related to the COVID 19 event as the center.

The degree is an important concept in communication networks we will use in this part. The degree of a node is defined as the number of edges connected to the node. In a graph, the number of all edges pointing to a certain node is called the node's in-degree, and the number of all edges starting from the node to other nodes is called the node's out-degree. The average degree of the network reflects the density of the network, and the degree distribution can describe the importance of different nodes. Degree centrality is the most direct metric for characterizing node centrality in network analysis. The greater the node degree of a node, the higher the degree centrality of the node, and the more important the node is in the network (Bonacich, 2007). Also, closeness centrality calculates the reciprocal of the shortest distance from a node to all other reachable nodes and normalizes the value after accumulation. The tight centrality can be used to measure the length of time the information is transmitted from this node to other nodes. The greater the closeness centrality of a node, the closer its position in the graph is to the center (Okamoto et al., 2008). In this study, we will refer to the calculation results of closeness centrality as the standard but mainly use degree centrality as the standard to filter words in social networks.

Word Embedding and Text Generation

In this part, Word2Vec, Doc2Vec, and Projection are used to help us understand the topics and influence of COVID19 in each of the three data sets and how the content may differ between different time and different states.

Word2Vec

First, this project use Word2vec methods, it is one of the ways of Word Embedding. And it is the process of transforming words into "computable" and "structured" vectors. To build models on texts, we need first one-hot encoding the words in the texts, and then assign a vector value to represent the relations between the words.

Word2Vec has two different methods: CBOW (Continuous Bag of Words) and Skip-gram. In figure 1, w(t) represents the current word, while w(t-2), w(t-1), etc. are adjacent words.

For CBOW, the goal is to predict individual words given neighboring words. Skip-gram is the opposite: we want to predict a certain range of words given a single word. Both methods use Artificial Neural Networks as their classification algorithms. First, each word in the vocabulary is a random N-dimensional vector. During the training process, the algorithm will use CBOW or Skip-gram to learn the optimal vector for each word.

In our project, we use Word2Vec similarity to different COVID 19 keywords to analyze the sentiment of the tweets.

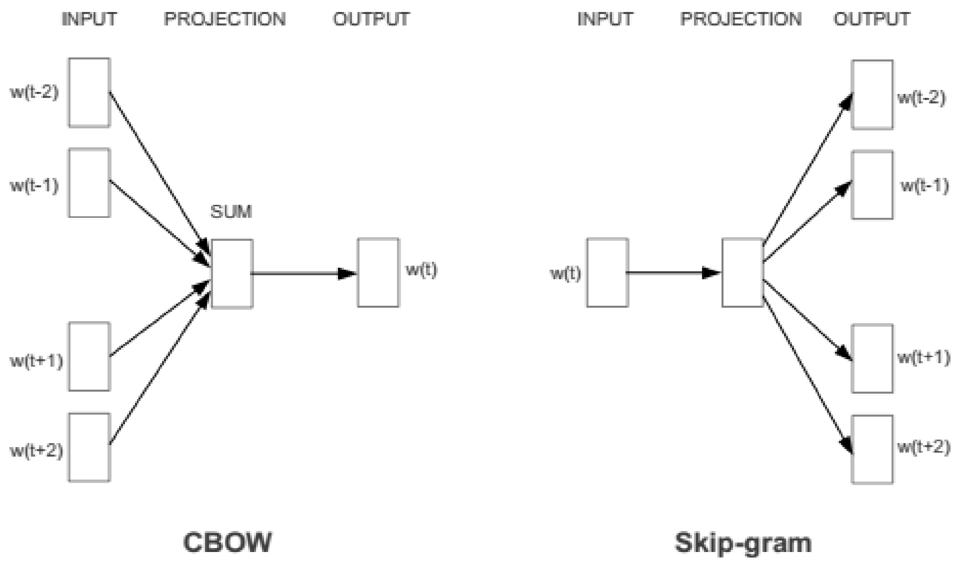


Figure 1: Two methods of Word2Vec

Doc2Vec

However, even if the above method of averaging word vectors is used, we still ignore the word order. Quoc Le and Tomas Mikolov (2014) proposed the Doc2Vec method to describe texts of different lengths. This method is the same as Word2Vec except adding the paragraph/document vector on the original basis. There are also two methods: DM (Distributed Memory, distributed memory) and distributed word bag (DBOW). DM tries to predict the following individual words given the previous word and paragraph vector. Even if the text context changes, the paragraph vector will not change, and word order information can be saved. DBOW uses paragraphs to predict a random set of words in a paragraph.

In our project, we use the most similar words and word relations to explore people's concerns in different aspects of influence brought by COVID19.

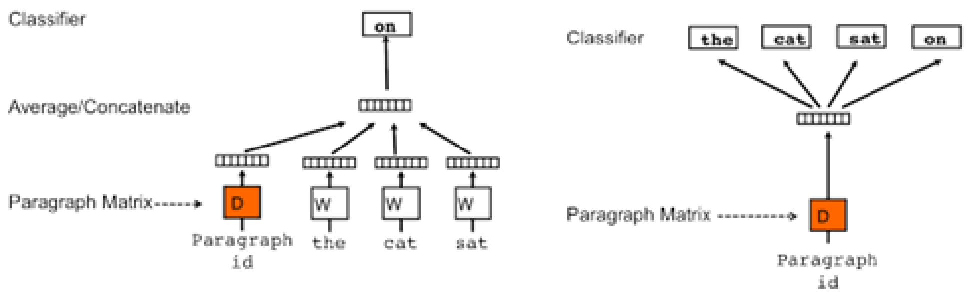


Figure 2: Two methods of Doc2Vec

Classification

Text classification techniques allow us to process texts and organize them into pre-defined classes (Akhter et al., 2020). In this part, the goal to explore in three different stages, which words can clearly express the emotional categories of the tweets. We will explore two classification methods：K-means/K-means++ clustering and hierarchical clustering.

In the K-means algorithm, we randomly select K samples from the data set as the initial clustering center. For each sample in the data set, we will calculate its distance to K cluster centers and classify them into the category corresponding to the cluster center with the smallest distance. Next, for each category, we will recalculate their category center. We will repeat the second and third steps until the cluster centers no longer change (Likas et al., 2003). In the K-means++ algorithm, we assume that the n initial cluster centers have been selected ($0<n<K$), when selecting the next cluster center: the farther from the current n cluster centers, the point will enjoy a higher probability to be selected as the next cluster center. When selecting the first cluster center ($n = 1$), the random method is also adopted (Arthur & Vassilvitskii, 2006).

Hierarchical clustering is also a kind of clustering algorithm. This algorithm creates a hierarchical nested clustering tree by calculating the similarity between data points of different categories. In a clustering tree, the original data points of different categories are the lowest level of the tree, and the top level of the tree is the root node of a cluster. Specifically, we can determine the similarity between the data points of each category and all data points by calculating the distance between them. The smaller the distance, the higher the similarity. We combine the two closest data points or categories to generate a clustering tree (Corpet, 1988). In this research, we employ hierarchical clustering with Wald's method.

At first, since the tweets are not labeled with the emotional categories, we have to label each of them. Therefore, we should decide how many classes do we need. The standard we choose to decide the cluster number in this unsupervised research is the average silhouette method, which measures how close each observation in their cluster is to points in the surrounding clusters (“Cluster Validation Statistics”, n.d.) We want to intend an optimal cluster number from 3, 4, 5, and 6. Therefore, we will calculate the silhouette score for different categories, and then select the cluster number with the largest silhouette score.

After determining the cluster number, we determine their emotion category according to the polarity compound score of each tweet. The value range of the polarity compound score is [-1,1]. The closer the polarity compound score is to -1, the stronger the negative emotion attribute of this text. If the polarity compound score is closer to 1, the positive emotion attribute of this text is stronger. If the best cluster number is 3 according to the silhouette score, we will divide [-1,1] into three intervals, and each interval corresponds to an emotion category. The same is true for other cluster numbers.

We will analyze whether these tweets present obvious sentiment attributes in three different stages, and for each sentiment category, whether there are words that can prominently express this sentiment category. In addition, we will classify tweets in three stages through hierarchical clustering. We will compare the classification results of the two methods to verify whether the sentiment category is a valid criterion for classifying tweets and whether these tweets show obvious differences at different stages.

Word Cloud

Word clouds are a popular and clear visualization method for content analysis. Word clouds aim to create a diagram in which words appearing more frequently in the text will be displayed in larger font sizes (KABIR et al., 2018).

The advantage of word clouds is that it can filter out a lot of text information so that readers can understand the main purpose of the text as long as they see it. In this study, we hope to build a basic understanding of these three stages of tweets through word clouds. We can deduce people's attitudes and focuses through the high-frequency words in these three stages of tweets. This will provide guidance and reference for our following analysis.

Sentiment Analysis

This part aims to find a perceptron algorithm among three ones with the optimal hyperparameters which can predict the emotional tendency of a review accurately and find words that strongly contribute to positive or negative emotion with a unigram bag-of-words model. This part is also an unsupervised machine learning process. We labeled the tweets of three stages as -1(referring to negative) or 1 (referring to positive), which shows their emotion category. The method is that if the polarity compound score of a tweet is bigger than 0.2, it will be labeled as 1, or it will be labeled as -1. For the tweets in each stage, we will divide them into a training set, validation set, and test set. The ratio of these three data sets is 6: 2: 2. We will use three different perceptron algorithms three perceptron algorithms including Perceptron, Average Perceptron, and Pegasos.

A perceptron algorithm uses a feature vector to represent a feed-forward artificial neural network. It is a binary classifier that maps the input (real value vector) on the matrix to the output value (a binary value). $\boldsymbol{w}$ is a vector of real weights, $\boldsymbol{w} * \boldsymbol{x}$ is a dot product. $\boldsymbol{b}$ is a fixed constant. $f(x)$ is used to classify and see if it is affirmative or negative. This is a binary classification problem. If $\boldsymbol{b}$ is negative, then the weighted input must produce a positive value and be greater than $\boldsymbol{b}$, so that the classification result is greater than the threshold $\mathbf{0}$. From a spatial perspective, $\boldsymbol{b}$ changes the position of the decision boundary. The formula $f(x)$ is:

$$f(x) = \begin{cases} 1, & \text{if } \boldsymbol{w} * \boldsymbol{x} + \boldsymbol{b} > 0 \\ 0, & \text{else} \end{cases}$$

The difference between the three perceptron algorithms used in this paper is mainly reflected in how to update the model coefficient vector θ according to the result of misclassification. For data $(\{(x^{(i)}, y^{(i)}), i = 1,2, \dots, n\}, \mathrm{T})$, The update methods of these three algorithms are:

$$\text{Perceptron: if } \; y^{(i)}(\theta * x^{(i)}) \leq 0, \theta = \theta + \; y^{(i)} x^{(i)}$$

$$\text{Average Perceptron: } \theta_{final} = \frac{1}{nT}(\theta^{(1)} + \theta^{(2)} + \cdots + \theta^{(nT)})$$

$$\text{Pegasos: } \theta = \begin{cases} (1-\eta\lambda)\theta + \eta y^{(i)} x^{(i)}, & if\ y^{(i)}(\theta * x^{(i)}) \leq 1 \\ (1-\eta\lambda)\theta, & \text{else} \end{cases},$$

where η and λ are fixed parameters.

We will apply these three algorithms to the tweets and tune their hyperparameters. Finally, we can select the algorithm that can obtain the highest accuracy on the test set as a tool to predict the sentiment of the tweets. This will contribute to the classification of the sentiment category of future tweets. And, with the final selected algorithm, we can also select the words that best reflect the positive and negative emotion of tweets in each stage.

Gensim Model and Dynamic Modelling

In this part, we want to explore the topics covered in each stage of the tweets and also the changes in the topics over time. Two models will be used in this part: the Gensim model and the dynamic model.

First, we will use the IDA model in the Gensim model to extract topics from the three-stage tweets. For each stage of tweets, we obtain the corresponding corpora and dictionary. After assigning an ID to each word, we can sort out the word frequency of each word and construct a sparse vector. With these sparse vectors, we use the LDA model for training. The purpose of this session is to extract and compare the themes of tweets in different stages from the perspective of examining tweets in different stages.

Dynamic topic models are generative models exploring the development of a group of un-predefined topics of given texts. These model series are introduced by David Blei and John Lafferty based on Latent Dirichlet Allocation (LDA). In LDA, the model does not pay attention to the order that a word shows in the document or the position of a text in the entire corpus. However, these factors are considered in a dynamic model. In the model, the texts are divided by time slice, and the documents from each time focus on a group of topics that are updated from the group of topics from the documents of the previous time (Karypis et al., 1999).

The difference between the Gensim model and Dynamic Modelling is that the former treats the subject of the text in different stages in isolation, while the latter links the three stages together. We hope to better understand tweets at different stages by comparing the results of these two models.

Results

Communication Networks

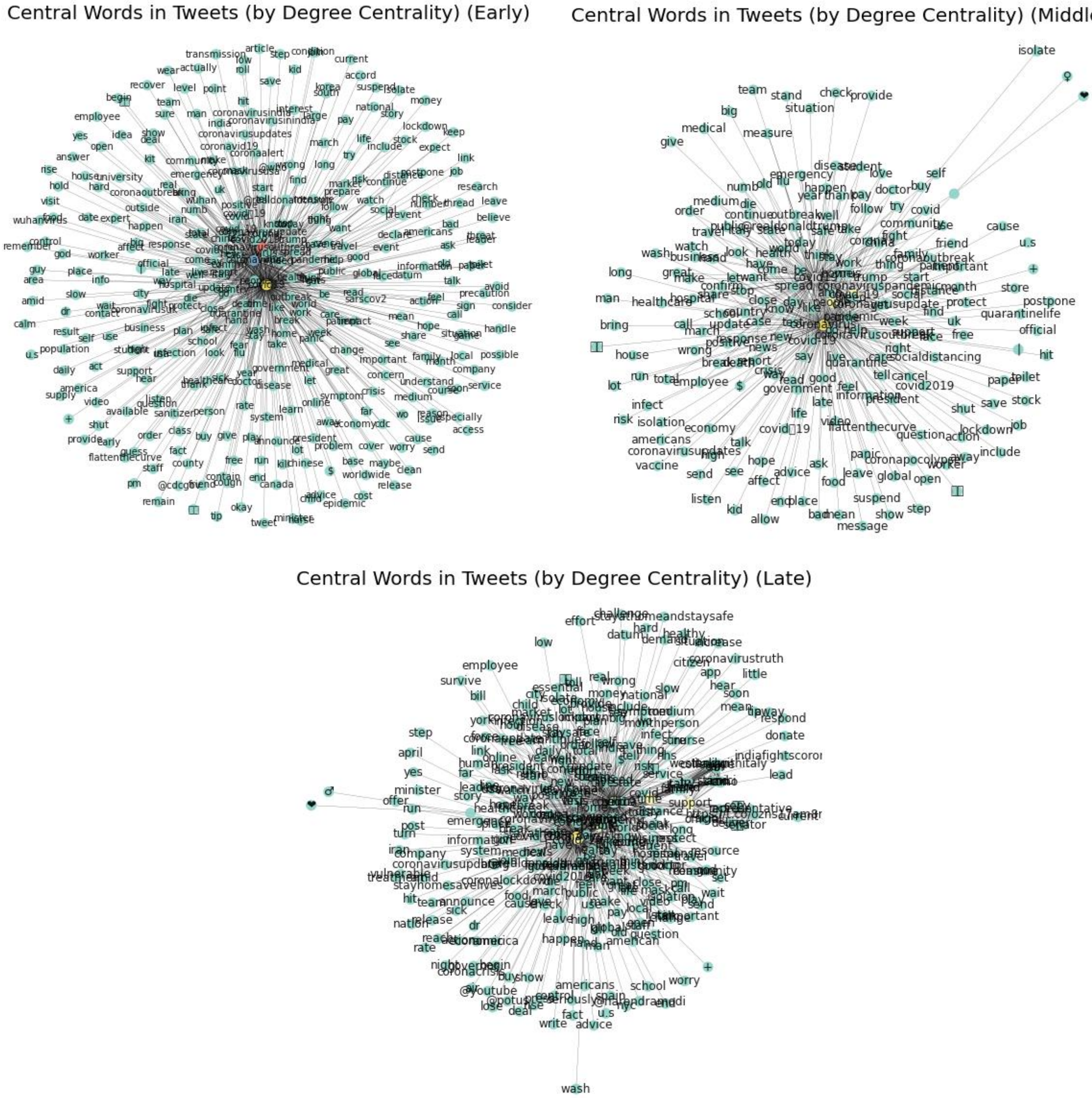


Figure 3: Central Words in Tweets (by Degree Centrality)

First, we compared the headwords of the three stages, and the related social networks are shown in Figure 3. No matter what stage of the tweet, the nouns related to COVID 19 are the core words in the social network, including ‘covid19’, ‘coronavirus’, ‘coronavirusoutbreak’ and so on. This shows that no matter what stage in March, people's attention is focused on the epidemic itself. Also, China and Italy, two regions with severe epidemics, are at the very core. We infer that people hope to learn about the relevant information of the epidemic in the two disease-stricken regions and draw on their epidemic prevention experience. Over time, the importance of words such as ‘stay’, ‘home’, ‘test’, ‘socialdistancing’, ‘lockdown’, and other measures related to responding to the epidemic has gradually increased. We believe that this is related to the gradual increase in the prevention and control measures of the epidemic in the United States, including the issuance of home orders and the increase in the number of tests. Furthermore, we noticed that in early March, words at the core of tweets were directly related to COVID 19. Over time, there are some politically relevant words in the tweets, such as nation, minister, etc. We infer that, in addition to increasing the intensity of prevention and control of the epidemic with the government, this phenomenon also reflects people's expectations for the government: people hope that the government will play a greater role in this public health crisis.

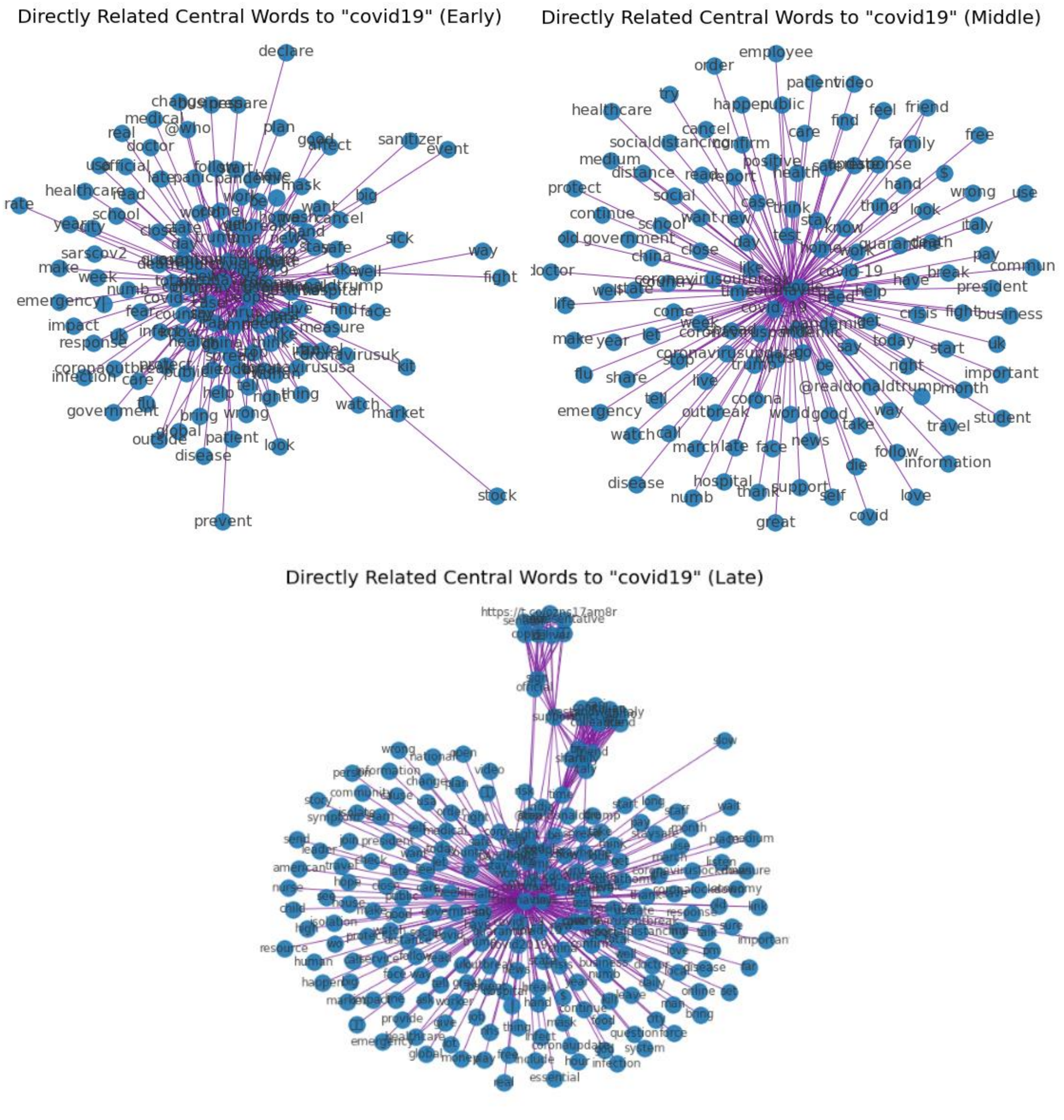


Figure 4: Directly Related Central Words to "covid19"

No matter what stage, 'covid19' is the core word in the text. Therefore, we studied which words are related to 'covid19', and the results are shown in Figure 4. As time goes on, the network becomes denser and denser, and the number of nodes is increasing. Given that the number of samples we use when drawing networks is of the same magnitude, we infer that COVID 19 is attracting more and more attention. In the early stage, the words related to the prevention of COVID 19 are mostly concentrated on the material level, such as sanitizer, mask, market, emergency, and so on. Social protection measures only appear in general terms such as government, official, etc. but do not involve specific measures. We also found that ‘sars’ appeared at this stage, which shows that people are also beginning to realize that COVID 19 may evolve into a very serious infectious disease crisis that affects the whole society. In the middle and late stages, the words ‘socialdistancing’, ‘lockdown’, ‘community’, ‘stay’, etc. appeared, and the President’s Twitter account was also directly related to ‘covid19’. This shows that in response to the crisis of COVID 19, government actions and the power of social groups gradually began to play a role.

After we cut down the nodes and edges with little importance in networks, we obtained more core networks. At the same time, in the new networks, we will study the connection between verbs and nouns. Since ‘covid19’ does not appear in the truncated networks, we will use ‘virus’ as a new research object. For the three stages of tweet data, we have established a ‘virus’-centric network. The words involved include words directly related to ‘virus’, words indirectly related to ‘virus’, and words related to two words separated by ‘virus’. The results are shown in Figure5-7. In these three social networks, the darker the node color of the word in the core position of the tweet, the larger the size.

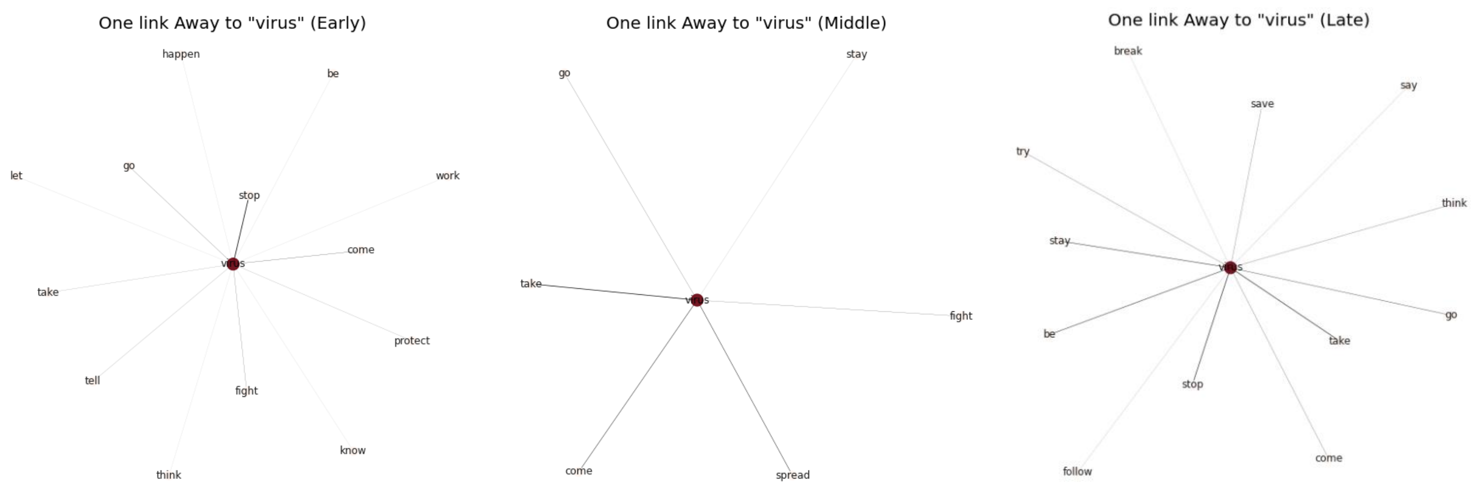


Figure 5: One Link Away to "virus"

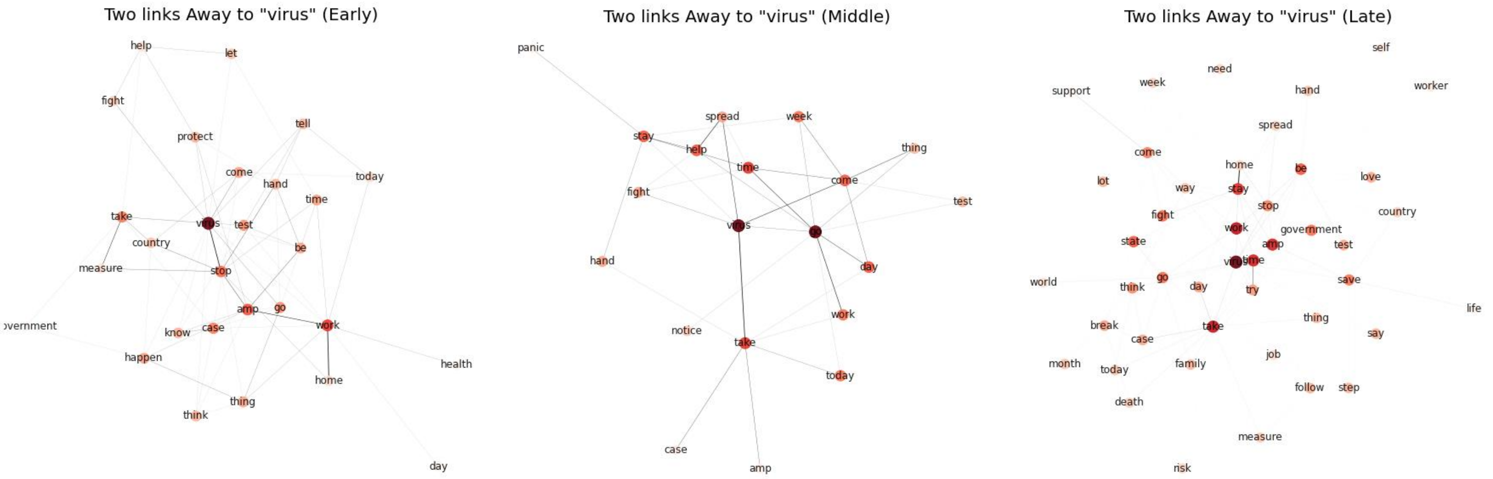


Figure 6: Two Links Away to "virus"

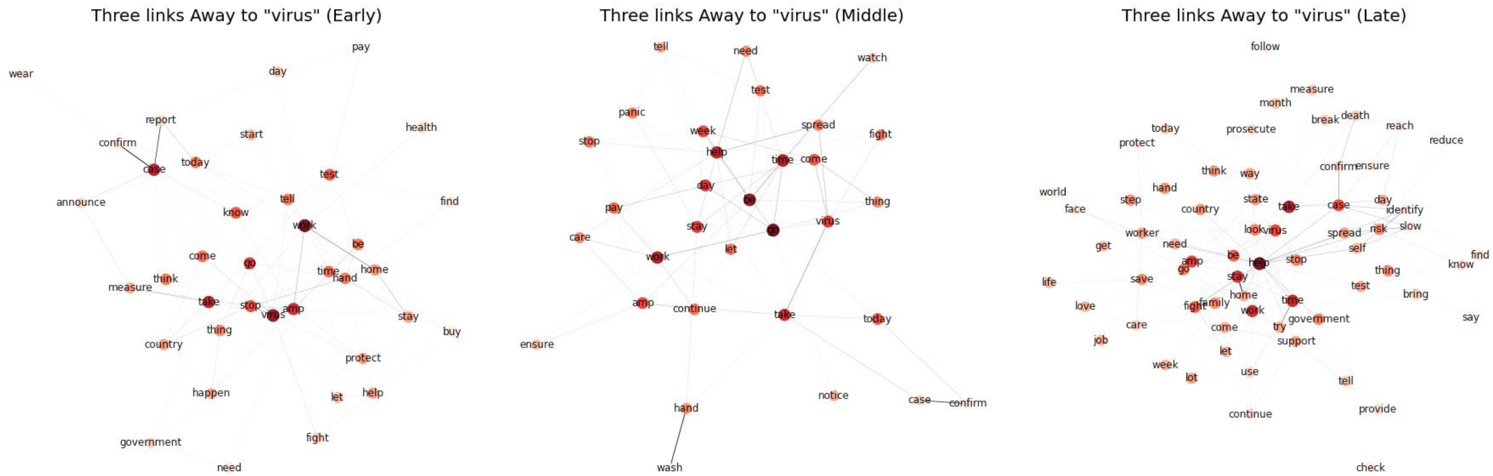


Figure 7: Three Links Away to "virus"

We study the relationship between verbs and nouns, so in one-link-away networks, we study which verbs are directly related to ‘virus’. According to Figure 5, we find that no matter at what stage, the three verbs ‘go’, ‘take’, ‘come’ are directly related to ‘virus’. Interestingly, the number of verbs directly related to ‘virus’ decreases first and then increases.

In the two-links-away networks and three-links-away networks, the third-stage networks were denser and more covered some legal vocabulary, such as prosecute and confirm. We infer that with the proliferation of COVID 19, social conflicts have increased, and more social conflicts need to be resolved through legal issues. At the same time, there is another possibility: local governments have perfected relevant management regulations and included some events involving COVID 19 in the management of laws and regulations. Also, in one-link-away networks, these verbs can be issued directly by individuals. When links increase, the main body involved in the words related to ‘virus’ is also expanding, including actions including ‘prosecute’, ‘pay’, etc. that must be completed by at least two individuals. Therefore, we believe that responding to the COVID 19 crisis requires the joint efforts of the entire social group. If the entire community is abstracted as a social network, then each individual is a node in this social network. Different individuals are connected and influence each other. If the whole community has not defeated the virus in terms of sorting, then the individual's safety cannot be finally guaranteed. In the top words in terms of degree centrality as shown in Table 1 – 3, we can find that the most relevant words to ‘virus’ do not change too much over time and the number of links. The main idea here is that the most common focus of people is to fight against the virus and try their best to keep the normality and standard of their previous life.

Table 1: Top 10 Words in Terms of Degree Centrality Related to ‘Virus’ in Early Tweets

One Link	***Two Links***	***Three Links***
protect	work	case
know	stop	amp
take	amp	work
come	take	hand
happen	hand	stop
go	case	today
think	go	take
left	come	time
tell	thing	test
stop	time	go

Table 2: Top 10 Words in Terms of Degree Centrality Related to ‘Virus’ in Middle Tweets

One Link	***Two Links***	***Three Links***
take	take	time
spread	come	help
come	virus	work
go	stay	be
stay	time	take
fight	help	day
	day	come
	week	virus
	spread	stay
	fight	week

Table 3: Top 10 Words in Terms of Degree Centrality Related to 'Virus' in Late Tweets

One Link	***Two Links***	***Three Links***
follow	stay	take
say	virus	time
save	be	stay
take	go	work
come	save	amp
go	stop	case
think	come	virus
break	work	be
try	time	go
stay	amp	save

Word Embedding and Text Generation

Word2Vec

When comparing the Word2Vec models trained on three periods, and the most similar concepts to 'China', 'mask' by plotting word embedding, the Word2Vec algorithm showed a large difference.

Firstly, in terms of countries of geographical locations associated with covid19, covid19 were often associated with Italy and China in early March, the United States in the middle and late periods.

To be specific, the similarity score between 'China' and 'Covid19' gradually fade away, with a similarity score drops from an early 0.99 to a mid-term 0.44 and finally to 0.29. The first case was discovered in Wuhan China in December 2019, and China is the source of the Coronavirus, so China attracted people's attention most when talking about Coronavirus. Also, at the beginning of March, the outbreak spread rapidly in Italy. The Italian government decided to close all schools across the country and close all other shops except pharmacies and supermarkets. So, people start paying more attention to Italy. And on March 12, the epidemic peak in mainland China has generally passed, the number of new incidences continues to decline, and the epidemic generally remains at a low level. This explains the reason why the similarity score between 'China' and 'Covid19' gradually fade away.

At the same time, the cumulative number of confirmed cases in the United States was about 3,000, and the United States began to close its borders, so people's attention turned on to its own country.

People's discussion on quarantine has focused more on health in the early stages and more on the economic situation and unemployment rate in the middle and later stages. For example, we analyze the word relations between ‘coronavirus' + 'unemployment' - 'quarantine'. In the early period, the most similar words of this word relation are ‘die’, ‘away’ and ‘fear’, however, when it comes to the middle and late period, the most similar words are ‘economic’, ‘funds’ and ‘financial’.

Table 4: Most Similar Words of the 'coronavirus' + 'unemployment' - 'quarantine'

Early		***Middle***		***Late***	
Most similar words	***Scores***	***Most similar words***	***Scores***	***Most similar words***	***Scores***
die	0.999	economic	0.824	global	0.808
away	0.999	global	0.813	humanitarian	0.781
fear	0.999	declared	0.810	funds	0.769
m	0.999	economy	0.810	current	0.766
italy	0.999	funds	0.806	combined	0.765
cancelled	0.999	drtedros	0.796	financial	0.760

Also, from the visualization results of three stages (Figure 8-10), in the late period (Figure 10), ‘quarantine’ and ‘lockdown’ are more obvious than in the early and middle tweets. And in the middle tweets, ‘outbreak’ and ‘pandemic’ are very close, and when we look at late

Figure 8: Dimensionality reduction using t-SNE (early tweets)

tweets, ‘outbreak’ disappeared and replaced with ‘spread’. This illustrates the outbreak and

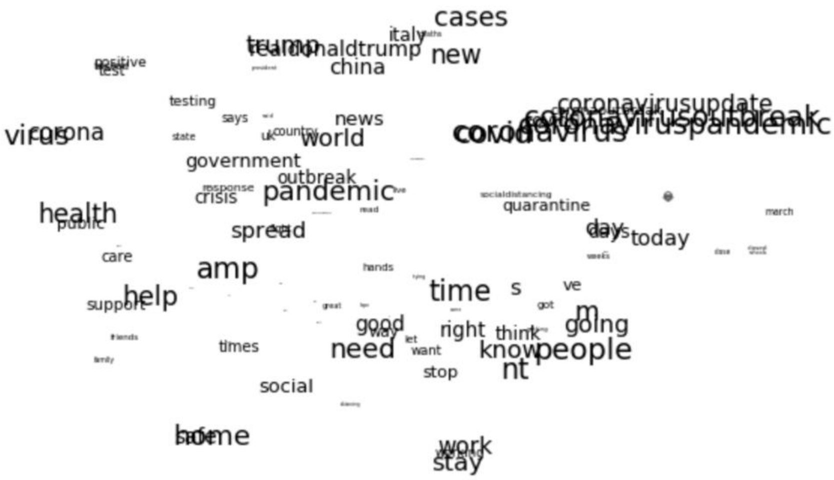


Figure 9: Dimensionality reduction using t-SNE (middle tweets)

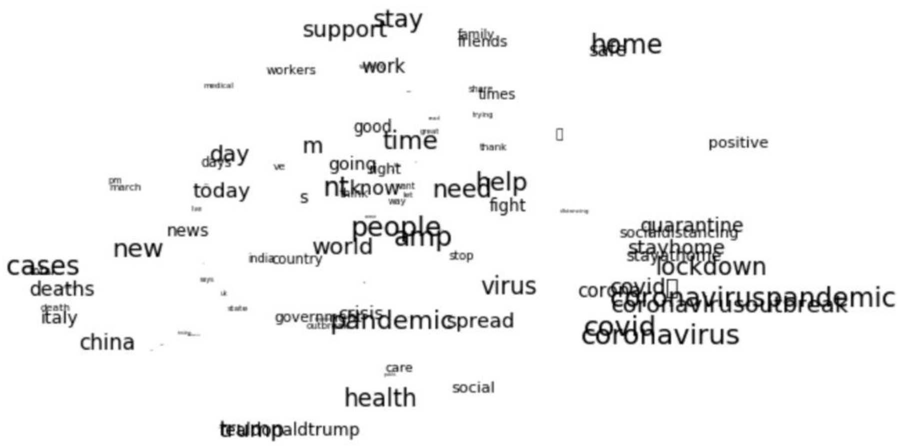


Figure 10: Dimensionality reduction using t-SNE (late tweets)

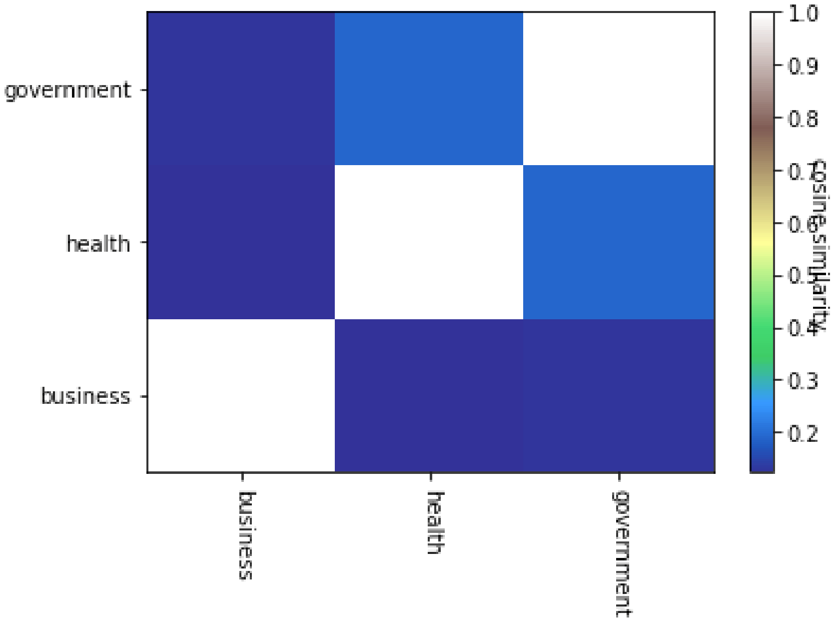


Figure 11: Heatmap of Doc2Vec

spread the timeline of COVID 19 in the US.

Doc2Vec

In this part, we analyze all the datasets ranging from the early stage to the late stage. But we choose two sets of keywords in this method.

For the first set of keywords, we choose ‘business’, ‘health’, and ‘government’, which are the three latitudes that people care about. Then the word relations for *'coronavirus'* + *'die'* - *'quarantine'* suggests that people in tweets blame the spread of COVID19 on China, most of them believe that the Communist Party of China has a responsibility to the rest of the world where suffering from COVID19 since CCP did not open the disease data to the world.

Table 5: Most Similar Words of the 'coronavirus' + 'die' - 'quarantine'

Most similar words	***Scores***
prc (People's Republic of China)	0.732
spread	0.701
conceal	0.692
capitalist	0.689
erase	0.683
mistake	0.677
choose	0.675
incompetents	0.675
Isis (Islamic State of Iraq and Syria)	0.667
ccp (Communist Party of China)	0.667

We also explore people’s concerns about unemployment, except for the impact of quarantine. During the pandemic, the unemployment rate rose to 5.3% in March, and people paying most of the attention to unemployment compensation and economic recovery. Table 3 shows that people mention ‘Davidndii’, it relates to an open letter to the President written by David Ndii. In this letter, he requested the president to take action on COVID19 and care about the suffering of the people. And ‘imf’ also has a high similarity score, we can infer that the IMF (International Monetary Fund) play a key role in responding to COVID19 by quickly deploying financial assistance.

Table 6: Most Similar Words of the 'coronavirus' + 'unemployment' - 'quarantine'

Most similar words	***Scores***
crippled	0.753

repo (repurchase agreement)	0.745
Davidndii	0.737
imf	0.736
worsens	0.735
monies	0.729
moderna	0.725
bil	0.721
investments	0.719
abrupt	0.713

Cosine similarity uses the cosine value of the angle between two vectors in the vector space as a measure of the difference between the two individuals. The closer the cosine value is to 1, the closer the angle is to 0 degrees, that is, the two vectors are more similar. And the heatmap of the keywords shows that the cosine similarity between ‘government’ and ‘health’ is around 0.25 and the cosine similarity between ‘government’ and ‘business’ is around 0.1. We can infer that people think that during the pandemic, government measures should focus more on people ’s health instead of business.

In this section, we project word vectors with a specific semantic dimension. We try to explore the states’ actions and attitudes toward COVID19. First, we created three dimensions, ‘attitude’, ‘economic’ and ‘quarantine’, then project tweets vectors to each dimension.

The value shows that all states' economies are in decline and all issued quarantine order to prohibit going out. And in Figure 12, we can infer that among seven states, Florida's economy is most affected, and Pennsylvania and Washington were less affected. The reason behind this might be that tourism is the main industry of Florida, and the banking and internet industry are the main industries of Pennsylvania and Washington.

Table 7: Dimensions for projection

dimension	***Keywords for one side***	***Keywords for another side***
economic	‘depression’ ‘crisis’	‘boom’ ‘growth’
quarantine	‘prohibit’ ‘ban’	‘welcome’ ‘accept’ ‘allow’ permit’

Figure 12： Projection results of two dimensions

Word Cloud

The word cloud we obtained in three different stages is shown in Figure 13. We can find that some common high-frequency words appear in different stages, such as https co and http(s). These two words represent references to other tweets in one tweet. In these quoted tweets, most of the content is mainly about the news of COVID 19 and selling daily necessities. We can infer that in March, people paid more attention to the development trend of COVID 19 and hoped to prepare for the infectious disease in terms of living supplies. 'COVID19', 'coronavirus', and other nouns describing this infectious disease are always high-frequency words in tweets. The name of the current US president also appeared frequently in tweets. Interestingly, in the word cloud of Middle March, social distancing began to become a word with a high frequency. Mid-March is also a period for people across the United States to quickly call on people to maintain social distance with others. In Late March's tweets, the frequency of 'lockdown' and 'socialdistancing' both increased. This reflects the fact that various regions have begun to take more stringent measures, and the people are also very concerned about how the government will respond to this crisis.

Based on the analysis of the word cloud in these three stages, we speculate that the people repeatedly mentioned the nouns describing the infectious disease, indicating that they may have little experience with such epidemics. The President's name appears more frequently in tweets, indicating that the people expect the government to help them better respond to this crisis. The frequency of 'socialdistancing' and 'lockdown' is on the rise, indicating that more and more people have accepted stricter epidemic prevention policies.

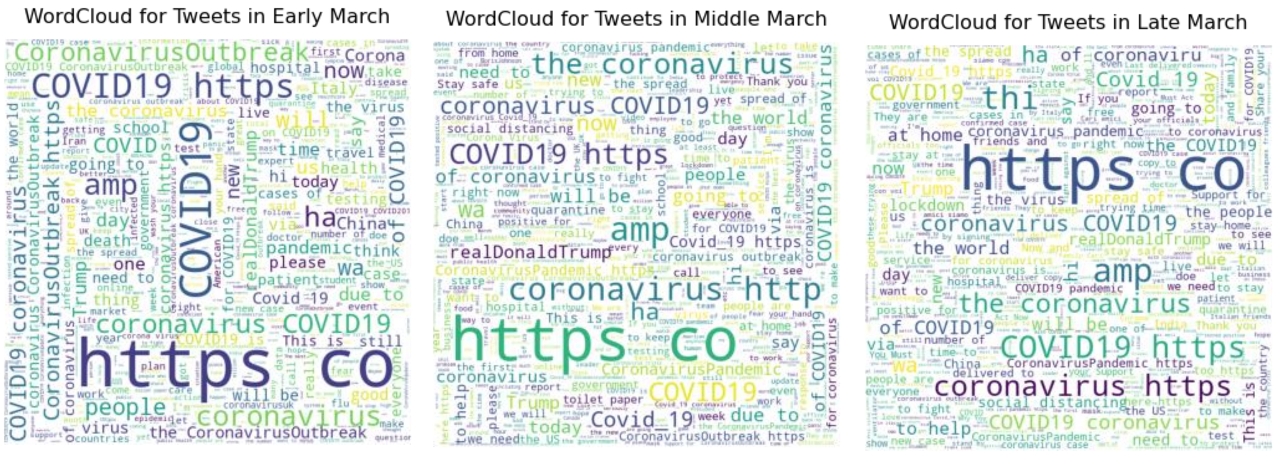
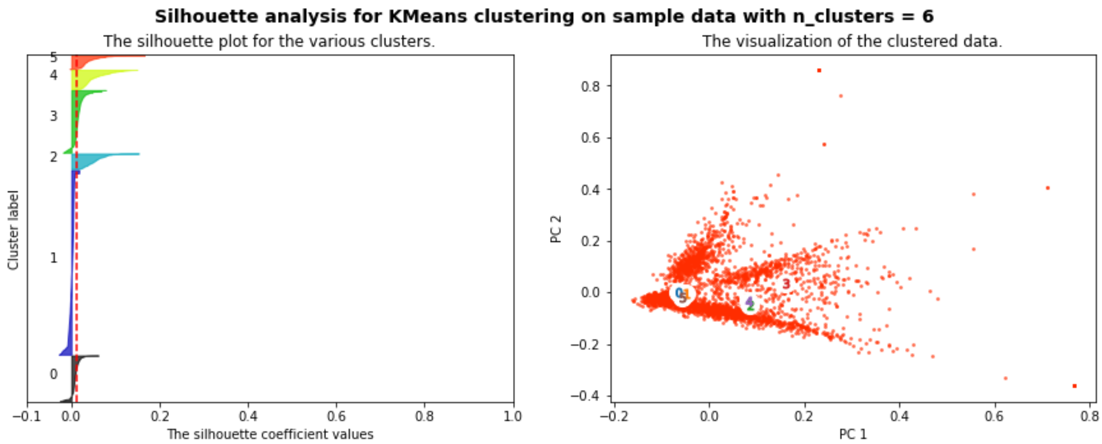


Figure 13: Word Clouds for Different Stages

Classification

We use the K-means algorithm to classify the tweets in the three stages and compare the average silhouette score as the standard. For the tweets in these three stages, the maximum average silhouette score is reached when the cluster number is 6. Therefore, we will divide the tweets into 6 categories and use the K-means ++ algorithm and Hierarchical Clustering with Wald's Method for classification. When we set k as 6, the silhouette analysis of the three stages is shown in Figure 2. The average silhouette score for the three stages is 0.011, 0.013, and 0.016.



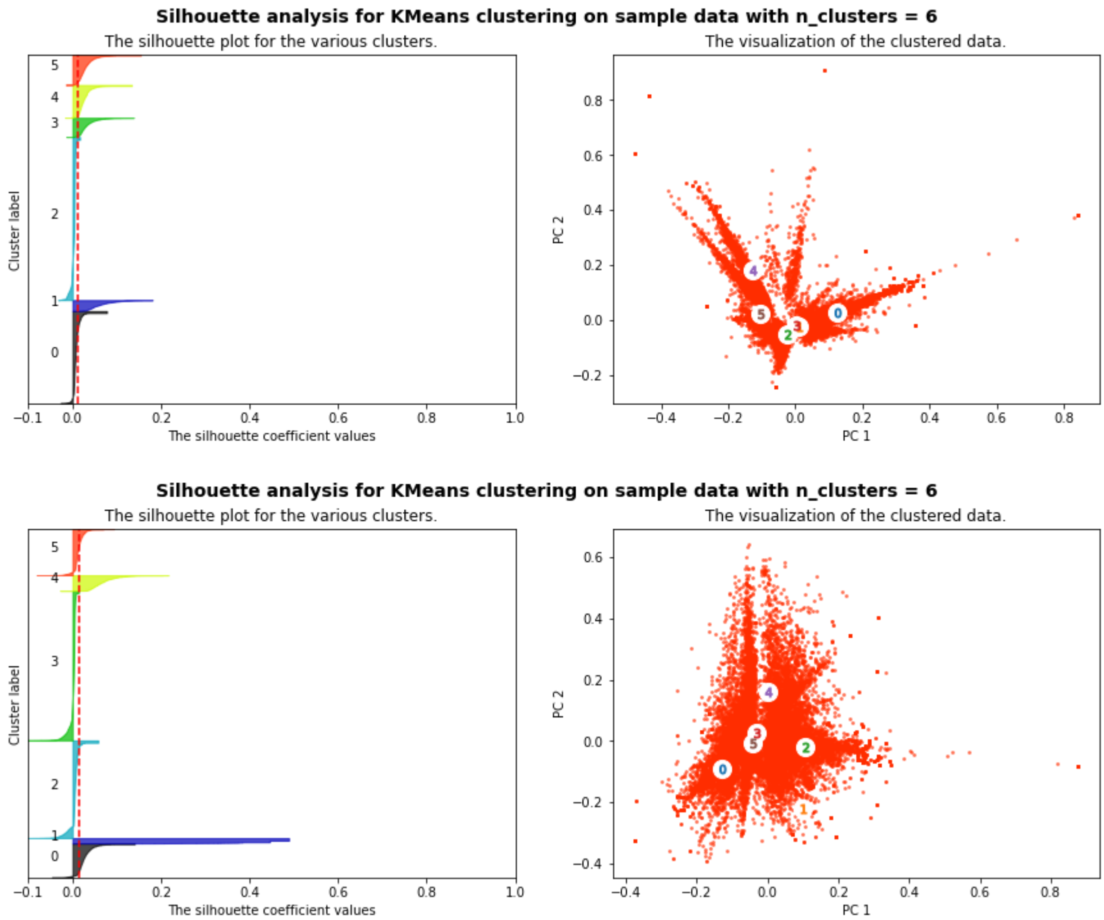


Figure 14: Silhouette Analysis for Different Stages

According to the previous analysis, when we are using K-means++ algorithm and Hierarchical Clustering with Wald's Method to cluster the datasets, we assigned a total of 6 different sentiment categories to each observation, including negative, pretty negative, slightly negative, slightly positive, pretty positive and positive. Related measurements are shown in Table 8.

Table 8: Classification Measurements for Different Stages (K-means++)

	Early	***Middle***	***Late***
Homogeneity	0.018	0.010	0.011
Completeness	0.018	0.009	0.010
V-measure	0.018	0.009	0.010
Adjusted Rand Score	0.016	0.001	0.001

In three different stages, the value of homogeneity is very close to 0. This shows that in these three stages, most of the tweets are not only subordinate to a cluster. In three different stages, the value of completeness is close to 0. This shows that in these three stages, we cannot say that all tweets belong to the same cluster. This means that these tweets belong to different clusters. These tweets present distinctly different textual features, and this difference is collective. Because the first two indicators are very close to 0, we are not surprised that the V-measure value of the tweets in the three stages is also close to 0. In three different stages, the adjusted Rand Score value is very close to 0. This shows that in these three stages, the labels of the tweets are independent of the cluster number and samples. These data provide us with another classification method below: dividing the tweets into two emotion categories and providing support for analysis and proof Our sampling of the original data set did not affect the accuracy of the research results.

As shown in Table 9 - 11, when we continue to look at the top words of different clusters in the three stages, we find that people's attitudes towards COVID 19 are changing. In the early stage, people paid more attention to the COVID 19 virus itself. We can find that in different clusters, the nouns representing this infectious disease are ranked very high. People also discussed the regions or countries where COVID 19 is serious, including China and Italy. In addition to the infectious diseases themselves, people also mentioned health-related and death-related topics. By mid-March, people began to discuss more measures to deal with the outbreak, such as ‘stay’, ‘home’, ‘help’, ‘need’, etc. Besides, the name of the President of the United States also appears more in the important position of different clusters. In late March, more official words appeared, including ‘lockdown’, ‘officials’, ‘senator’, etc. This also shows that the government has begun to take more active measures to deal with the crisis.

Table 9: Top 10 Terms in Clusters in Early March Tweets

Cluster 0	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***	***Cluster 5***
covid	people	coronavirus	amp	cases	coronavirusoutbreak
19	coronavirus	covid_19	coronavirus	confirmed	coronavirus
coronavirus	coronavirusoutbreak	covid2019	just	new	covid2019
coronavirusoutbreak	don	Covid-19	pandemic	total	coronavirusupdate
pandemic	italy	corona	like	coronavirus	corona
outbreak	just	coronavirusupdate	time	deaths	covid一19
covid_19	need	coronavirusoutbreak	don	coronavirusoutbreak	virus
covid2019	virus	china	health	reported	realdonaldtrump
cases	think	covid	virus	china	just
health	infected	pandemic	trump	number	trump

Table 10: Top 10 Terms in Clusters in Middle March Tweets

Cluster 0	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***	***Cluster 5***
coronavirusoutbreak	covid_19	amp	covid	covid19	home
people	coronaviruspandemic	covid19	19	pandemic	stay
just	coronaoutbreak	people	covid19	people	safe

like	coronavirusupdate	health	coronavirusoutbreak	time	covid19
time	corona	covid_19	covid_19	help	work
cases	just	time	cases	just	working
trump	people	like	test	health	covid_19
virus	virus	trump	pandemic	need	people
new	like	need	health	like	time
don	covid19	help	frontline	friends	don

Table 11: Top 10 Terms in Clusters in Late March Tweets

Cluster 0	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***	***Cluster 5***
coronaviruspandemic	covid_19	amp	covid19	covid	cases
people	quarantine	stay	pandemic	19	new
coronavirusoutbreak	people	home	time	covid19	deaths
covid19	lockdown	covid19	help	pandemic	total
just	stayathome	safe	support	coronavirusoutbreak	confirmed
trump	coronaviruspandemic	friends	need	covid_19	covid19
lockdown	time	family	ll	health	number
covid一19	like	people	stayhome	new	000
like	stayhome	support	like	coronaviruspandemic	reported
pandemic	corona	share	health	positive	positive

As shown in Figure 15, Figure 16 and Figure 17, we can see the comparison of predicted clusters and true clusters of tweets in different stages. The results of the classification also show that under this classification standard, the boundaries of different types of tweets are not obvious.

Figure 15: True Clusters and Predicted Clusters of Early Tweets

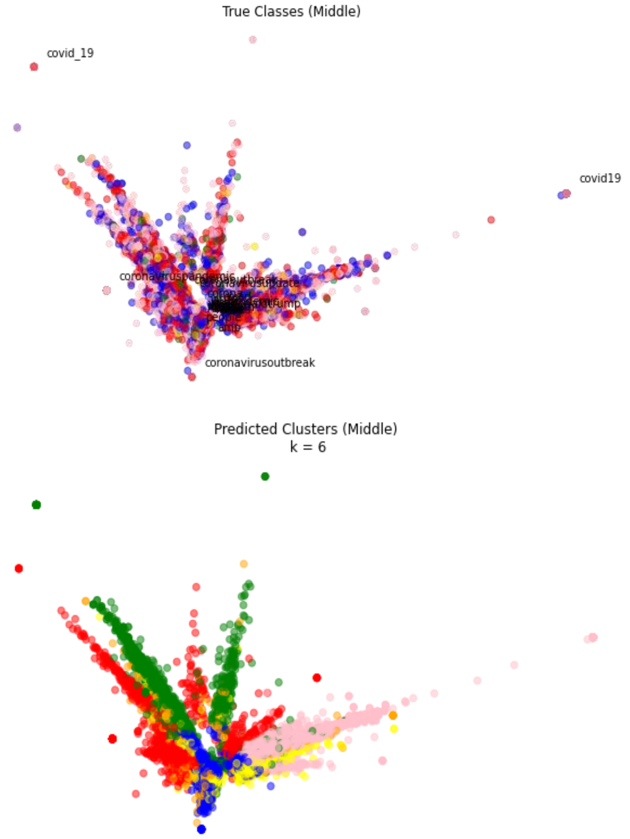


Figure 16: True Clusters and Predicted Clusters of Middle

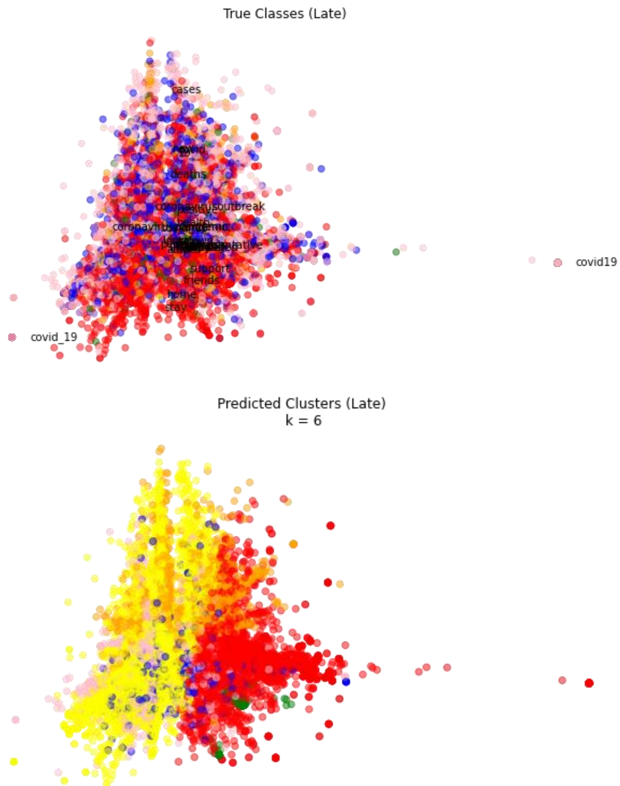


Figure 17: True Clusters and Predicted Clusters of Late

Table 12 shows the measurements of the hierarchical clustering results. Based on these data, the classification effect of hierarchical clustering is better than k-means++. However, this classification method has not yet achieved a good classification effect. As shown in Figure 18, we can see the visualized results of the three stages of hierarchical clustering. The abscissa shows different samples, while the ordinate shows the distance between these samples. Looking from the bottom up, the earlier tweets are connected, the higher similarity they have. The results show that tweets in late March are more diverse. We can find another interesting phenomenon: when we set the maximum distance of different clusters to about 6, the tweets in the three stages can be divided into 2 categories. Therefore, we can infer that maybe when we set the number of clusters to 2, we can get better classification results. This logic supports our computation in sentiment analysis, where we will divide the tweets into negative ones and positive ones.

Table 12: Classification Measurements for Different Stages (Hierarchical clustering)

	Early	***Middle***	***Late***
Homogeneity	0.041	0.047	0.052
Completeness	0.043	0.043	0.054
V-measure	0.042	0.045	0.053
Adjusted Rand Score	0.004	0.013	0.007

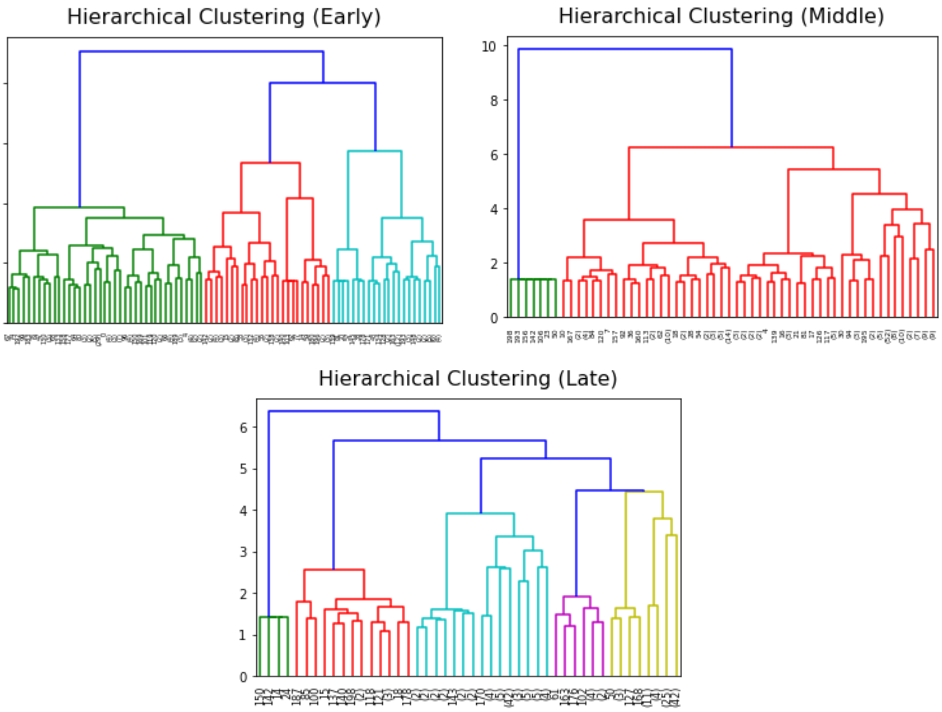


Figure 18: Hierarchical Clustering of Tweets of Different Stages

Sentiment Analysis

We apply three perceptron algorithms to tweets in the three stages, and the classification accuracies are shown in Table 13 - 15 respectively. The accuracies are relatively high no matter in the training set or the verification set. Therefore, we tune hyperparameters to find the best combination of algorithms and related hyperparameters to predict the sentiment category of the tweet.

Table 13: Three Algorithms on Early Tweets (Before Tuning)

	train accuracy	***cv accuracy***
perceptron	0.972647	0.728589
average_perceptron	0.996782	0.740651
pegasos	0.809332	0.718938

Table 14: Three Algorithms on Middle Tweets (Before Tuning)

	train accuracy	***cv accuracy***
perceptron	0.980761	0.699935
average_perceptron	0.996974	0.749190
pegasos	0.784911	0.691510

Table 15: Three Algorithms on Late Tweets (Before Tuning)

	train accuracy	***cv accuracy***
perceptron	0.992037	0.752986
average_perceptron	0.997133	0.757764
pegasos	0.733397	0.693741

We separately trained the three algorithms in three stages of tweets. For each stage of tweets, we will select the algorithm with the highest accuracy and corresponding parameters in the test set. After tuning the hyperparameters, the average perceptron algorithm achieved the highest accuracy for tweets in the middle of March, and the perceptron algorithm works better for tweets in the other two stages. Therefore, we use this algorithm and the corresponding hyperparameters to predict the tweets in three stages. According to the prediction results, we have drawn the top 10 words of positive/negative attitude that best reflect the tweets' emotion category of different stages. The results are as follows:

1. Top 10 words that strongly indicate positive emotion include (early):

['positive', 'best', 'safe', 'sure', 'good', 'help', 'hand', 'chance', 'Please', 'protect']

2. Top 10 words that strongly indicate negative emotion include (early):

['sick', 'stop', 'die', 'emergency', 'fear', 'crisis', 'deaths', 'dying', 'infected', 'isolation']

3. Top 10 words that strongly indicate positive emotion include (middle):

['best', 'safe', 'positive', 'support', 'hope', 'important', 'good', 'Please', 'save', 'care']

4. Top 10 words that strongly indicate negative emotion include (middle):

['crisis', 'death', 'avoid', 'fighting', 'alone', 'kill', 'crisis.', 'worst', 'stop', 'infected']

5. Top 10 words that strongly indicate positive emotion include (late):

['great', 'positive', 'best', 'love', 'hand', 'Thank', 'hope', 'please', 'free', 'safe.']

6. Top 10 words that strongly indicate negative emotion include (late):

['crisis', 'crisis.', 'death', 'infected', 'critical', 'stop', 'wrong', 'ass', 'fuck', 'isolation']

From the above results, we find that words that reflect people's emotional attitudes have not changed much. People's attitude towards COVID 19 is positive. At the same time, people hope to avoid the crisis brought by this infectious disease and the threat to life and health.

Gensim Model and Dynamic Modelling

From Table 16 - 21, we can see the top words that can reflect the various topics of the three stages of tweets. We can further look at how each topic is displayed in each tweet. Besides, as shown in Figure 17, we can find that according to the results of the dynamic model, the topics are pretty scattered, and they enjoy similar importance. This shows that in March, according to tweets, people talk about several major issues in terms of COVID 19.

The results produced by the two models have some similarities. For example, according to these tables, no matter which model we use to analyze tweets in these three periods, there are many overlapping words in the calculation results of the two models. Second, when we change some parameters, the results can change. Figure 19 through Figure 23 compares these changes well.

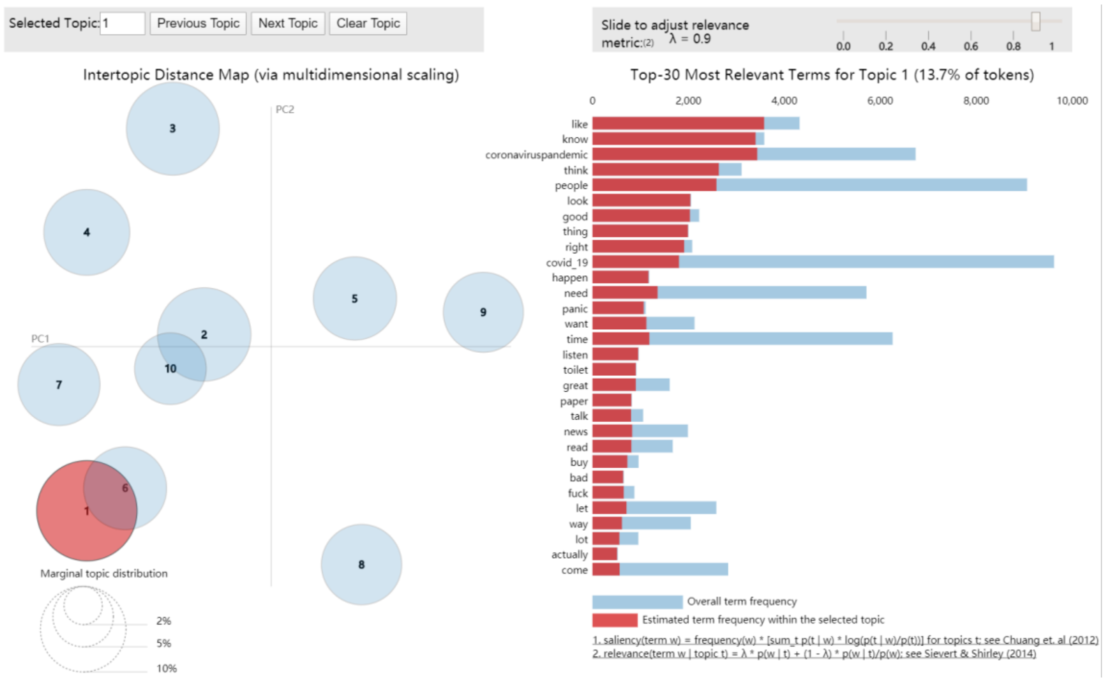


Figure 19: Dynamic Modelling (λ = 0.9)

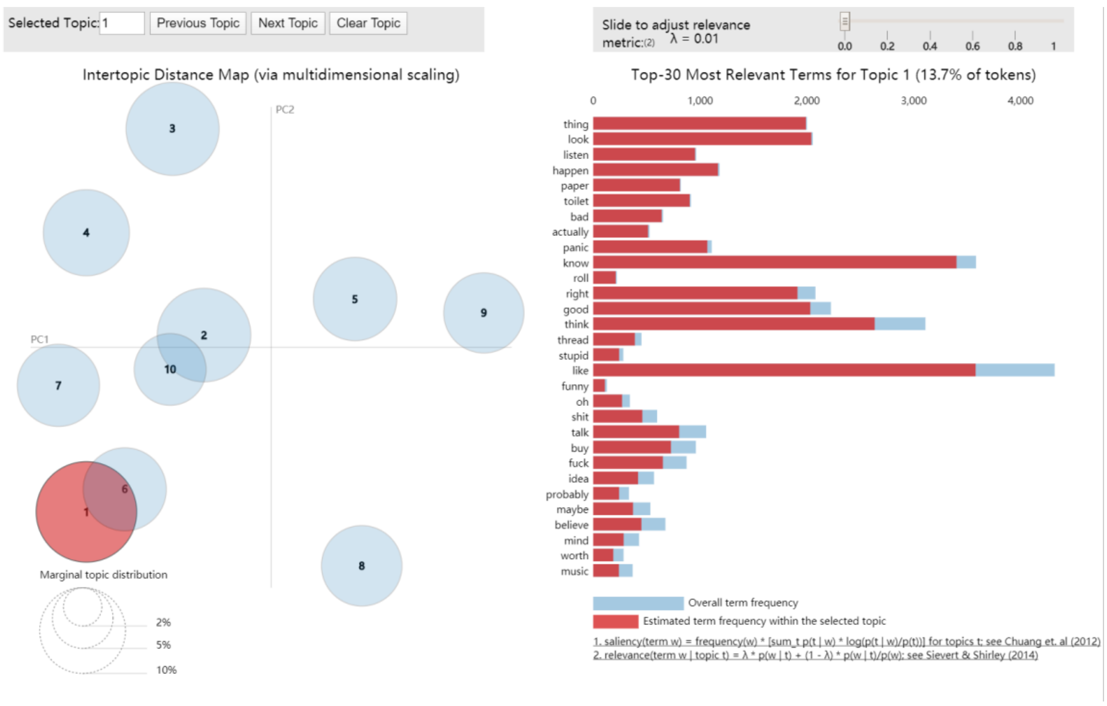


Figure 20: Dynamic Modelling (λ = 0.1)

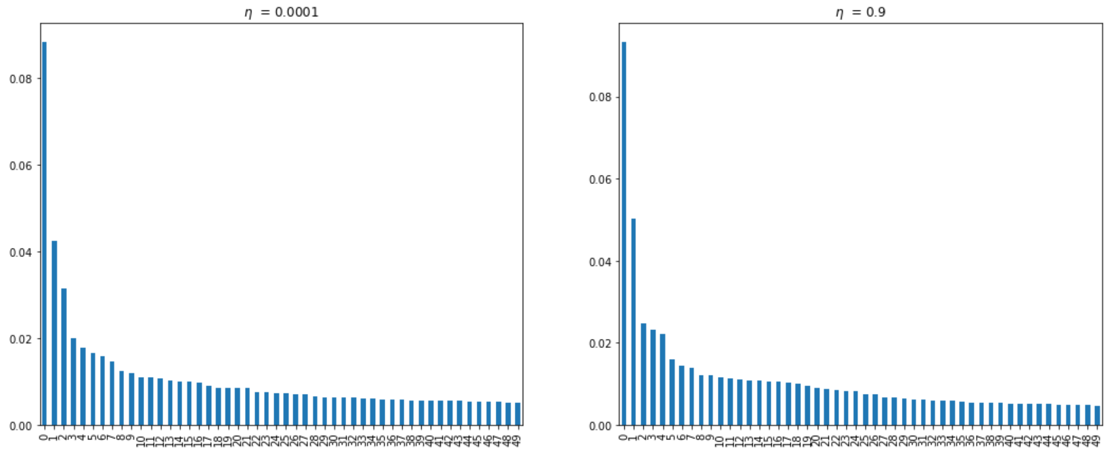


Figure 21: Different Topic Structures (Early)

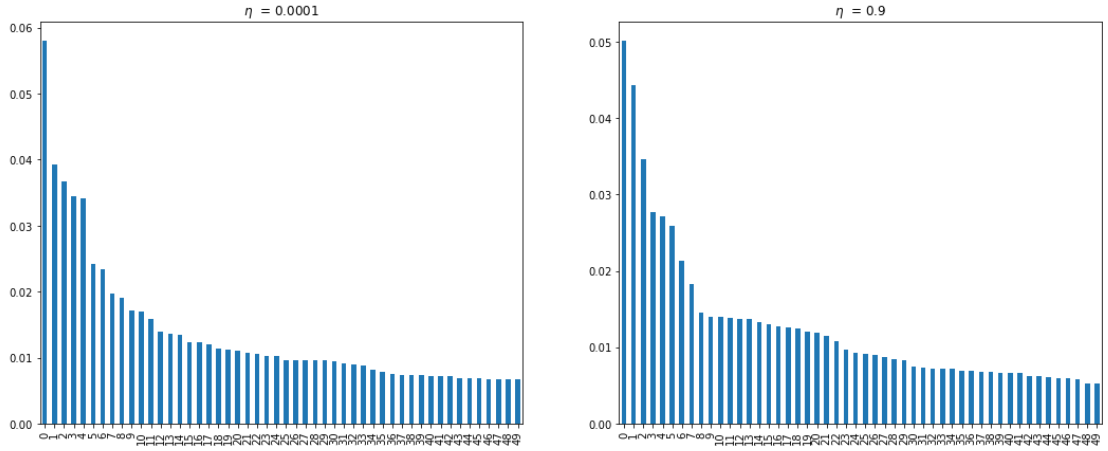


Figure 22: Different Topic Structures (Middle)

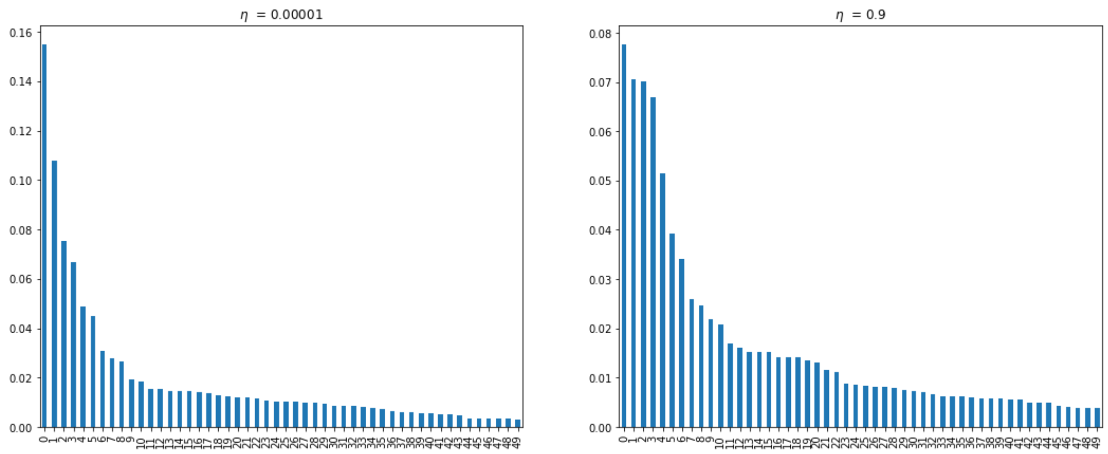


Figure 23 : Different Topic Structures (Late)

At the same time, the specific words in the topics calculated by the two models are quite different. The Gensim models better reflect the characteristics of the content of tweets at each stage. According to the three Gensim models, many terms describing COVID 19 appeared in tweets in the early stages of March, such as coronavirus, coronavirusoutbreak, etc. We speculate that in early March, people were still shocked by the relevant news of COVID 19, and people did not get accustomed to the life changes brought about by the pandemic, nor did they know how to respond to this crisis. However, in the middle and late March, COVID 19 related terms appeared more concentrated in 1-2 topics. This is similar to the results of dynamic modeling. It’s deduced that by the middle and late March, people gradually recovered their emotions and began to view the crisis rationally. The number of tweets in these three phases also shows how people's attitudes change. With the spread of COVID 19, people pay more attention to COVID 19. The average daily tweet count for people in mid-to-late March is about 10-12 times the average daily tweet count before mid-March. In the meanwhile, during this period, more measures related to epidemic prevention appeared in the topics of tweets, such as ‘socialdistancing’, ‘isolation’, ‘lockdown’ and so on. In the dynamic model, the change of keywords in each topic in each period is not as large as the results of the Gensim model.

Discussion

The good characteristics of big data include big, always-on, and nonreactive (Salganik, 2018). In the project, the datasets include huge volumes of data. The original entire more than 760,000 Tweets. This not only allows us to infer more accurate models but also allows us to explore some subtle differences. Furthermore, since we only used data in and before March, the data will not change in the future. Finally, whether we do this research or not, people's opinions on COVID 19 will not change.

The bad characteristics of big data include incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, dirty, and sensitive (Salganik, 2018). Although these characteristics are sometimes unavoidable in social science research with big data, we have solutions to reduce their negative impact on this project.

1. The data we use includes all the tweets available before and in March, so there is no problem with incompleteness.

2. The datasets are all open data that are accessible online. Therefore, we do not have the problem of inaccessibility.

3. Our datasets do not contain any personal information. Thus, they are not sensitive.

Validity means how much the results of our project can deduce general conclusions. Internal validity focuses on if researchers conduct the experimental procedures properly, and external validity measures how much the conclusions of our project can be generalized to the population (Salganik, 2018). In this project, people can freely express their opinions on Tweets, so there is no question of authenticity. External validity is determined by how we select unbiased samples of Tweets that can represent people's thoughts. Due to the limited computing power of our computers and the lack of our team members' ability to recognize other languages, we only extracted 1.25% of the English tweets for research, and they are also resampled in the research process. Furthermore, Twitter users do not cover all civic groups, and our research results cannot represent the ideas of all groups of people, and it is difficult to even popularize all English Speakers. However, Twitter is still one of the platforms with the largest number of users on all social media. When a computer with more computing power is available, we can analyze all the data.

Our research results are consistent with our common sense, and also with the actual situation. For example, after the WHO declared COVID 19 as a global pandemic on March 11, the number of tweets in the middle and late March increased rapidly. We have used different methods in our research, but the results of different research methods can mutually confirm the validity of other results. For example, in the word cloud that high-frequency words and changes in high-frequency words in different periods are also reflected in dynamic models and communication networks. Therefore, we believe that our research results make sense.

Conclusion

The coronavirus is spreading rapidly around the world, bringing deep fear and anxiety to individuals and society.

No matter what stage in March, people's attitude towards COVID 19 is still based on ensuring personal safety and normal life. People are more concerned about how individuals, society, and the government should take action to better protect people's health in this public health crisis. In the early stages of March, people's attitude towards COVID 19 is still at a stage where they do not understand and accept new news. People also pay more attention to areas with severe epidemics such as China and Italy. By the middle and the end of March, citizens and the government began to take stricter measures to deal with COVID 19. More people are staying at home to work and study, and the government has introduced stricter control measures including quarantine and lockdown. At the same time, the corresponding laws and regulations are constantly improving, the government began to use legal weapons to regulate people's behavior. It can be said that March is the period when people's lives have changed the most because of COVID 19. Also, we discovered from social networks that we live in an interconnected world, and the new coronavirus has demonstrated this. Everyone's health is closely related to the health of other people in the society during the pandemic. Actions to prevent the spread of the virus need to reach everyone. This fact requires that we must work together to deal with the crisis, and no one can solve this problem alone.

Reference

Ahmed Imran KABIR, Ridoan KARIM, Shah NEWAZ, & Muhammad Istiaque HOSSAIN. (2018). *The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R*. https://doi-org.proxy.uchicago.edu/10.12948/issn14531305/22.1.2018.03

Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., Mehmood, A., & Sadiq, M. T. (2020). Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network. *IEEE Access, Access, IEEE*, *8*, 42689–42707. https://doi-org.proxy.uchicago.edu/10.1109/ACCESS.2020.2976744

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding*. Stanford.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. Social networks, 29(4), 555-564.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, *16*(22), 10881-10890.

Cluster Validation Statistics: Must Know Methods. (n.d.). Retrieved from https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11), 2169-2188.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68-75.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

Monge, P. R., Peter, R., Contractor, N. S., Contractor, P. S., & Noshir, S. (2003). Theories of communication networks. Oxford University Press, USA.

Salganik, Matthew J., Bit by Bit: Social Research in the Digital Age. Princeton University Press, 2018.

Okamoto, K., Chen, W., & Li, X. Y. (2008, June). Ranking of closeness centrality for large-scale social networks. In International workshop on frontiers in algorithmics (pp. 186-195). Springer, Berlin, Heidelberg.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196.

Table 16: Gensim Topic Model for Early Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
think	case	coronavirus	coronavirusoutbreak	coronavirus	trump	people	coronavirus	travel	coronavirus
coronavirus	coronavirusoutbreak	hand	coronavirus	coronavirusoutbreak	lot	case	coronavirusoutbreak	new	coronavirusoutbreak
health	coronavirus	wash	covid2019	close	coronavirusoutbreak	say	corona	coronavirus	covid_19
time	new	people	case	kill	like	tell	covid_19	case	pandemic
people	total	amp	italy	test	thing	need	coronavirusupdate	state	amp
amp	like	coronavirusoutbreak	covid—19	wake	response	coronavirusoutbreak	covid	cancel	coverage
public	bring	come	india	virus	coronavirus	work	covid2019	city	outbreak
coronavirusoutbreak	report	break	coronavirusinindia	country	way	quarantine	people	test	news
organization	virus	sick	care	school	say	coronavirus	virus	spread	virus
fast	question	home	govt	emergency	look	infection	case	people	people

Table 17: Dynamic Topic Model for Early Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
help	like	coronavirusoutbreak	hand	test	amp	stay	day	pandemic	case
health	coronaviruspandemic	virus	spread	close	work	home	quarantine	trump	new
need	know	corona	covid_19	school	business	people	week	world	death
support	think	covid一19	watch	positive	pay	safe	year	country	report
care	people	coronavirusupdate	ask	state	food	time	feel	crisis	lockdown
community	look	coronaviruspandemic	wash	say	home	social	self	live	italy
thank	good	china	question	order	company	try	covid_19	president	total
hospital	thing	covid2019	stop	uk	help	family	time	response	update
medical	right	covid_19	video	march	service	let	month	say	country
information	covid_19	covid	soon	people	essential	share	house	global	india

Table 18: Gensim Topic Model for Middle Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
home	test	amp	case	covid19	covid19	covid19	covid19	covid19	coronavirusupdate
stay	help	covid19	covid19	update	pandemic	social	support	like	covid_19
covid19	covid19	know	new	close	quarantine	let	trump	fight	covid19
work	friend	people	death	business	crisis	people	share	day	coronavirusoutbreak
safe	family	hand	italy	school	health	distance	time	china	coronaviruspandemic
time	positive	question	virus	march	self	right	care	look	corona
people	community	time	patient	learn	isolation	need	need	world	lockdown
house	people	wash	china	state	public	come	video	watch	socialdistancing
feel	need	mask	government	amp	work	stop	news	play	coronacrisis
love	try	tell	spread	city	company	know	thank	amp	covid2019

Table 19: Dynamic Topic Model for Middle Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
help	like	coronavirusoutbreak	hand	test	amp	stay	day	pandemic	case
health	coronaviruspandemic	virus	spread	close	work	home	quarantine	trump	new
need	know	corona	covid_19	school	business	people	week	world	death
support	people	covid—19	watch	positive	pay	time	year	country	report
care	think	coronaviruspandemic	ask	state	food	safe	feel	crisis	lockdown
community	look	china	wash	say	home	social	self	live	italy
thank	good	coronavirusupdate	question	order	help	try	covid_19	president	total
hospital	right	covid2019	video	uk	company	family	time	response	update
medical	thing	covid_19	stop	march	job	let	month	say	india
doctor	covid_19	covid	soon	people	service	share	house	global	country

Table 20: Gensim Topic Model for Late Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
covid19	amp	mask	covid19	case	covid19	covid19	covid19	covid19	covid19
stay	fight	family	lockdown	covid19	people	pandemic	deliver	trump	amp
covid_19	people	share	people	death	covid_19	virus	support	like	help
home	come	try	thank	test	time	people	hospital	coronavirustruth	pandemic
stayhome	covid19	time	covid_19	new	food	doctor	official	say	business
corona	pm	friend	world	report	pay	world	worker	covid_19	read
stayathome	month	covid19	country	positive	watch	god	act	people	service
safe	govt	wear	day	total	amp	cause	work	president	response
quarantine	protect	stand	woman	china	happen	listen	sign	stop	need
coronavirusoutbreak	join	face	know	update	make	story	patient	know	crisis

Table 21: Dynamic Topic Model for Late Tweets

Topic_0	***Topic_1***	***Topic_2***	***Topic_3***	***Topic_4***	***Topic_5***	***Topic_6***	***Topic_7***	***Topic_8***	***Topic_9***
help	like	coronavirusoutbreak	hand	test	amp	stay	day	pandemic	case
health	coronaviruspandemic	virus	spread	close	work	home	quarantine	trump	new
support	know	corona	covid_19	positive	business	people	week	world	death
need	people	covid—19	watch	state	pay	safe	year	country	lockdown
care	think	coronaviruspandemic	video	say	food	time	feel	crisis	report
thank	right	china	ask	school	home	family	self	live	italy
deliver	look	covid2019	wash	order	job	try	covid_19	president	total
hospital	good	covid_19	question	uk	company	social	time	response	update
medical	thing	covid	stop	march	help	share	month	say	india
mask	covid_19	coronavirusupdate	soon	people	essential	let	house	global	rate