Healthcare Concerns During a Pandemic: Public Sentiment Analysis of COVID-19 Related Tweets

Student:Vincent Feng (19075325)
Primary Supervisor: Samaneh Madanian

Auckland University of Technology

Abstract. As from the first outbreak of Novel Coronavirus (COVID-19) was reported from Wuhan, China, on 31 December 2019, more than 127 million people have been confirmed with the COVID-19, and over 2.7 million people have lost their lives. Meanwhile, many worldwide issues began to surface because of the impact of COVID-19, such as rising unemployment, economic crisis and public panic. Since then, social media platforms have experienced an exponential growth regarding the content related to the COVID-19. In the recent few years, Twitter as an indispensable tool has been applied in variety of fields in terms of crisis warning, sentiment analysis and commercial activities.

The objective of this paper is to use Machine Learning (ML) and Natural Language Processing (NLP) techniques to extract public sentiments of the epidemic disease from Tweets. Compared with traditional data analysis methods, a novel cloud computing scenario Amazon AWS will be utilised in this paper to analyze the Twitter datasets. The outcome of Tweets analysis could help us to identify the general population concerns of the COVID-19, in order that, the governments and communities could get a better understanding of current situation and implement appropriate policies.

To achieve the objective of this research, a systematic literature review and the quantitative data analysis method will be talked about in detail in next sections. Eighty related literature will be reviewed systematically in the literature review section. The COVID-19 Tweets Dataset (Lamsal, 2020a) published on IEEE DataPort will be used to implement the sentiment analysis.

Keywords: Novel Coronavirus, COVID-19, Social Media, Twitter, Tweets Analysis, Sentiment Analysis, Machine Learning, Natural Language Processing, Amazon AWS, Cloud Computing.

Table of Contents

1 Introduction

In this section, the background of this research will be presented in brief, followed with problem statement, objective of the research and research question, research methodology, significance of the study and structure of the research.

1.1 Background of the research

1.1.1 Social media and crisis events

When a crisis occurs, no matter what kinds of it, people prefer to spend more time on social networking and social media than normal. Normally, people get or update the information of the crisis and make informal conversations with their relatives and friends including posting their current situations, querying the status of their families and sharing the real-time description of the crisis, etc (Castillo, 2016; Imran, Castillo, Diaz, & Vieweg, 2015). To acquire the latest situation of others around you, Facebook released a novel functionality called **'I am safe'**, in which people living in the same area where a disaster has occurred would receive a message from Facebook asking if they are safe. People who receive that message should click **'I am Safe'**. Under Friends in the area, you will see a list of your friends who are **'Marked Safe'** and who are **'Not Marked Safe'**. As crisis happens, the volume of data related to the crisis would have an explosive increase on social media platforms such as Twitter and Facebook (Imran et al., 2015), which have a quicker response than emergency response agencies and official media channels (Imran, Ofli, Caragea, & Torralba, 2020). The continuous information sharing process on those social media platforms will generate sheer volume of metadata in the database of the platforms. The large scale of the metadata could range from millions to billions (Kalyanam, Quezada, Poblete, & Lanckriet, 2016). The metadata produced from social medias could be processed and analysed to extract useful information which could help the decision makers to make more effective decisions for future crisis and improve their response to it.

Currently, there are several mainstream social media platforms in terms of Twitter, Facebook, TikTok, WeChat, etc. Among these, Twitter provide a friendly and well-documented Application Programming Interface (API), with which we could programmatically analyze, learn from, and engage with the conversation on Twitter. Due to the user-friendly features, Twitter has become an indispensable source of information for data analyst to analyze the data in the Social Computing field. Several research studies (Carley, Malik, Landwehr, Pfeffer, & Kowalchuck, 2016; Chatfield, Scholl, & Brajawidagda, 2013; Cheong & Lee, 2011; Earle et al., 2010; Imran, Castillo, Lucas, Meier, & Vieweg, 2014; Landwehr, Wei, Kowalchuck, & Carley, 2016; Takahashi, Tandoc Jr, & Carmichael, 2015; Z. Wang, Ye, & Tsou, 2016; Zahra, Imran, & Ostermann, 2020; Zou, Lam, Cai, & Qiang, 2018) have been conducted and demonstrated that the specific crisis related Tweets could provide a better insight regarding the situation. For

deeply analysing the crisis related information, large-scale tweets related to the events including Pakistan Floods, India Floods, Palestine Conflict, Nepal Earthquake, Japan Earthquake, Flight MH370, etc. have been collected and made accessible online (Imran, Mitra, & Castillo, 2016; Burel & Alani, 2018). Moreover, the metadata of Twitter have also been utilised in modeling the Machine Learning architecture (Imran et al., 2014; Lamsal, 2020c; D. Nguyen et al., 2017) for categorizing the invisible tweets into specific categories such as volunteering efforts, community needs, infrastructure damages and loss of lives. These well-labeled tweets could be summarized (Olariu, 2014; Rudra, Goyal, Ganguly, Imran, & Mitra, 2019; Shou, Wang, Chen, & Chen, 2013; Z. Wang, Shou, Chen, Chen, & Mehrotra, 2014) or trimmed (Purohit et al., 2013) and delivered to research departments for further analyzing. The alert-level heat maps could also be illustrated based on the location information embedded in the tweets dataset.

In addition, tweets could also be used to diagnose the the fake news and to block their dissemination (Bondielli & Marcelloni, 2019; Bovet & Makse, 2019; Inuwa-Dutse, Liptrott, & Korkontzelos, 2018; B. Wang & Zhuang, 2018). Once fake news and unauthorised posts are detected before the widely spread on the social media platform, they could be labeled as junk posts or block their accounts. What is more, analysing the tweets based on NLP techniques could help to extract the sentiment of public about the crisis, and have a better understanding of the dissemination level towards a specific event.

1.1.2 Novel Coronavirus (COVID-19)

Currently, the pandemic has caused over 120 million confirmed cases and almost 3 million deaths around the world (Worldmeter, 2020). All countries and regions are trying their hardest to prevent the dissemination of the disease by conducting variety of nationwide strategies including curfew and lockdown. As individuals, people have to work at home, keep social distance in public places and wear face mask in public transports. Due to the continuously increasing confirmed cases and quarantine rules, people become more active to express their sentiments through the social media platforms. Several hot topics regarding the Covid-19 have been accumulating large scale of discussions and arguments so far. For that reason, the metadata of Twitter are a worth referencing and studying resource for researchers to conduct researches in the field of Social Computing, such as behaviour analysis, topic modeling, fact-checking, sentiment analysis and analytical visualisation.

The large-scale data always play an important role in training machine learning models and the analysis in the next stage. The conclusion extracted form small-scale dataset could not be convincible enough due to the volume limitation of tweets and coverage of locations. To address the above problem, a large-scale Covid-19 related tweets called ***COVID19Tweets Dataset*** will be analysed in this research.

There are more than 310 million tweets in this dataset and it is accessible to public on IEEE DataPort (Lamsal, 2020b). The dataset is collected form March 20, 2020 to now and is still up to date.

1.2 Techniques of sentiment analysis

Artificial Intelligence (AI) has been widely used in the domain of Social Computing. In the recent few years, there are quite a few algorithms and methods highlighted in traditional AI models, in terms of Deep Learning and Machine Learning. The overview of AI methods are presented in Fig.1. The Machine Learning method is classified into three different learning models including Supervised Learning, Unsupervised Learning and Reinforcement Learning. Deep Learning is a subset of Machine Learning in AI that has networks capability of learning from unsupervised data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network. Deep Learning has been universally utilized in a variety of field and achieved many notable outcomes. The most highlighted advantage of Deep Learning is that it could process and make sense of different kinds of datasets, even unstructured and large-scale datasets. For instance image, audio and video data.

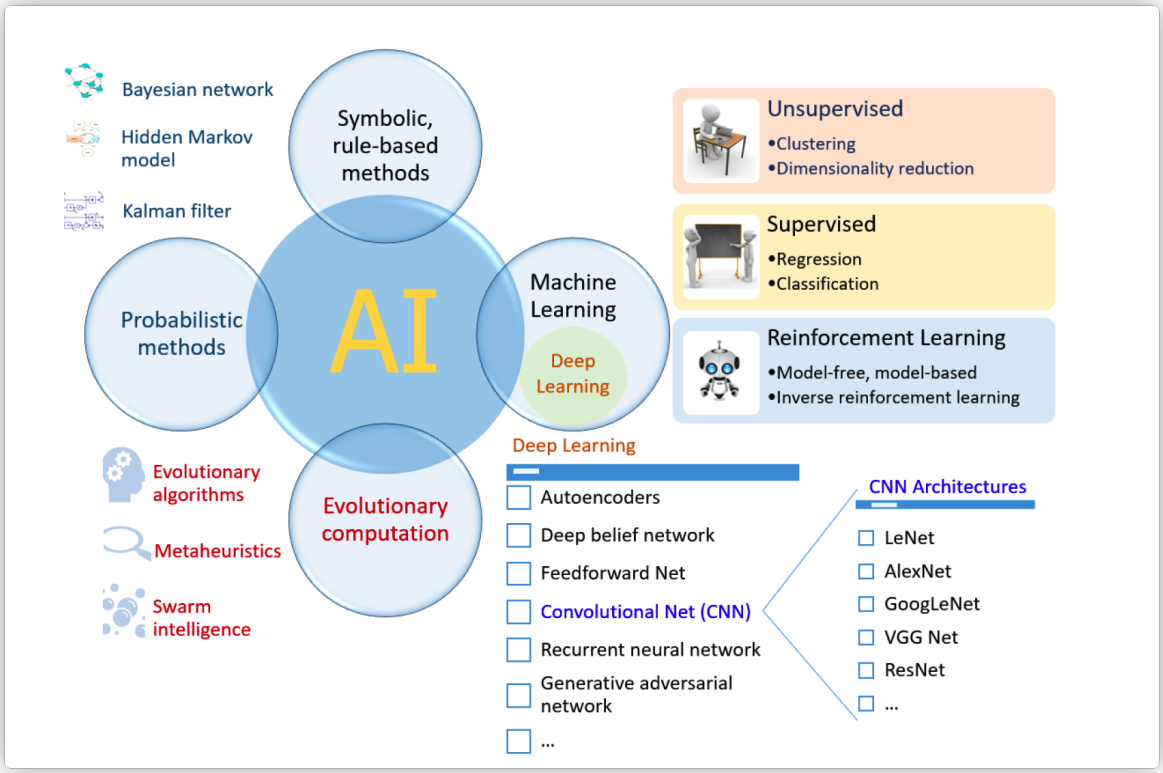


Fig. 1. The overview of AI methods (*Figure from Hussain et al. (2020)*).

1.2.1 Supervised Learning

Supervised Learning is a relatively common scenario which is adopted in healthcare and clinical areas (Patro, Padhy, & Chiranjevi, 2020; Y.-H. Hu, Wu, Lo, & Tai, 2012; Maes, Twisk, & Johnson, 2012). The method could extract the specific standard and patterns from the fed training dataset, which has been pre-labeled the output value of an input value. In the next stage, the well-trained model will be used to predict the output of the testing dataset. Fig.2 demonstrates the whole process of Supervised Learning method. The training data and testing data collected form COVID-19 related tweets are as input X, and the sentiment types are as output Y. The CovidNet model is trained by the well-labeled training dataset, and then the well-trained model will be used to predict the testing dataset and label the tweets with positive symbol or negative symbol.

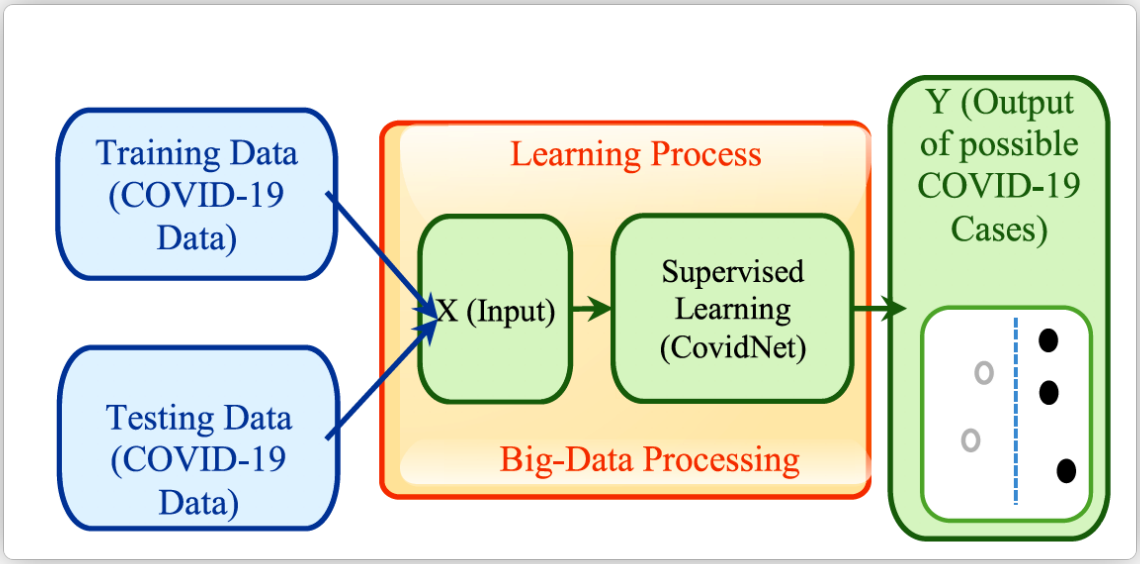


Fig. 2. The workflow of supervised learning method (*Figure from Hussain et al. (2020)*).

1.2.2 Unsupervised Learning

Unsupervised Learning is another Machine Learning method which is different form the Supervised Learning. There is no necessary to pre-train the model using training dataset at the initial stage, which is the substantial difference with Supervised Learning. Unsupervised Learning method is usually used to analyse the veiled data and categorized them into different clusters depending on the features. K-means as a member of Unsupervised Learning methods, is widely applied in medical system to monitor the actions of patients (Gozes et al., 2020). As the workflow presented in Fig.3, the COVID-19 related tweets as input X, and then the model will classify the input tweets into two clusters (positive and negative), which is pre-labeled manually. The clusters generated by the model are described as Y.

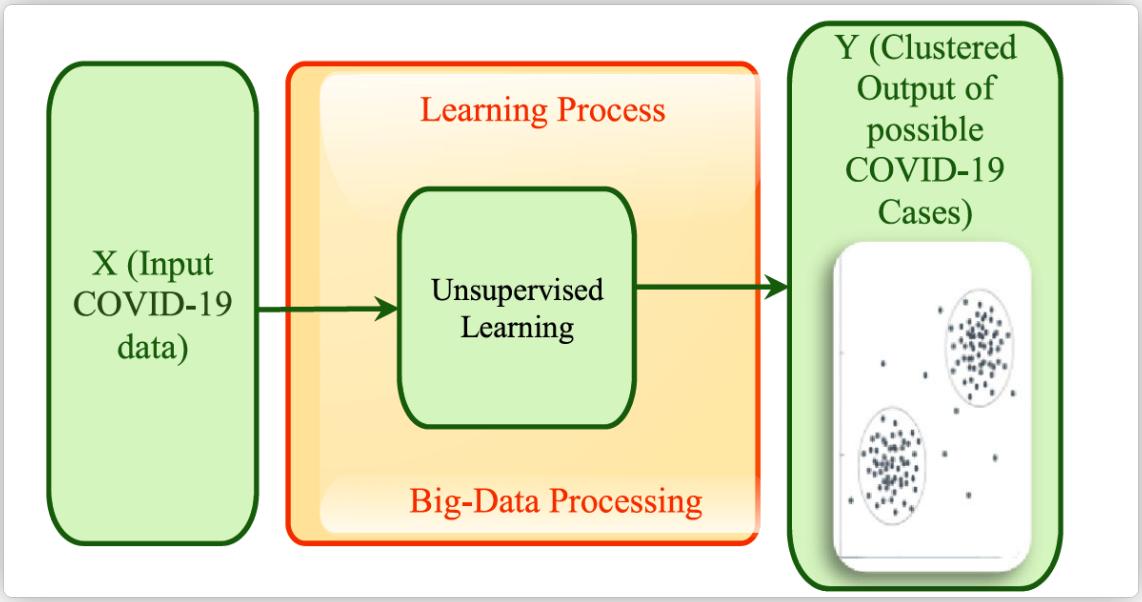


Fig. 3. The workflow of unsupervised learning method (*Figure from Hussain et al. (2020)*).

1.2.3 Reinforcement Learning

Different form Supervised and Unsupervised Learning methods, Reinforcement Learning method mainly focus on how software agents could perform in an model to maximize the notion of accumulative reward. This method do not need the pre-labeled training dataset to train the model, instead, it primarily focus on exploring the balance between exploitation (of current techniques) and exploration (of uncharted domains). It is a wise choice for the increasingly groundbreaking solutions in the medical domains, in which detection choices or treatment methods are usually described by the successive and delayed strategy (L. Li, Qin, et al., 2020). In Fig.4, the training dataset collected form COVID-19 related tweets as input X, then the raw dataset is fed to semi-supervised model DarkCovidNet. The testing data as input fed to DarkCovidNet model and the sentiment types are generated as output Y.

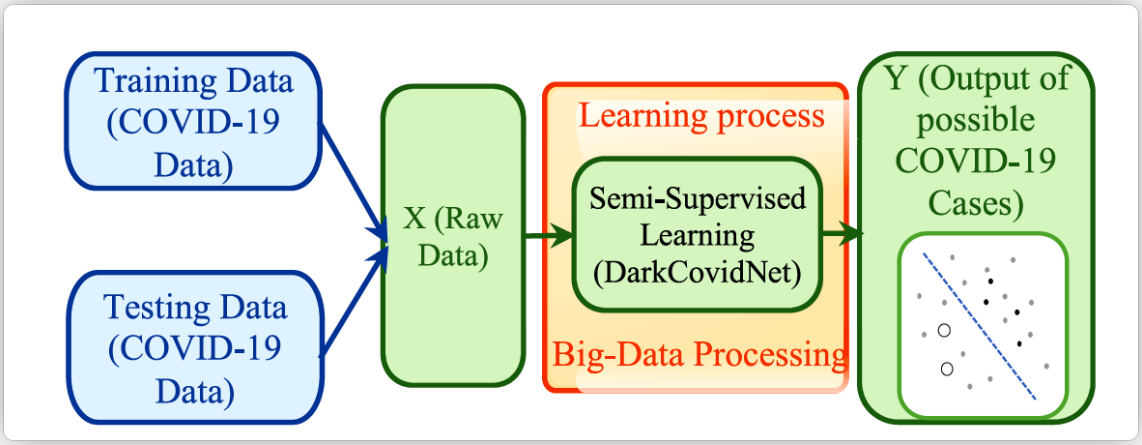


Fig. 4. The workflow of reinforcement learning method (*Figure from Hussain et al. (2020)*).

1.2.4 Deep Learning

Deep Learning is a subset of Machine Learning techniques that mimic the functions of the human brain in processing data and exploring patterns for application in decision making. The technology of Deep Learning could be used to predict the developing trend of COVID-19 (Naudé, 2020). Fig.5 illustrates the how the model works with three layers including input layer, hidden layer and output layer. The X's are inputs, and W's are weights per neuron. The output Y is calculated based on utility function f as follows: $y = f(u) = \sum_{i=1}^{N}(w(i)x(i) + b)$, b is the bias, and u represents the internal signals between the neurons. This model is roused by the human cerebrum and comprises of different associated neurons. The system comprises of a layer of information neurons that is the input neuron, and a layer of yield neurons that is the output neuron and various alleged shrouded layers in the middle, known as the hidden layer.

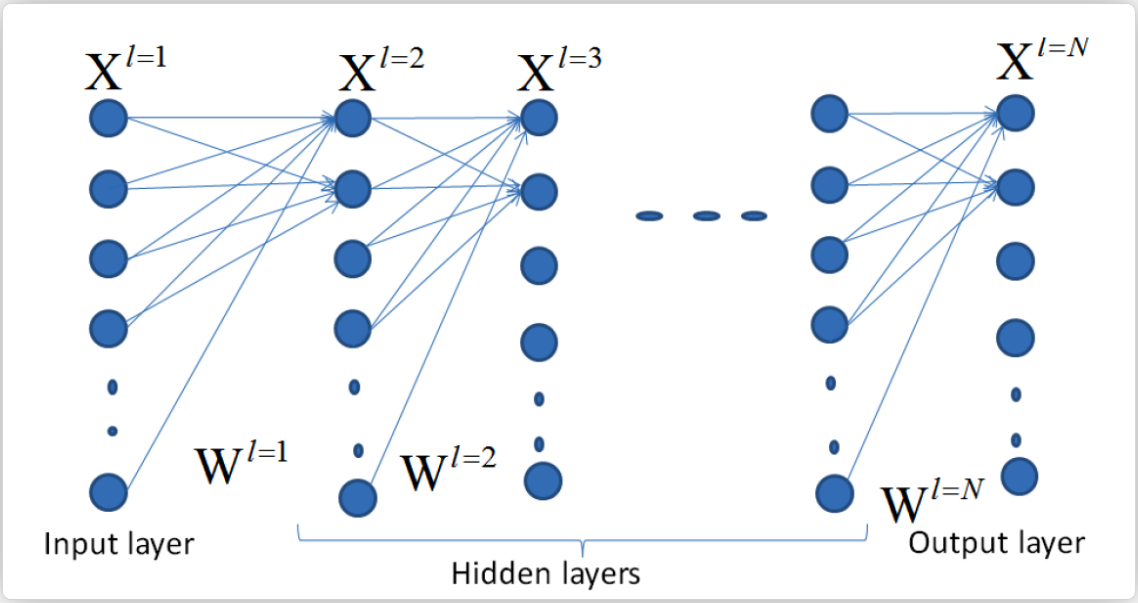


Fig. 5. The workflow of neural network (*Figure from Hussain et al. (2020)*).

1.2.5 Amazon Comprehend

Different from traditional Machine Learning methods, the E-commerce giant Amazon provide a novel Cloud Computing Services solution called Amazon AWS, in which Amazon provides very comprehensive functions to meet the different needs of end users. For example, Amazon EC2 provides the virtual servers in the cloud, Amazon S3 provides the scalable storage in the cloud, Amazon Comprehend provides Natural Language Processing service with built-in Machine Learning methods.

Amazon Comprehend could recognize the significant information in dataset, in terms of places, people and references to language, moreover, it could also classify the text files into different clusters depending on the similar topics. What is most important, Amazon Comprehend could automatically and accurately detect the sentiments of the content in real time. Amazon Comprehend is fully managed, thus you could get up and running quickly, without having to train models from scratch. Start processing millions of documents in minutes by leveraging the power of Machine Learning. Fig.6 presents the general processing flow of Amazon Comprehend.

Amazon Comprehend uses a pre-trained model to analyze and examine a document or set of documents to extract insights about it. The model is continuously trained with a large volume of input dataset so that there is no need for you to provide training data. When you feed your input data to

Amazon Comprehend, the model will call the API functions, they will automatically analyse and extract useful information from the text.

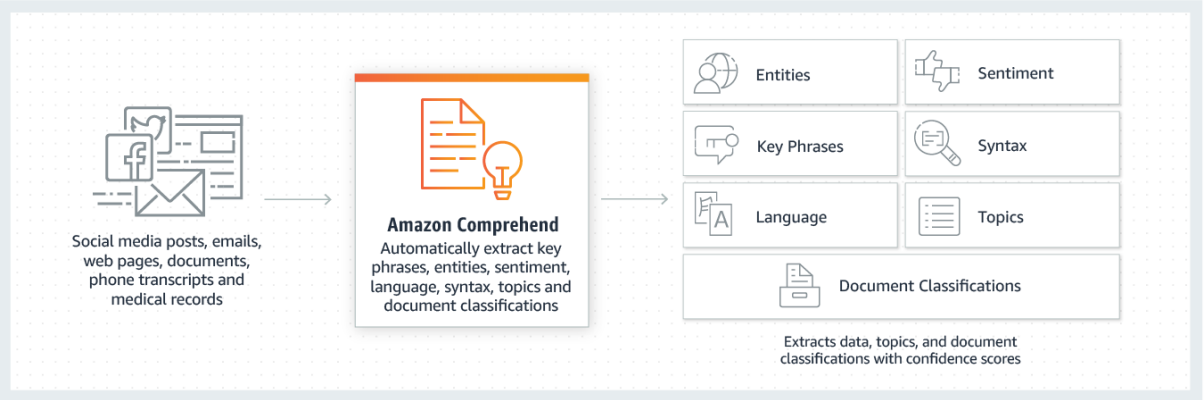


Fig. 6. The workflow of Amazon Comprehend.

1.3 Objective of the research and research questions

The objective of this research is to extract public sentiment and concerns from the tweets dataset during the outbreak of COVID-19 by using data analysis techniques including ML, NLP and Cloud Computing service Amazon AWS. To achieve the objective, firstly, we should identify the possibility of extracting public sentiment and concerns from tweets dataset. There were two outbreak waves of COVID-19 last year, the first wave is from mid-March to the end of June, and the second wave is form mid-September to the end of December. In the next stage, we will explore the differences of public sentiment between the firs and second waves of the COVID-19.

- **RQ1:**

  Can we identify the general public healthcare concerns related to COVID-19 pandemic through mining social media – the case of Twitter?

- **RQ2:**

  What are differences of public sentiment between the first and second waves of the COVID-19 pandemic?

1.4 Overview of the research methodology

To achieve the research objective, a combination method is applied in this research, a systematic literature review combines with the quantitative data analysis approach based on the state-of-the-art Machine Learning, Natural Language Processing techniques and Cloud Computing platform. Reviewing the topic

related literature could help us to find the limitations and research gaps within them in order that we could try to address them in our research. Quantitative analysis is a method which uses statistical and mathematical modeling, research and measurement to understand the behavior of data. The method has been widely used in many scenarios in terms of performance evaluation, measurement, and predicting real-world events.

There are several tools adopted in this research to conduct the systematic literature review, such as AUT Library, Google Scholar, IEEE, Scopus, etc. The detailed content of literature review will be talked about in section 2. In order to extract valuable information from the tweets dataset, the quantitative analysis method is introduced into this research. There are four steps in this approach: data collection, data pre-processing, data analysis and data visualization. The specific aspects of the quantitative analysis method will be presented in section 3.

1.5 Significance of the study

After the outbreak the COVID-19 around the world, many countries and regions has been suffering the disordered situation because of the increasing number of confirmed cases and deaths. In order to prevent the continuous large-scale dissemination of the virus, many counties have to introduced the restrictive rules and quarantine measures. For example, New Zealand instantly promulgated lockdown rules when the first community case was confirmed. Auckland city followed the level 4 restrictive rules and others followed the level 3 restrictions (Oxholm, Rivera, Schirrman, & Hoverd, 2021). During the lockdown period, people's movement and business activities were restricted: individuals had to work at home, all schools and commercial places were closed besides supermarket where people could buy essential stuff.

As billions of lives were affected by the epidemic, social media platforms such as Twitter and Facebook have become an channel for the public to share their opinions, concerns and feelings regarding the COVID-19 related events. Social media has become an indispensable and significant platform for the public to share the health related information, and a significant proportion of people have used social media to post healthcare concerns during the period of COVID-19 (Organization et al., 2020; Shannon & Kent, 2020). It has been confirmed in the research that social media as a dominant role of healthcare related information sharing is indisputable (Silver, Huang, & Taylor, 2019). With the fast prevalence of social media, sharing and consuming healthcare related information has become a kind of normality.

By knowing the general population reaction to COVID-19 and all the introduced precaution by governments and scientists, we can understand people emotions and identify their real concerns about the COVID-19. This can help governments to come up with more effective strategies and action response

to address general population concern. Analysing the healthcare related data sharing on social media platforms not only could help us to grab the firs-hand information of health crisis but also enable quicker access to the real-time situation which could help the governments and organizations to make more appropriate and efficient policies and measures.

1.6 Structure of the research

There are 5 sections in this research. The first section presents the introduction of this research including background information and knowledge of this research, the objective and research questions of this paper and the objective and significance. In the second section, related literature will be talked about in detail. The third section will illustrate the methodology of this research in terms of data collection, data pre-processing, data analysis and data visualisation. Followed by the fourth section, finding and results will reported in this section. In the last section, in-depth and extensive discussion will be demonstrated and the conclusion and future work as well.

2 Literature Review

In this section, related researches will be comprehensively reviewed. The applications of sentiment analysis will be talked in the first subsection, followed by the applications of tweets analysis. Research gaps will be presented in the last subsection.

2.1 The applications of sentiment analysis

Sentiment analysis (also known as emotion AI or opinion mining) is an analysis of using text analysis, natural language processing, biometrics and computational linguistics to systematically extract, identify, study and quantify subjective information and emotional states. Sentiment analysis is widely utilised in many domains such as marketing, finance, clinical medicine, customer service, politics etc.

2.1.1 Sentiment analysis and customer service

With the rapid growth of customer-generated reviews on the internet. Sentiment analysis of customer reviews has become a urgent need for service and product providers to understand the general sentiment and opinions of the customers and help them to make better decisions. A novel unsupervised and domain-independent model was proposed by (Bagheri, Saraee, & De Jong, 2013) to identify implicit and explicit aspects in reviews for sentiment analysis, the novel model could be easily used in the environment of non-English languages and other domains without pre-training the model with labeled dataset. Paddeu et al. conducted a research to explore the role of sentiment analysis for the application of customer loyalty analysis (Paddeu, Fancello, & Fadda, 2017). They launched a sentiment analysis on the comments of 120 banks' customers, the results showed that the customer loyalty was positively related with customer satisfaction and negatively related with customers' intention of leaving. Another related research conducted by Candelieri et al. in a transportation company, they proposed a model based on Support Vector Machin (SVM) to extract the insights between the commuters' sentiment and their services. The result of the research study could help the company to improve their services which could actually meet the commuters' need (Candelieri & Archetti, 2015). In another research, Mohsan et al. did a research to find out the role of sentiment analysis for marketing application. They analysed the thoughts and tweets related to their products and services, in order that they could launce more targeted and effective marketing promotion campaigns to build a good reputation (Mohsan, Nawaz, Khan, Shaukat, & Aslam, 2011).

2.1.2 Sentiment analysis and stock market

Understanding the polarity of stock market related news could help investors to make more appropriate decisions in stock market. To achieve the goal, an automatically dictionary construction approach was

proposed by (Mizumoto, Yanagimoto, & Yoshioka, 2012), the approach was based on semi-supervised learning and could determine correct polarities of 45% of all news. In another research, Ren et al. presented a machine learning method based on support vector machine to analyse the investor-generated textual content on the Internet and they achieved the accuracy of forecasting the movement direction of the SSE 50 Index can be as high as 89.93% (Ren, Wu, & Liu, 2018). Their findings could assist investors in making wiser decisions and they also imply that sentiment probably contains precious information about the asset fundamental values and could be regarded as one of the leading indicators of the stock market. Predicting the stock price is a challenge task because the stock price is affected by many factors. Nguyen et al. proposed a method based on Support Vector Machine (SVM) with the linear kernel. The proposed method could predict the stock price movement with more than 60% accuracy for a few stocks, and performs much better than other methods for the stocks that are difficult to predict with only past prices (T. H. Nguyen, Shirai, & Velcin, 2015). In another research, a combination method was proposed by Khedr et al., naïve Bayes algorithm was used to identify the polarity of the news in the first stage and the second stage incorporates the output of the first stage as input along with the processed historical numeric data attributes to predict the future stock trend using KNN algorithm (Khedr, Yaseen, et al., 2017). They model for predicting the future behavior of stock market obtained accuracy up to 89.80%.

2.1.3 Sentiment analysis and politics

A large number of researches have tried to predict the election results through analyzing the sentiment of social media platforms. Anjaria and Guddeti (2014) used five supervised machine learning classifiers to classify the sentiment dataset of social media platform for 2012 US presidential election and 2013 Karnataka state assembly election. Their research revealed that Support Vector Machine achieved the highest accurate score within the five machine learning classifiers for both of the elections. Ceron, Curini, and Iacus (2015) analyzed two election campaigns (2012 Italian Centre-Left Coalition election and 2012 US presidential election) on social media platform using sentiment analysis. The research applied the Hopkins and King (HK) method (Hopkins & King, 2010), which is supervised machine learning method to extract the voting intention of citizens. HK algorithm analyzes all the word vectors of the test datasets to evaluate the aggregate distribution of opinions directly rather than evaluating the sentiment of individuals. In this research, they acquired better understanding of textual dataset and achieved more reliable and accurate outcomes. In another research, a context and semantic based method was proposed by Singhal et al. to analyze sentiment of Twitter users for National Capital Territory (NCT) Delhi elections 2014 (Singhal, Agrawal, & Mittal, 2015). A hybrid model (dictionary-based and contextual rule based) was utilised in this study, which achieved more accurate score than other existing methods. Meanwhile, the novel method could also to predict explicit and implicit negation, conjunctions and positive result and handle sarcasm.

2.2 The applications of tweets analysis

Tweets are mostly used these days for expressing the views or opinions about any topic, product, event, or any breaking news from anywhere at anytime. Thus tweet dataset is meaningful and valuable for researchers and analyst to analyze and extract useful information and patterns.

2.2.1 Tweets analysis and emergency response

The existing research has assessed the performance of Twitter dataset in situation analysis under emergent circumstances, for example, identifying bushfire-related Tweets in San Diego County by analyzing more than 41,000 correlated Tweets in 2014 (Z. Wang et al., 2016). The outcome of this research found that the highest frequency (six of nine) of wildfires emerged in May 14 and the largest post-zones were in the business district of San Diego. This research illustrated that there is geographical association with tweets. In another research, Chatfield et al. (2013) analyzed and studied Indonesian Agency for Meteorology, Climatology Geophysics (BMKG), the central disaster agency and the upstream supplier of Indonesia's disaster information value chains, and its Twitter Tsunami Early Warning Civic Network as an integral part of the agency's RII. They found clear evidence which BMKG tweets have an effective mechanism for the direct participation of net-savvy resident in early warning of tsunami. A novel framework was proposed by (Cheong & Lee, 2011), they applied Twitter microblogging service as a multifaceted data source to conduct demographic analysis and sentimental data in public response to terrorism activities. The outcomes of their research could help the law enforcement agencies and homeland security authorities to make quicker and proper response to terror threats. Landwehr et al. (2016) proposed a real-time scenario called TWRsms, which could collect and analyze the tweets in real time, in order to provide early tsunami warning to the public. Location information could also be collected form the real-time tweets, which is quite valuable for government and organisations to conduct rescue and evacuation activities.

According to Takahashi et al. (2015), a typology of Twitter use was tested to examine Twitter use during and after Typhoon Haiyan pummeled the Philippines. The outcomes show that different stakeholders used social media mostly for dissemination of second-hand information, in coordinating relief efforts, and in memorializing those affected. Z. Wang et al. (2014) analyzed the wildfire-related Twitter activities including their attributes pertinent to space, content, time, and network, so as to gain insights into the usefulness of tweet data in revealing situational awareness. Their outcomes show that tweet could characterize the wildfire across space and over time, and thus are applicable to provide useful information on disaster situations. Second, people have strong geographical awareness during wildfire threats and are interested in communicating situational updates related to wildfire damage (e.g., burned acres and containment percentage), wildfire response (e.g., evacuation), and appreciation

to firefighters. Third, news media and local authorities are opinion leaders and play a dominant role in the wildfire retweet network. Zahra et al. (2020) proposed a machine learning classifier, which was trained by characteristics and labeled data. The results showed that real-world Twitter datasets reveal textual features (bag-of-words) when combined with domain-expert features could achieve better classification performance. Their approach provided a successful scenario for combining crowd sourced and machine learning analysis, and improved our understanding and capability of detecting valuable eyewitness reports during disasters. In another related research, the spatial–temporal patterns of Twitter activities during Hurricane Sandy was analyzed by (Zou et al., 2018). 126 counties affected by Hurricane Sandy were contained in this research, and findings show that common indexes derived from Twitter data, including normalized ratio, ratio, and sentiment, could enable comparison across regions and events and should be documented. The insights gained from this research could provide useful information into strategies for applying tweets dataset to increase resilience to disasters.

2.2.2 Tweets analysis and public opinion

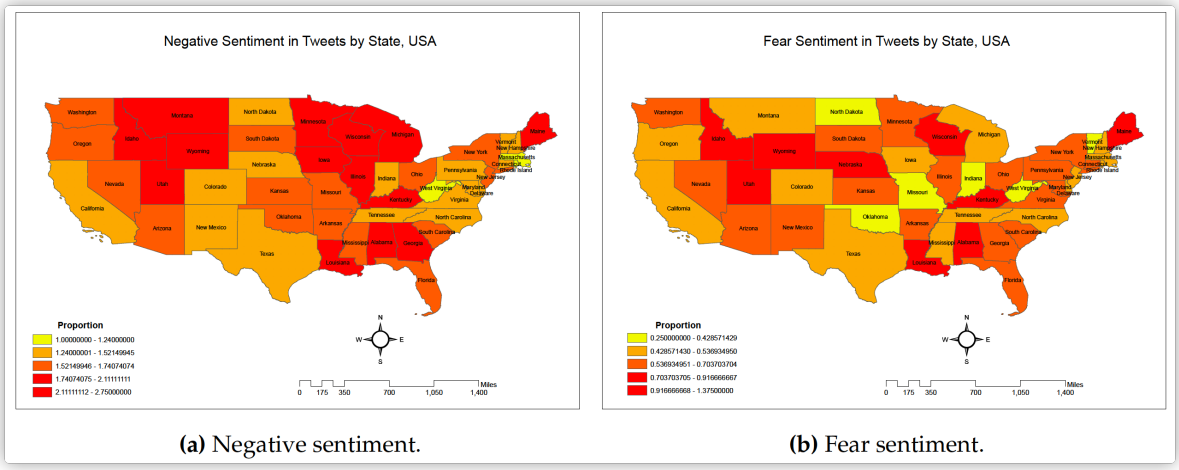
Tweets could be utilised as data source for mining the general opinions of the public on a wide variety of domains such as polls, election, climate change, vaccination, customer service etc. According to O'Connor, Balasubramanyan, Routledge, and Smith (2010), they used NLP techniques to analyze some surveys regarding public political opinion and confidence over the 2008 t0 2009 period, and they found some correlations between polls related tweets and sentiment word frequencies. Their results showed the correlation as high as 80%, and captured critical large-scale trends as well. As regards their outcomes, it reveals the possibility of social media platform such as Twitter as a supplement or substitute for traditional polling activities. D'Andrea, Ducange, Bechini, Renda, and Marcelloni (2019) established a intelligent model to automatically detect general treads in public opinion in regard to their stance towards the vaccination. Their model make it possible to monitor the significant opinion variation of public, which could be possibly explained with the happening of specific social context-related events. From multiple combinations of different text representations and classification approaches, they achieved the best accuracy by adopting the method of bag-of-word, with stemmed n-grams as tokens, SVM model for classification. The proposed novel model could be used by the authorities to monitor and track public opinions related vaccination decision making in a real-time and low-cost way.

The dictionary-based method (also called lexicon-based approach), utilises a large volume of pre-labelled words to extract the semantic orientation of the text (Thelwall, Buckley, & Paltoglou, 2011). For instance, M. Hu and Liu (2004) produced a large number of adjective synonyms and antonyms (opinion words) via bootstrapping process applying the WordNet dictionary. In the next stage, the collection of opinion words were used to detect the sentiment polarity of electronic product reviews at a sentence level.

The dictionary-based approach typically counts the numbers of negative and positive opinion words in one sentence. If negative opinion words prevail, the orientation of the sentence is negative and otherwise positive. Sentiment analysis can be used not only at a sentence level, but also at others levels: paragraph-, document-, or attribute-level. As the level of granularity increases, so does its complexity. Attribute-level sentiment analysis aims to associate opinions associated with certain features (Chiu, Chiu, Sung, & Hsieh, 2015). In another research, a large-scale dataset of geo-tagged tweets, which contain specific keywords related to climate change, were analysed by applying text mining and volume analysis techniques such as sentiment analysis and topic modeling (Dahal, Kumar, & Li, 2019). The used these technique to contrast and compare the public opinion of climate change between different countries. The results showed that overall sentiment was negative, especially when people were reacting to political or extreme weather events.

2.2.3 Tweets analysis and COVID-19

With respect to COVID-19 pandemic, a research study conducted by Samuel et al. showed the relationship between COVID-19 related Tweets and locations, which was presented in Fig.7. Such relationship can be explained to some extent by the fact that people in urban areas have better access to information and communication technologies, resulting in a higher number of tweets from urban areas (Samuel et al., 2020). A research study was done by Boberg et al. to analyse the public fears on the social media application such as Twitter and Facebook in the early stage of COVID-19 (Boberg, Quandt, Schatto-Eckrodt, & Frischlich, 2020). In another search, COVID-19 related tweets were analysed by Jahanbin et al., in order to identify the public response to the epidemic over time (Jahanbin, Rahmanian, et al., 2020). Li et al. intend to acquire a insight about the situational information of the pandemic and explore the developing trend on the social medias including Twitter and Facebook (L. Li, Zhang, et al., 2020).



(a) Negative sentiment. **(b)** Fear sentiment.

Fig. 7. The sentiment proportion in each state of USA. Figure from Samuel et al. (2020)

According to Mohammad et al., they used Naïve Bayes algorithm to analyse the Arabic tweets related to COVID-19. Their findings revealed that the majority of public hold positive sentiment towards the pandemic, which could help the government to acquire the insights of public response to the epidemic and overcome the difficult situation (Alhajji, Al Khalifah, Aljubran, & Alkhalifah, 2020). Sharma et al. proposed an in-depth model on shifting topics with regard to trends and sentiments to identify the false and fake information about COVID-19 spreading on Twitter (Sharma et al., 2020). Open marking annotation was left to optimise the classification techniques in this research. Their outcomes could help to identify and elimination the fake and false information of COVID-19 to avoid further deterioration of the current situation in a early stage. Ghafarian and Yazdi (2020) presented a novel model to identify informative tweets by utilising distributional assumptions. In their model, every single tweet is considered as a "distribution", and significant and meaningful outcomes were achieved in detecting informative tweets towards a crisis event.

2.3 The state-of-the-art techniques towards sentiment analysis

Stance and sentiment analysis could be accomplished by a variety of methods: lexicon-based, machine learning and hybrid approaches (Medhat, Hassan, & Korashy, 2014). With regard to texts, they could be fed to classification models in different ways which should be able to encode the text with the prospective precision. Considering different scenarios, text could be regarded as vector of numbers, for instance, word embeddings (Tang et al., 2014), bag-of-words (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015), and hybrid methods (Mohammad, Sobhani, & Kiritchenko, 2017). Machine learning methods contain supervised and unsupervised, which could automatically analyse and extract patterns from the texts. Deep learning as a kind of machine learning method has been widely used in sentiment analysis in recent years (Zhang, Wang, & Liu, 2018). In contrast, lexicon-based approach rely on predefined lexicon, for example SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), WordNet (Miller, 1995) and SenticNet (Cambria, Havasi, & Hussain, 2012).

In the following, we will review recent researches related to state-of-the-art approaches and techniques of sentiment analysis. As regards machine learning approaches, supervised learning method is the predominant solution in recent related works. Chen and Tseng (2011) used SVM to evaluate the quality of information in product reviews. To help the company make proper marketing decisions, a model for summarisation has been proposed and SVM classification of sentiments on tweets (Y.-M. Li & Li, 2013). Naive Bayes was applied by H. Wang, Can, Kazemzadeh, Bar, and Narayanan (2012) to analyse the tweets of 2012 USA election in real-time, aiming to infer the general public sentiments of the president candidates. In another research, Aisopos, Papadakis, and Varvarigou (2011) presented a different method of text representation, using n-gram graphs with distance-weighted edges, and employing two classifiers

(C4.5 decision tree and MNB) to carry out two types of classification (two-way and three-way) about tweets posted in the last seven months. According to Valdivia, Luzíón, and Herrera (2017), the accuracy of classification, which achieved from a low-level model consisting of SVM classification and text elaboration, could be approved by utilising a majority vote to several approaches for filtering neutral sentiments out.

Deep learning as a kind of machine learning has also been widely used in sentiment analysis. Among deep learning methods applied in the related domain, Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Convolutional Neural Networks (CNNs) (LeCun, Bengio, & Hinton, 2015) regarded as the state-of-the-art methods. In Cliche (2017), the authors pre-trained three popular unsupervised machine learning architecture (Word2vec, GloVe and FastText), feeding with 100 million unlabeled tweets. A distant supervision feeding with 5 million negative and 5 million positive tweets was used to fine-tune the word vectors to enrich the word representation with polarity information. In the next stage, the fine-tuned word vector was used to initialise the CNN and LSTM model. Their well-train model achieved significant outcomes and performance in the SemEval-2017 International Workshop on Semantic Evaluation, task 4 (Sentiment analy- sis in Twitter) (Rosenthal, Farra, & Nakov, 2017). In another related research, Xiong, Lv, Zhao, and Ji (2018) studied sentiment-related word embedding through applying both distant supervised and lexicon information. A word representation combining word-level sentiment (such as lexicon information) and tweet-level sentiment (such as emoticon and hashtag) were fed to several neural networks to achieve a multi-level sentiment-enriched word embeddings. With regard to sentiment analysis in Twitter, Dey, Shrivastava, and Kaushik (2018) presented a two-stages text classification scenario. In the first stage, a specific tweet was classified as subjective or neutral regarding the given topic. In the next stage, the subjective tweet was classified as against or in favor about the topic. In the two stages, LSTM was used as classification model. To achieve a significant result applying Deep Learning methods, massive training works should be conducted to support the learning process in advance.

2.4 Research Gap

As regards the topic of sentiment analysis, a bunch of related literature and researches has been deeply reviewed in the previous section. However, there also are some limitations and gaps in their works, and these limitations and gas will be talked about in this section.

2.4.1 Health concerns

In the previous topic related literature reviews, sentiment analysis and tweets analysis are widely applied in a wide variety of domains in terms of customer service, stock market, politics, emergency response

and public opinions. Nevertheless, there are limit researches talk about how to use sentiment analysis techniques to extract health concerns of the public. Health has always been a topic that people pay more attention to, particularly after the outbreak of COVID-19. The clinical data manifest that people infect with COVID-19 are characterized by cough, fever, dyspnea, and bilateral infiltrates on chest imaging (Yang et al., 2020). COVID-19 outbreak could cause physical health problems as well as mental health issues. Fear of being infected due to close contact with confirmed patients, prolonged working schedules without enough rest, and disturbed wake and sleep routines have increased the risk of stress and anxiety in the healthcare workers (Khan et al., 2020).

What is more, although several researches has been conducted in the sentiment analysis of COVID-19 related tweets. Such as extracting relationship between COVID-19 related tweets and locations (Samuel et al., 2020), analysing public fears on tweets (Boberg et al., 2020), extracting developing trends of COVID-19 (L. Li, Zhang, et al., 2020) and identifying fake and false information of COVID-19 spreading on Twitter (Sharma et al., 2020). Limited researches pay their attention on public health concerns as well. In this research, we will try to use state-of-the-art techniques extracting public health concerns from COVID-19 relate tweets.

Furthermore, little researches try to examine the differences between the first wave and second wave of COVID-19. As we know, there are two outbreak waves of COVID-19 in 2020. The first wave is form from mid-March to the end of June, and the second wave is form mid-September to the end of December. During the two waves of the epidemic, people's health concerns may have some changes due to they already know enough knowledge about the pandemic and know how to prevent it to protect themselves. So it is meaningful and necessary to explore the differences of public sentiment between the first and second waves of the COVID-19 pandemic, which could present an insight to the authorities and organisations to help them optimise and improve their strategies and response towards the global crisis.

2.4.2 Technique limitations

- **Data Sensitivity:**

  The previous related works mainly focus on analysing the sentiment of tweets by using machine learning and deep learning techniques. However, these techniques highly rely on structured data, and in the real world, almost 80% of data is unstructured. In order to process the unstructured data, many fussy works should be conducted manually. In cloud computing scenario such as Amazon AWS and Microsoft Azure, this problem could be addressed properly by using Data Lake. Data Lake is a repository or system that stores data in its original format. It stores the data as it is, without the need to structure the data in advance. A data lake can store structured data (such as tables

in relational databases), semistructured data (such as CSV, logs, XML, JSON), unstructured data (such as emails, documents, PDF), and binary data (Such as graphics, audio, video).

AWS data lake can be divided into three stages to process data. The first stage of batch processing: By loading various types of raw data on Amazon S3, and then processing the data in the data lake through AWS Glue, you can also use Amazon EMR for advanced data processing and analysis. The second stage of stream processing and analysis, this task is based on Amazon EMR, Amazon Kinesis to complete. The third level is machine learning. Data is processed in depth through Amazon Machine Learning, Amazon Lex, and Amazon Recognition to form usable data services.

– **Scalability:**

Traditional sentiment analysis tools commonly run on domain host, the performance of data processing highly rely on the hardware such as CPU, memory, GPU, hard disk etc. The users will spend lots of money and time to upgrade their devices. However, in cloud scenario, users could establish their own virtual computer via cloud services depending on their actual needs. The users could manually choose the hardware and software they will used in the task, such as CPU, GPU, memory, operating system, APIs and Networking features.

– **Cost:**

The sentiment analysis work commonly require specialised knowledge and devices, that is really a huge investment for mid-size and small companies to deploy. Cloud services give them a chance to conduct their requirements. For instance, the AWS charge the fees based on the actual usage of the task. This service is known as a "Pay-as-you-go" model.

3 Methodology

There are several approaches to achieve the objectives of this research including research methodology, data analysis methodology. In this section, the two methodologies will be talked about in detail followed with the reason of adopting the two specific methodologies. There are three subsections in the methodology part in terms of research methodology, related literature review and quantitative data analysis.

3.1 Research methodology

There are three types of approaches towards research methodology including quantitative, qualitative and hybrid-method (Creswell & Creswell, 2017). The quantitative research method uses the statistic techniques to extract the specific pattern and meaningful information from a great volume of dataset. Qualitative research method uses qualitative techniques to analyse the words (spoken and written language). Quantitative approaches are applied to justify the theory, while the quantitative ones are utilised to generate theory (Braun & Clarke, 2013). The hybrid-method is consist of both quantitative and qualitative methods in a single research to achieve a extensive and deeper insights and understandings (Johnson, Onwuegbuzie, & Turner, 2007). We adopt the quantitative methods to answer the research questions presented in previous section.

3.1.1 Quantitative research method

Quantitative approaches highlight objective measurements and the mathematical, statistical, or numerical analysis of data collected from questionnaires, polls, and surveys, or by processing pre-existing statistical data using computational techniques. Quantitative research focuses on collecting numerical data and generalizing it via groups of people or to explain a specific pattern (Muijs, 2010). The quantitative research method is applied in this research because its advantages listed below:

- The outcomes are based on larger volume dataset that are representative of the population.
- Data are in the form of statistics and numbers, often arranged in charts, tables, figures, or other non-textual forms.
- The method could be utilised to generalize concepts more widely, predict future results, or investigate specific relationships.
- The research study can usually be repeated or replicated, given its high reliability.
- The approach is to classify features, count them, and construct statistical models in an attempt to explain what is observed.

As mentioned above, quantitative research is data-oriented, there are two main methods of quantitative research: Primary quantitative research methods and Secondary quantitative research methods. Primary quantitative research is widely used in the researches of marketing, it mainly focused on collecting data directly rather than relying on data collected from previously published researches. There are several tracks included in primary quantitative research: Survey Research, Cross-sectional surveys, Longitudinal surveys, Experimental research etc. In contrast, Secondary quantitative research or desk research is a research method which uses already existing data or secondary data. There are five popular secondary quantitative research methods in terms of Data available on the internet, Government and non-government sources, Public libraries, Educational institutions, and Commercial information sources (Bhat, 2019). Due to the selection of appropriate quantitative research method depends on the research questions presented in the previous section, thus in our research, all the methods of primary quantitative research will be applied for literature review, and for secondary quantitative research, the Data available on the internet will be used in the data collection part.

3.1.2 Literature review

In this research, the already existing data source is considered applicable as several related studies have been conducted to illustrate the successful factors towards different technologies sentiment analysis method in different kinds of contexts adopting various research approaches. The issue is that there is little studies explore public healthcare concerns using sentiment analysis techniques form tweets, and most of studies based on traditional machine learning and NLP techniques which have some limitations in terms of data sensitivity, scalability and high cost. The method of this study is to explore public healthcare concerns from COVID-19 related tweets using sentiment analysis techniques and Amazon cloud services. What is more, using published literature has some advantages including time efficiency for collecting data, and cost-effective for easily accessing database through the internet (Rotmans, Kemp, & Van Asselt, 2001).

The procedures of reviewing related published literature and extracting useful information from those studies is called literature review process. Literature review is applied in this research, which could help to find out topic-related information of applications in different domains, to review and analyse the techniques utilised in their studies (Okoli & Schabram, 2010). Literature review could encapsulate the existing achievements, identify the limitations and gas within the research, establish a framework for achieving research purpose. Recently, there are several different kinds of approaches to conduct literature review, for instance, narrative review, critical review and systematic review (Grant & Booth, 2009). In this research, systematic literature review will be applied to conduct the review process.

According to Grant and Booth (2009), systematic literature review is (SLR) the most widely used method, which could be used to evaluate, search, and identify the literature systematically. The purpose of SLR is to collecting all existing knowledge and information related to a specific topic, and the most highlight advantage of SLR is that the process if repeatable because of its transparent feature in reporting of its approaches (Grant & Booth, 2009). To achieve the SLR process, some principles and guidelines you should follow with includes the NHS Centre for Reviews (Tacconelli, 2010) and Dissemination and he Cochrane Collaboration (Chandler, Cumpston, Li, Page, & Welch, 2019).

3.2 Systematic Literature review

There are two main reasons for selecting systematic literature review in this research. One reason is that it could capture all the knowledge and information regarding a specific topic from the online dataset, another reason it that the process is replicatable by other researchers for the reason that the reporting method is transparent (Grant & Booth, 2009).

There are three databases selected in this research including Google Scholar database, Scopus database and IEEE DataPort. Google Scholar and Scopus databases were used to search topic related literature and published articles. IEEE DataPort was use to find out COVID-19 related dataset. In order to grab as much literature from the selected databases, several keywords were identified based on the research questions of this study. All the used keywords are listed in Fig.8.

1. COVID-19	6. Novel Coronavirus	11. Machine Learning
2. Twitter	7. Healthcare Concern	12. Cloud Computing
3. Tweet	8. Social Media	13. Deep Learning
4. Healthcare	9. Tweets Analysis	14. Social Media Mining
5. AWS	10. Sentiment Analysis	15. Natural Language Processing

Fig. 8. The keywords of systematic literature review

3.3 Quantitative data analysis

After reviewing topic related literature, quantitative data analysis will be conducted in the next stage. There are four steps in quantitative data analysis including data collection, data pre-processing, data analysis and data visualisation. The detailed information will be given in the next subsections.

3.3.1 Data collection

As mentioned in previous section, the dataset used in this research captured form IEEE DataPort called ***CORONAVIRUS (COVID-19) TWEETS DATASET*** published by Rabindra Lamsal (Lamsal, 2020a). The submitted tweets dataset was from March 20, 2020 till now, but the dataset of March 29, 2020 was not available due to some technical faults. Among the whole datasets, we only used two period datasets: March 20, 2020 to June 30, 2020 and September 15, 2020 to December 21, 2020, which present the two outbreak waves of the COVID-19 pandemic. The keywords and hashtags of COVID-19 related tweets collection is list in Fig.9.

Keywords[a]
corona, #corona, coronavirus, #coronavirus
covid, #covid, covid19, #covid19, covid-19, #covid-19, sarscov2, #sarscov2, sars cov2, sars cov 2, covid_19, #covid_19, #ncov, ncov, #ncov2019, ncov2019, 2019-ncov, #2019-ncov, #2019ncov, 2019ncov
pandemic, #pandemic, quarantine, #quarantine, flatten the curve, flattening the curve, #flatteningthecurve, #flattenthecurve, hand sanitizer, #handsanitizer, #lockdown, lockdown, social distancing, #socialdistancing, work from home, #workfromhome, working from home, #workingfromhome, ppe, n95, #ppe, #n95

Fig. 9. The keywords of COVID-19 related tweets collection

The COVID-19 related dataset includes CSV files which only contain tweet IDs. Due to the content redistribution policy set by Twitter, users can not share Twitter data with others other than tweet IDs. Therefore, to acquire the complete tweet metadata, we have to use the tool ***Hydrator*** to hydrate the tweet IDs. Before hydrating operation, we should login our Twitter account and input the Access Key and Access Key Secret. The operation interface of ***Hydrator*** is demonstrated in Fig.10.

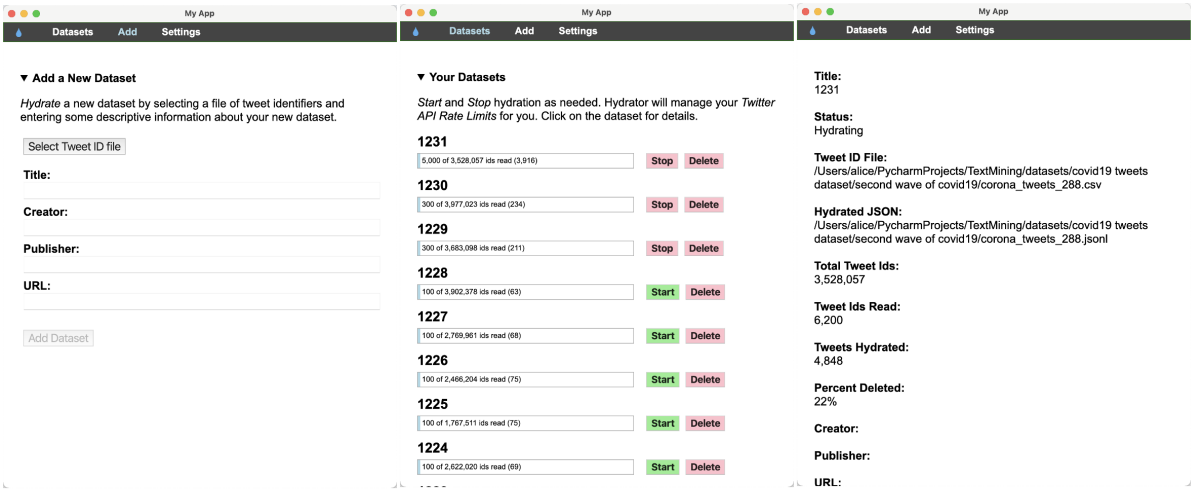


Fig. 10. The Hydrator operation interface.

3.3.2 Data pre-processing

After gathering all the metadata through hydrating operation, the following step is data pre-processing operation, which should be implemented in the initial stage to acquire better quality of dataset to commence the next stage of Data Analysis. Data Pre-processing includes removing non-English tweets, removing duplicated tweets, removing http links, removing bitly links, removing user mentions, replacing hashtags with corresponding words, removing punctuation and converting capital words to lower case. The Fig.11 illustrates the Python code of data pre-processing operation.

```
'''
Put all the tweets together and remove the duplicated tweets
'''
first_wave_tweets = pd.concat([tweet0320_p1, tweet0320_p2, tweet0320_p3], axis=0, ignore_index=True).drop_duplicates()
print(first_wave_tweets.shape)

'''
Choose the English tweets
'''
first_wave_tweets = first_wave_tweets.loc[first_wave_tweets["lang"] == "en"]
print(first_wave_tweets.shape)
# print(first_wave_tweets)

'''
Reset the index of the tweets
'''
first_wave_tweets = first_wave_tweets.reset_index(drop=True)
# print(first_wave_tweets)

'''
Extract the text of tweets
'''
first_wave_tweets = first_wave_tweets["text"]
# print(first_wave_tweets)

'''
Convert the tweets format to tabular data
'''
first_wave_tweets = pd.DataFrame(first_wave_tweets)
# print(first_wave_tweets)
```

```
def cleanTweetsAttribute(text):
    # Removing urls
    text = re.sub(r'http\S+', '', text)  # remove http links
    text = re.sub(r'bit.ly/\S+', '', text)  # remove bitly links
    text = text.strip('[link]')  # remove [links]
    text = text.strip('RT ')  # remove retweet sign 'RT'

    # Removing user mentions
    text = re.sub('(@[A-Za-z]+[A-Za-z0-9-_]+)', '', text)

    # Removing hashtags
    text = re.sub(r"#(\w+)", '', text, flags=re.MULTILINE)

    return text

def tweetsPreprocessing(text):
    text = cleanTweetsAttribute(text)

    # Converting to lowercase
    text = text.lower()

    # Removing punctuations
    text = text.translate(str.maketrans('', '', string.punctuation))

    return text

first_wave_tweets["text"] = first_wave_tweets["text"].apply(tweetsPreprocessing)
print(first_wave_tweets)
```

Fig. 11. The pre-processing operation of tweet metadata.

3.3.3 Data analysis

After the operation of Data Pre-processing, the process of Data Analysis will be conducted in this section. NLP method, machine learning and deep learning algorithms will be used to analyse the pre-processed

data to extract useful information, which will be fed to data visualization procedure. A series of libraries were utilized to achieve the purpose of Data Analysis, in terms of Pandas, re, nltk, collection, wordcloud etc.

– **The comparison of two waves' tweets volume:**

Analysing the trend of the volume of COVID-19 related tweets posted on twitter could give us a insight of how the clout of COVID-19 develops with the development of the pandemic. In this research, we will compare the developing trend of two waves' tweets volume posted by the users, which could show us the difference of the clout's shifting between the two period of COVID-19 outbreak. The Fig.12 and Fig.13 presents the detailed Python code.

```
'''
Read the daily tweets of the first wave of covid19 (03.20 ~ 06.30)
'''
tweet_volume = []
tweet0320_p1 = pd.read_csv(path + 'corona_tweets_0320_p1.csv')
tweet0320_p2 = pd.read_csv(path + 'corona_tweets_0320_p2.csv')
tweet0320_p3 = pd.read_csv(path + 'corona_tweets_0320_p3.csv')
tweet_volume.append(tweet0320_p1.shape[0] + tweet0320_p2.shape[0] + tweet0320_p3.shape[0])
tweet0321 = pd.read_csv(path + 'corona_tweets_0321.csv')
tweet_volume.append(tweet0321.shape[0])
tweet0322 = pd.read_csv(path + 'corona_tweets_0322.csv')
tweet_volume.append(tweet0322.shape[0])
tweet0323 = pd.read_csv(path + 'corona_tweets_0323.csv')
tweet_volume.append(tweet0323.shape[0])
tweet0324 = pd.read_csv(path + 'corona_tweets_0324.csv')
tweet_volume.append(tweet0324.shape[0])
tweet0325 = pd.read_csv(path + 'corona_tweets_0325.csv')
tweet_volume.append(tweet0325.shape[0])
tweet0326 = pd.read_csv(path + 'corona_tweets_0326.csv')
tweet_volume.append(tweet0326.shape[0])
tweet0327 = pd.read_csv(path + 'corona_tweets_0327.csv')
tweet_volume.append(tweet0327.shape[0])
tweet0328 = pd.read_csv(path + 'corona_tweets_0328.csv')
tweet_volume.append(tweet0328.shape[0])
tweet0330 = pd.read_csv(path + 'corona_tweets_0330.csv')
tweet_volume.append(tweet0330.shape[0])
tweet0331 = pd.read_csv(path + 'corona_tweets_0331.csv')
tweet_volume.append(tweet0331.shape[0])
tweet0401 = pd.read_csv(path + 'corona_tweets_0401.csv')
tweet_volume.append(tweet0401.shape[0])
tweet0402 = pd.read_csv(path + 'corona_tweets_0402.csv')
tweet_volume.append(tweet0402.shape[0])
tweet0403 = pd.read_csv(path + 'corona_tweets_0403.csv')
tweet_volume.append(tweet0403.shape[0])
tweet0404 = pd.read_csv(path + 'corona_tweets_0404.csv')
```

```
'''
The trend of the volume of COVID-19-related tweets (first wave)
'''
labels = ["20/3", "21/3", "22/3", "23/3", "24/3", "25/3", "26/3", "27/3", "28/3", "30/3", "31/3",
          "01/4", "02/4", "03/4", "04/4", "05/4", "06/4", "07/4", "08/4", "09/4", "10/4", "11/4",
          "12/4", "13/4", "14/4", "15/4", "16/4", "17/4", "18/4", "19/4", "20/4", "21/4", "22/4",
          "23/4", "24/4", "25/4", "26/4", "27/4", "28/4", "29/4", "30/4", "01/5", "02/5", "03/5",
          "04/5", "05/5", "06/5", "07/5", "08/5", "09/5", "10/5", "11/5", "12/5", "13/5", "14/5",
          "15/5", "16/5", "17/5", "18/5", "19/5", "20/5", "21/5", "22/5", "23/5", "24/5", "25/5",
          "26/5", "27/5", "28/5", "29/5", "30/5", "31/5", "01/6", "02/6", "03/6", "04/6", "05/6",
          "06/6", "07/6", "08/6", "09/6", "10/6", "11/6", "12/6", "13/6", "14/6", "15/6", "16/6",
          "17/6", "18/6", "19/6", "20/6", "21/6", "22/6", "23/6", "24/6", "25/6", "26/6", "27/6",
          "28/6", "29/6", "30/6"]

plt.bar(labels, tweet_volume, align="center", alpha=1, color="blue")
plt.ylabel('Volume')
plt.xlabel("Date")
plt.title('First wave of COVID19')
plt.show()
```

Fig. 12. The tweet volume trend of first outbreak wave.

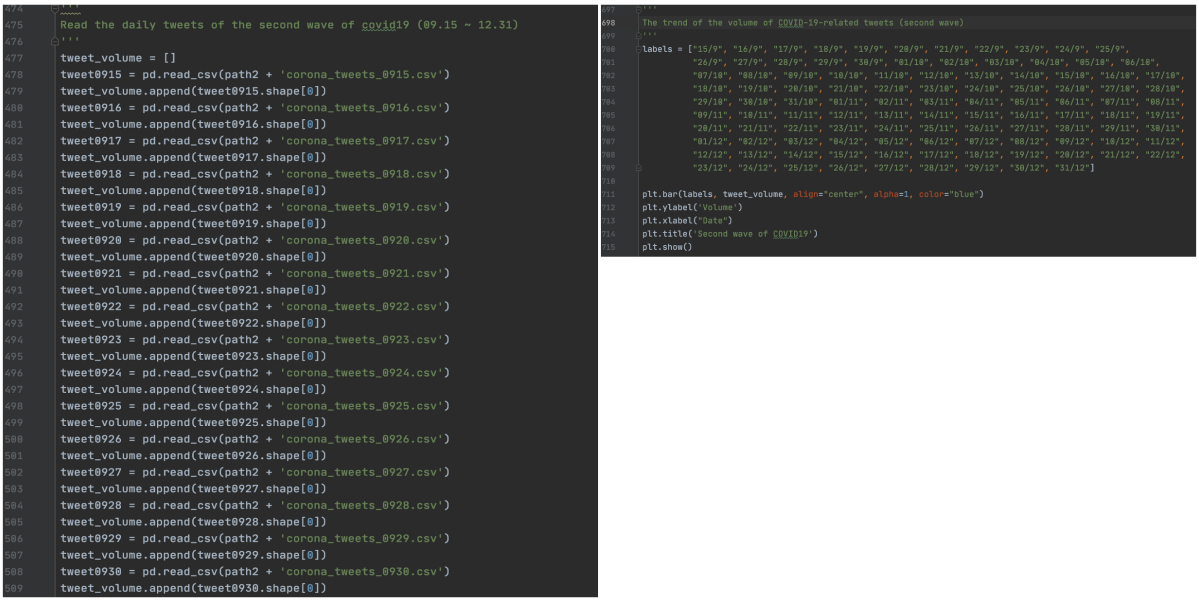


Fig. 13. The tweet volume trend of second outbreak wave.

– **The sentiment analysis of COVID-19 related tweets:**

In this research, TextBlob library was used to conduct the sentiment analysis of tweets, the result of the sentiment analysis will be classified into three polarities including positive, negative and neutral. TextBlob's output for a polarity task is a float within the range [-1.0, 1.0], where -1.0 is a negative polarity and 1.0 is positive. This score can also be equal to 0, which stands for a neutral evaluation. Fig.14 demonstrates the detailed Python code of the sentiment analysis. The public sentiment of policies adopted by the government such as lockdown, wearing face mask, keeping social distance etc, will also be analysed in this research.

```python
def sentiment_analyzer(input_text):
    score = TextBlob(input_text).sentiment.polarity
    return score

first_wave_tweets['sentiment'] = first_wave_tweets["text"].apply(sentiment_analyzer)
print(first_wave_tweets)

positive_count = 0
negative_count = 0
neutral_count = 0
output = open("datasets/twitter_sentiment.txt", "w")

for i in first_wave_tweets["sentiment"]:
    if i < 0:
        negative_count += 1
        output.write("neg")
        output.write('\n')
    if i == 0:
        neutral_count += 1
    else:
        positive_count += 1
        output.write("pos")
        output.write('\n')

x = [positive_count, negative_count, neutral_count]
tot = positive_count + negative_count + neutral_count
positive_count_per = (positive_count / tot) * 100
negative_count_per = (negative_count / tot) * 100
neutral_count_per = (neutral_count / tot) * 100
print(positive_count_per, negative_count_per, neutral_count_per)


labels = "positive", "negative", "neutral"
sizes = [positive_count_per, negative_count_per, neutral_count_per]
explode = (0, 0.1, 0.1)
fig, ax = plt.subplots()
ax.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%', shadow=False, startangle=90)
ax.axis("equal")
plt.show()
```

Fig. 14. The Python code of sentiment analysis.

– **The developing trend of public sentiment:**

Having a good insight of the developing trend of public sentiment towards COVID-19 related events could help the government improve their policies and emergency measures. Thus, the analysis of the developing trend of public sentiment towards the pandemic will be taken into account in our research. The Python code is showed in Fig.15.

```
'''
Trend of public sentiment towards COVID-19
'''
def tweet_sentiment(tweets):
    output = open("datasets/twitter_sentiment.txt", "w")

    for tweet in tweets["text"]:
        sentiment_value, confidence = s.sentiment(tweet)
        print(tweet, sentiment_value, confidence)

        if confidence * 100 >= 80:
            output.write(sentiment_value)
            output.write('\n')

    output.close()
    return True

tweet_sentiment(first_wave_tweets)


fig = plt.figure()
ax1 = fig.add_subplot(1, 1, 1)

def animate(i):
    pullData = open("datasets/twitter_sentiment.txt").read()
    lines = pullData.split("\n")

    xar = []
    yar = []

    x = 0
    y = 0

    for l in lines:
        x += 1
        if "pos" in l:
            y += 1
        elif "neg" in l:
            y -= 1

        xar.append(x)
        yar.append(y)

    ax1.clear()
    ax1.plot(xar, yar)

ani = animation.FuncAnimation(fig, animate, interval=1000)
plt.show()

```

Fig. 15. The Python code of the developing trend analysis of public sentiment towards COVID-19.

– **The most frequent words mentioned by the public:**

Through identifying the most frequent words and phrased mentioned in the tweets could help to find out the hot topics regarding the epidemic, which could be regarded as the evidences of the healthcare concerns of the public. In this research, 100 most frequent words will be identified to help to extract the healthcare concerns of the public. The library of WordCloud will be utilised to conduct the process. Fig.16 illustrates the Python code of the operation.

```
'''
Word cloud of 100 most frequent words
'''
stopwords = set(stopwords.words('english'))
stopwords.update(["40", "jan", "ppl", "amp", "may"])

cv = CountVectorizer(stop_words = stopwords)
words = cv.fit_transform(first_wave_tweets["text"])

sum_words = words.sum(axis=0)

words_freq = [(word, sum_words[0, i]) for word, i in cv.vocabulary_.items()]
words_freq = sorted(words_freq, key = lambda x: x[1], reverse = True)

frequency = pd.DataFrame(words_freq, columns=['word', 'freq'])

frequency.head(50).plot(x='word', y='freq', kind='bar', figsize=(15, 7), color = 'blue')
plt.title("Most Frequently Occuring Words - Top 50")
plt.show()


wordcloud = WordCloud(max_words=50, width=1500, height=1250,
                      background_color="black").generate_from_frequencies(dict(words_freq))

# Display the generated image:
plt.figure(1, figsize=(12, 10))
plt.imshow(wordcloud, interpolation='bilinear')
# plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

Fig. 16. The Python code of word cloud.

3.3.4 Data visualisation

Data visualisation is an interdisciplinary domain, which using graphics to represent the relationship and pattern of data. Data visualisation utilises plots, statistic graphics, information graphics and other tools to deal with the data efficiently and clearly. Numerical data could be encoded with lines, dots, bars etc. to visually reveal the uncovered information in a quantitative research (Few, 2004). Effective data visualisation methods could help researchers analyse and reason about evidence and data. Moreover, it could make numerous and complex data more understandable, accessible and usable. In this research, the library of **Matplotlib** and **Seaborn** will be used to visualise the well-analysed data, in which the data is much easier for people without any background to understand and analyse.

4 Findings and Results

This section presents the findings and results of this research based on the quantitative data analysis method. The detailed visualised graphics and the deep discussion based on the graphics will be comprehensively talked about in the following subsections.

4.1 Global situation

Fig.17 shows the current global situation of the pandemic. Since the first confirmed case was reported, there have been more than 167 million confirmed cases and over 3.4 million deaths caused by the epidemic. From Fig.17 we could find that there are two outbreak waves of COVID-19 in 2020. The first wave is from mid March to the end of June, the daily confirmed cases raised from 158 thousand to 1.3 million and the daily deaths raised from 7 thousand to 32 thousand. The second wave is from mid September to the end of December, the daily confirmed cases raised from 2.1 million to 4.5 million and the daily deaths raised from 36 thousand to 80 thousand. In the following subsections, we will try to extract the sentiment and healthcare concerns of the public, moreover, the comparison between the two outbreak waves of COVID-19 will be conducted which is not comprehensively analysed and studied in the existing works.

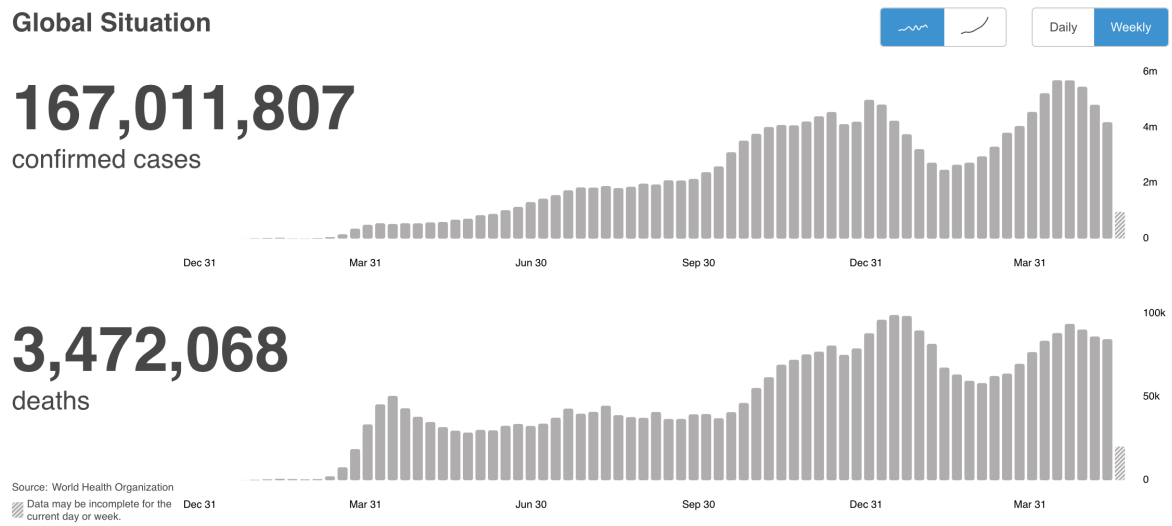


Fig. 17. The global situation of COVID-19 (*figures from WHO Coronavirus (COVID-19) Dashboard - https://covid19.who.int/*).

4.2 The volume trend of COVID-19 related tweets

Fig.18 and Fig.19 show the volume trend of COVID-19 related tweets during the first and second wave of the pandemic. It could be seen that there are obvious differences between the two graphs. As stated in Fig.18, the volume of tweets posted by users has a stepped ascent during the first wave of COVID-19. At the beginning of the first wave, people posted the least volume of tweets, the average daily volume of tweets is about 50 thousand. In the middle of first wave, the average volume of the tweets posted by the users tripled to 150 thousand compared to the beginning of the first wave. At the end of the first wave, the average volume has raised five times to 250 thousand compared to the beginning stage of the first wave.

As reported by Fig.19, the tweet volume trend of the second wave is much more steady than the first wave. the highest volume of tweet is more than 350 thousand posted at 12 October, and the lowest volume of tweet is about 120 thousand posted at 24 December. The average volume of tweets posted during the second wave is around 250 thousand which is equal to the third stage of the first wave.

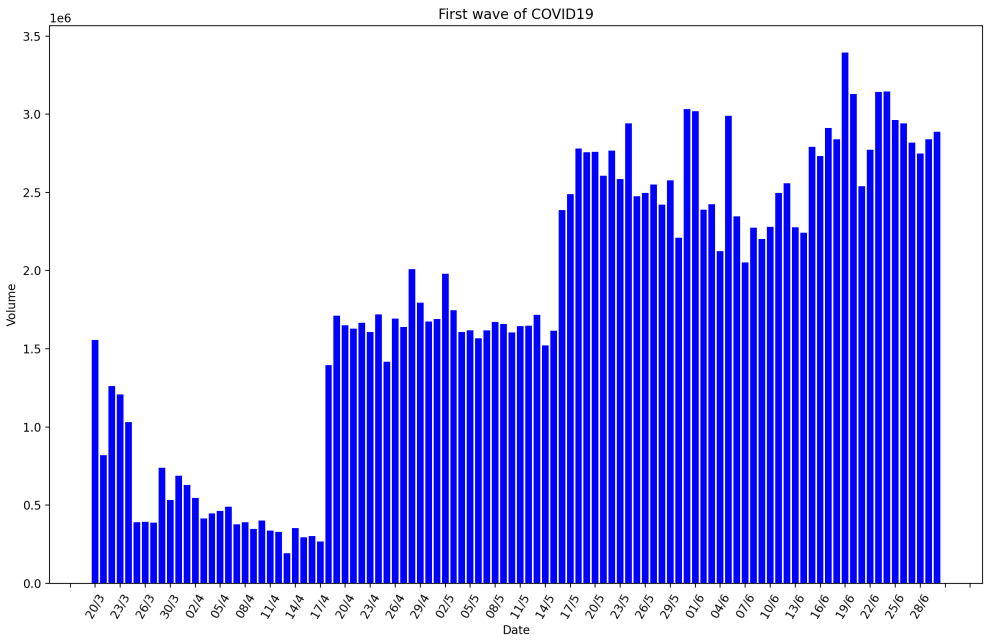


Fig. 18. The volume trend of COVID-19 related tweets during the first wave.

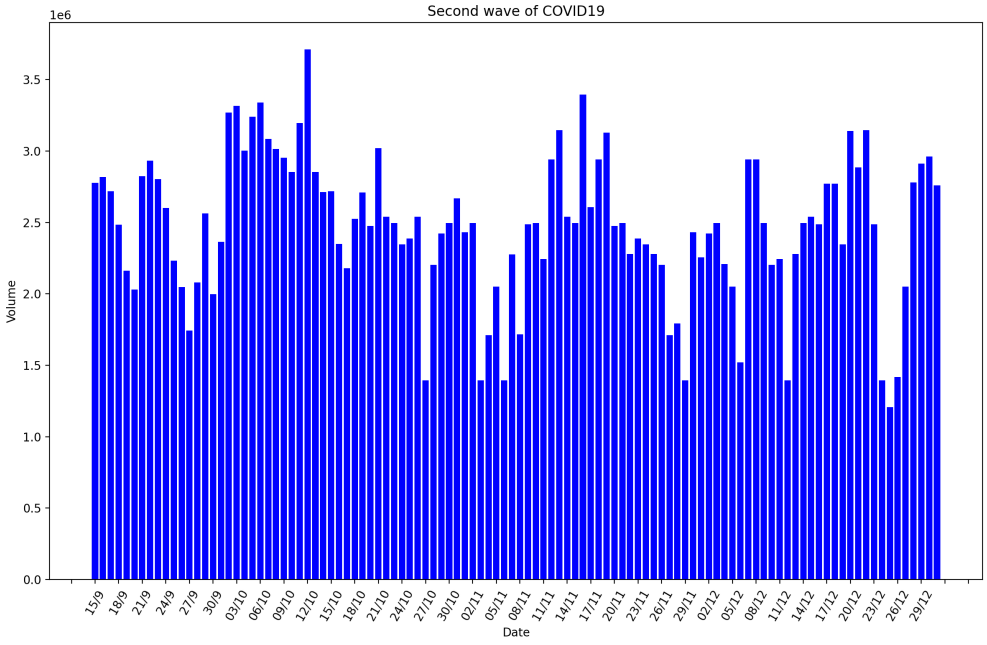


Fig. 19. The volume trend of COVID-19 related tweets during the second wave.

4.3 Sentiment polarity of COVID-19 related tweets

Fig.20 illustrates the sentiment percentage of the COVID-19 related tweets during the first and second waves of the epidemic. During the first wave of COVID-19, almost a half of people (46.6%) held positive sentiment, over one third of them (33.4%) held neutral sentiment, and one fifth of them (20.0%) held negative sentiment. In contrast, During the second wave of COVID-19, more than a half of people (52.3%) held positive sentiment, almost one third of them (30.4%) held neutral sentiment, and almost one fifth of them (17.3%) held negative sentiment. Comparing the two waves of COVID-19, the percentage of people held positive sentiment in the second wave is higher than the one in the first wave, and the percentage of people held neutral and negative sentiments in the second wave are a little lower than the ones in the first wave.

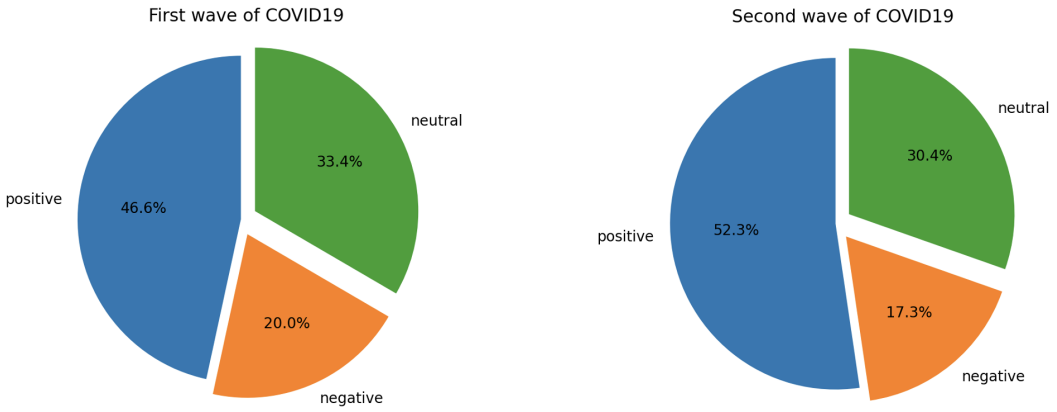


Fig. 20. The sentiment polarity of COVID-19 related tweets during the first and second waves.

4.4 Sentiment polarity of Lockdown rule

Fig.21 demonstrates the public sentiment polarity towards Lockdown rules during the first and second waves of the COVID-19. During the first wave of the outbreak, over a half of the public (55.0%) were positive about the Lockdown rules. However, the number of people supporting the Lockdown policies in the second wave has dropped slightly to 50.6%, as well as the number of public held negative sentiment towards the Lockdown rules, dropped from 19.8% to 17.6%. In contrast, the number of people held neutral sentiment of the Lockdown rules raised from 25.2% to 31.8%.

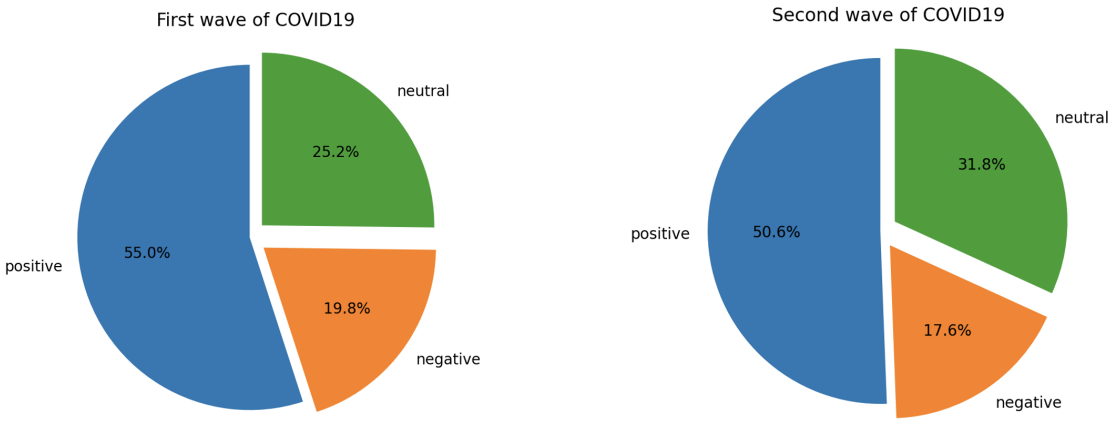


Fig. 21. The sentiment polarity towards Lockdown rules during the first and second waves.

4.5 Sentiment polarity of wearing face mask

Fig.22 presents the public sentiment polarity towards wearing face mask in public areas during the first and second waves of the pandemic. The percentage of population held the three different sentiments are almost similar between the two waves of COVID-19. Over a half of the public thought that wearing face mask in public areas has positive effect to prevent the spread of the epidemic. By contrast, near one fifth of the public did not think wearing face mask in public areas works effectively. And there are another 30% of the population held neutral sentiment towards the prevention strategy.

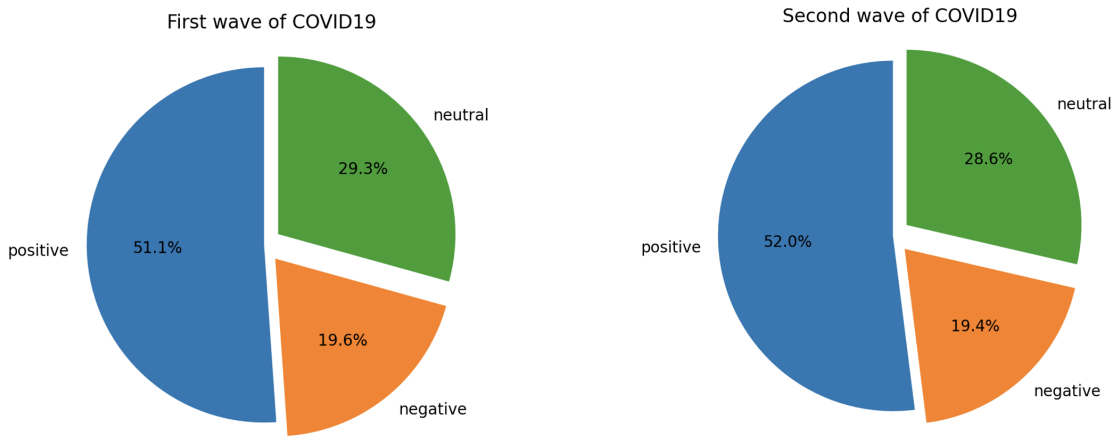


Fig. 22. The sentiment polarity towards wearing face mask during the first and second waves.

4.6 Sentiment polarity of keeping social distance

Fig.23 describes the public sentiment polarity towards keeping social distance in public areas during the first and second waves of the pandemic. The number of the population held the three different sentiments are particularly comparable between the two outbreak waves of the COVID-19. More than a half of the public the method of keeping social distance in public areas could prevent the dissemination of the infectious disease. Around 30% of the population held neutral sentiment towards the prevention method. And almost 20% of the public held negative sentiment to the strategy.

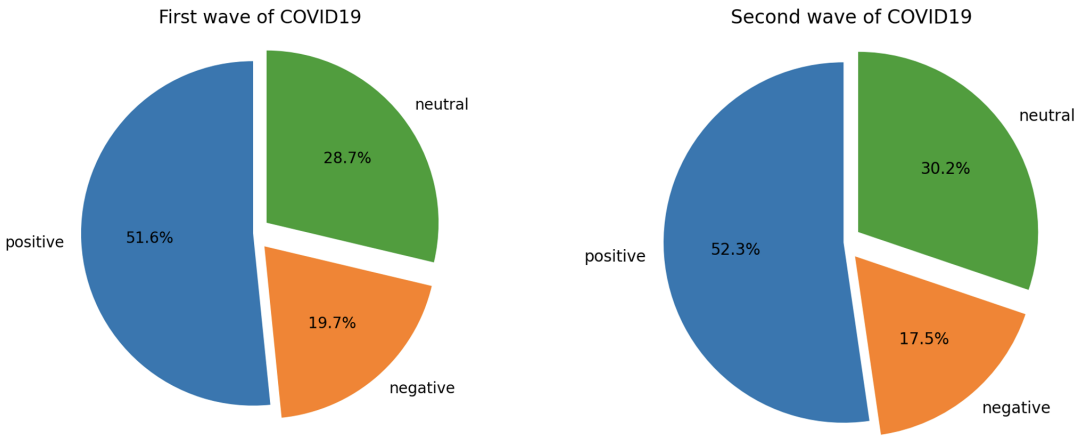


Fig. 23. The sentiment polarity towards keeping social distance during the first and second waves.

4.7 Top 50 most frequently mentioned words

Fig.24 and Fig.25 presents the top 50 most frequently mentioned words by the public during the first and second waves of COVID-19. It could be seen that there are several same words most frequently mentioned by people within the two waves are related to the pandemic including **'covid19', 'pandemic', 'covid', 'coronavirus', 'mask', 'lockdown', 'health', 'deaths'**. The differences between the two waves are that people also mentioned the words in terms of **'testing', 'government', 'businesses'** in the first wave, and in the second wave, people also mentioned **'quarantine', 'vaccine', 'food', 'work'**.

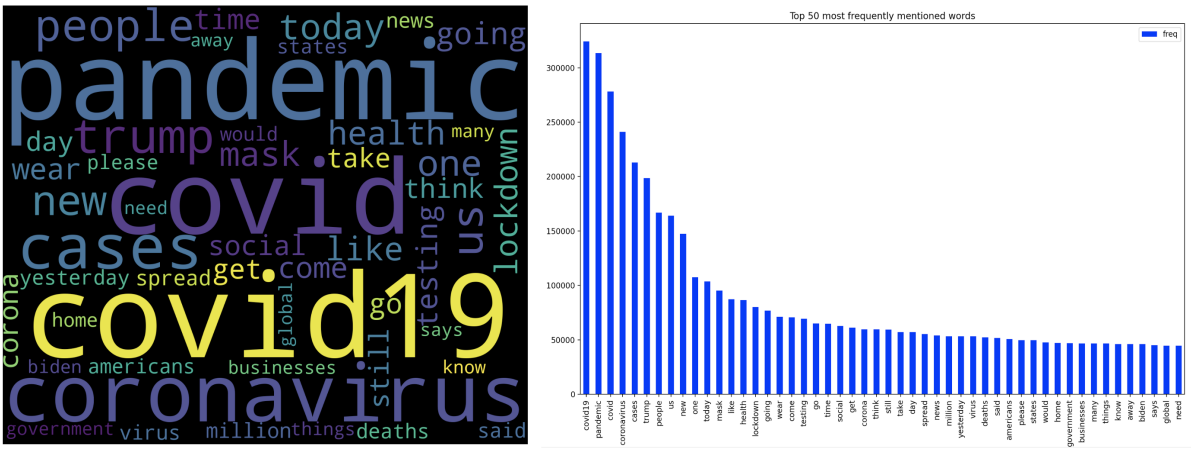


Fig. 24. The Top 50 frequently mentioned words during the first wave of COVID-19.

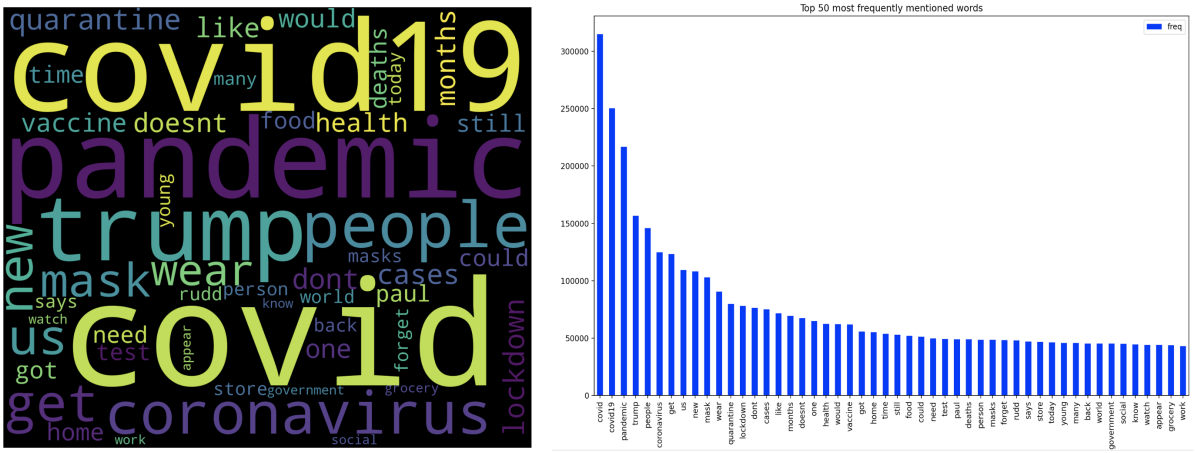


Fig. 25. The Top 50 frequently mentioned words during the second wave of COVID-19.

4.8 Public sentiment of geo-location

Fig.26 shows the overall public sentiment of different countries around the whole world, Fig.27 presents the public sentiment of New Zealand and three metropolises. It could be seen from Fig.26 that the majority of the tweets are originating from North America, Indian, and the Europe. What is more, the public sentiment of New Zealand and its three metropolises including Auckland, Wellington and Christchurch is also analyse in this research. The detailed situation could be seen from Fig.27.

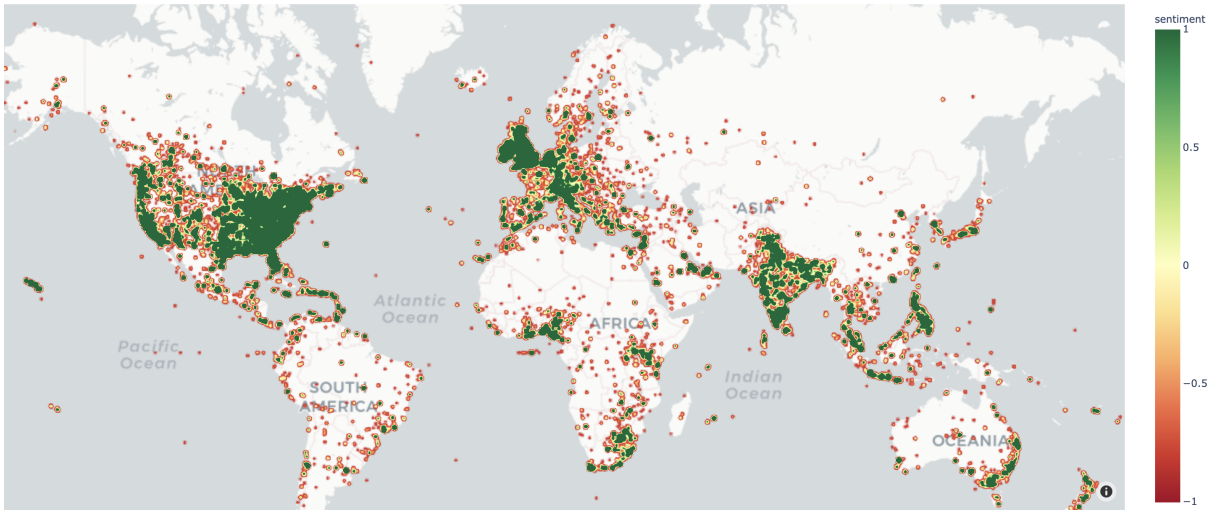


Fig. 26. The public sentiment of geo-location.

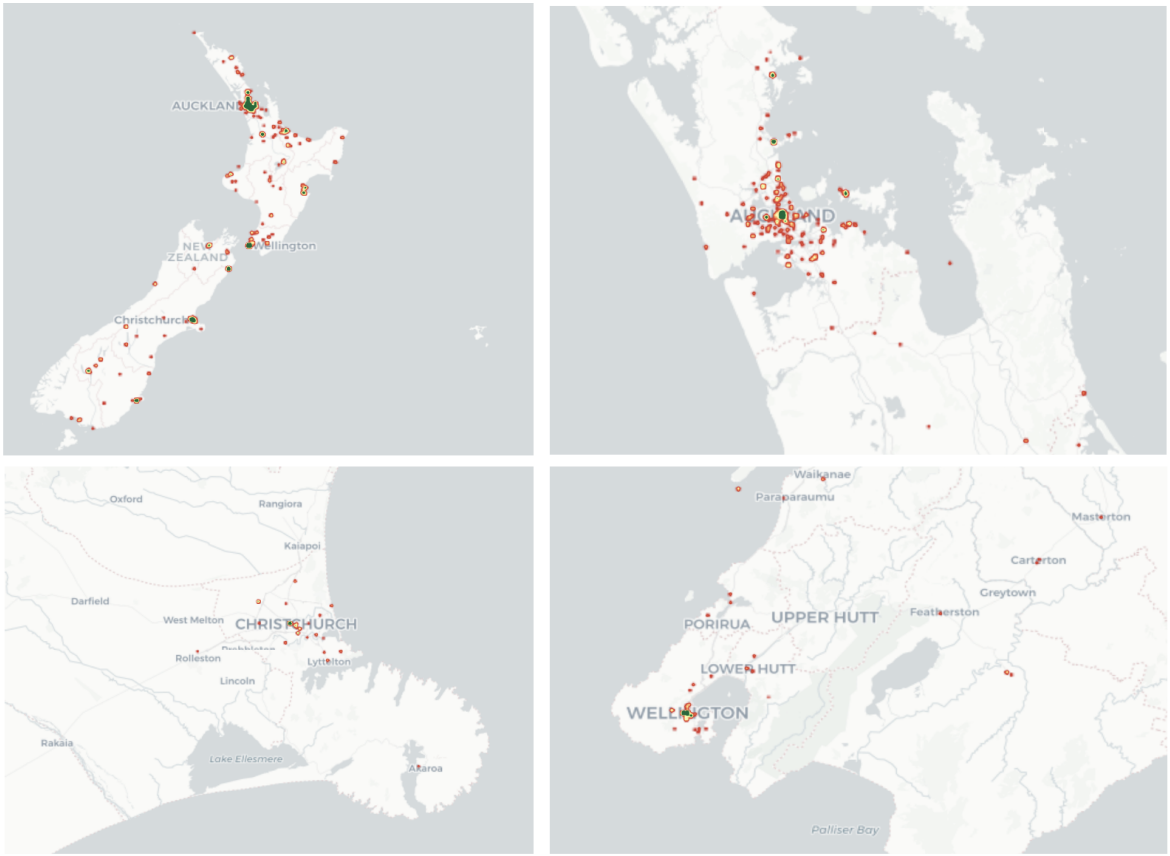


Fig. 27. The public sentiment of New Zealand and metropolises.

5 Discussion and Conclusion

In this section, we will comprehensively talk about the findings of this research and its value to the society and healthcare industry. Meanwhile, the comparison and contrast between the findings in this research and the ones in literature review section will be presented in the following subsections. What is more, we will analyse that if the findings of this research could adequately answer the research questions asked in previous section. Eventually, the limitations and future works will be illustrated at the end.

5.1 Answer research questions

At the beginning of this research, we asked two research questions presented below:

- **RQ1:**

  Can we identify the general public healthcare concerns related to COVID-19 pandemic through mining social media – the case of Twitter?

- **RQ2:**

  What are differences of public sentiment between the first and second waves of the COVID-19 pandemic?

As regards the first research question, the answer is yes. The general public healthcare concerns could be identified and extracted from the COVID-19 related tweets through text mining techniques. For instance, from Fig.18 and Fig.19, we could find that the volume variation of the tweets users posted during the two outbreaks of the COVID-19. By analyzing the changes in the number of tweets sent by users, the government and organisations could have a better insight of the user's concern degree to the COVID-19 crisis. From Fig.20, we could identify the public sentiment of the COVID-19. Fig.21, Fig.22 and Fig.23 tell us the public sentiment of the policies conducted by the government such as Lockdown rules, wearing face mask and keeping social distance in public areas. The result could insist the government to review and improve the policies and strategies. We could identify hot topics and concerns people are talking about during the outbreak of the pandemic from Fig.24 and Fig.25. During the two wave of COVID-19, people pay more attention on these concerns in terms of **'mask', 'lockdown', 'health', 'deaths', 'businesses', 'quarantine', 'vaccine', 'food', 'work'**. Fig.26 and Fig.27 provide a bird-eye view of geo-location related sentiment overview, which could help the authorities to create fist-hand sketches of tentative locations to launch response and reactions to the COVID-19.

With respect to the second research question, the answer is obvious from the visualised graphics. From Fig.18 and Fig.19, we could see that the concern level of the public towards the epidemic has been raising gradually in the firs wave. However, in the second wave, people has been keeping high level

concerns towards the crisis. From Fig.20, Fig.21, Fig.22 and Fig.23, we could also find the variation of the public sentiment regarding the pandemic and policies conducted by the government. The public concerns also have some changes during the two wave of COVID-19. In the first wave, people paid more attention on **'testing', 'government', "businesses**, in contrast, the public cared more about **'quarantine', 'vaccine', 'food', 'work'**. Through analysing the differences and variation between the first and second wave of COVID-19, the authorities could adjust and improve their policies and strategies.

5.2 Our findings vs literature review

In previous literature review, the technique of sentiment analysis was used in several domains such as customer service, stock market and politics. Bagheri et al. proposed a novel unsupervised and domain-independent model, which could identify implicit and explicit aspects in reviews for sentiment analysis. Ren et al. presented a machine learning method based on support vector machine to analyse the investor-generated textual content on the Internet and they achieved the accuracy of forecasting the movement direction of the SSE 50 Index can be as high as 89.93%, and their findings could assist investors in making wiser decisions. Anjaria and Guddeti used five supervised machine learning classifiers to classify the sentiment dataset of social media platform for 2012 US presidential election and 2013 Karnataka state assembly election. Their research revealed that Support Vector Machine achieved the highest accurate score within the five machine learning classifiers for both of the elections.

Moreover, the technique of tweets analysis was applied in a variety of fields including emergency response, public opinion extraction and COVID-19 crisis. Cheong and Lee proposed a novel framework, which applied Twitter microblogging service as a multifaceted data source to conduct demographic analysis and sentimental data in public response to terrorism activities. The outcomes of their research could help the law enforcement agencies and homeland security authorities to make quicker and proper response to terror threats. O'Connor et al. used NLP techniques to analyze some surveys regarding public political opinion and confidence over the 2008 t0 2009 period, and they found some correlations between polls related tweets and sentiment word frequencies. Their results showed the correlation as high as 80%, and captured critical large-scale trends as well. As regards their outcomes, it reveals the possibility of social media platform such as Twitter as a supplement or substitute for traditional polling activities. With regard to COVID-19 crisis, Sharma et al. proposed an in-depth model on shifting topics with regard to trends and sentiments to identify the false and fake information about COVID-19 spreading on Twitter. Ghafarian and Yazdi presented a novel model to identify informative tweets by utilising distributional assumptions. In their model, every single tweet is considered as a "distribution", and significant and meaningful outcomes were achieved in detecting informative tweets towards a crisis event.

In this research, we identified two main findings. The first one is general public healthcare concerns could be identified and extracted from analysing data of the social media platform such as Twitter. The state-of-the-art NLP and ML techniques could be applied in the research to achieve this purpose. What is the most valuable and meaningful of the outcomes is that the findings and results could help the authorities to have a better ans accurate insight of a ongoing crisis, which could help them to adjust and improve their policies and strategies. Moreover, the findings and outcomes are also significant guidable evidences for future out-breaking crises.

The second finding of this research is that the cloud services such as Amazon AWS could make up the shortcomings of the traditional ML and NLP techniques. For example, the traditional ML and NLP techniques are sensitive to the structure of the dataset, these traditional techniques highly rely on structured dataset, nevertheless, almost 80% of data in the real world is unstructured, Amazon AWS could resolve this issue by using Data Lake, Data Lake is a repository or system that stores data in its original format. It stores the data as it is, without the need to structure the data in advance. The Data Lake support a variety of data types including tables, CSV, XML, Json, Logs, emails, PDF, graphics, video, audio etc. Another shortage of traditional techniques is scalability. As we know large volume data analysing highly relies on the performance of the computer including memory, CPU, GPU, hard disk etc. A lot of money and time should be invested to upgrade the performance of the computer. However, the problem could be easily addressed in cloud scenario, users could establish their own virtual computer via cloud services depending on their actual needs. The users could manually choose the hardware and software they will used in the task, such as memory,CPU, GPU, operating system, APIs and Networking features. What is more, Cloud services give them a chance to conduct their requirements. For instance, the AWS charge the fees based on the actual usage of the task. This service is known as a "Pay-as-you-go" model.

5.3 Limitation and future works

In this research, we collected the tweets posted by the users during the two outbreak waves of COVID-19 and tried to extract public healthcare concerns from the dataset. Due to the limited time and resources, there are several limitations within our research, and these limitations will be the next objectives of our future research. The first limitation is language of the tweets, the language of tweets analysed in this research is only English, and other languages are not taken into account. For that reason, the findings and outcomes may not be reasonable enough to represent all the users' sentiment of the pandemic. In the future works, we will analyse other languages as well, in order that the findings and results could be more representative and reasonable for all users of Twitter. Another future work will be taken into account is emojis analysis. Emojis are a better way to express the sentiment and attitude of the users towards

an event, people prefer to use emojis to present their opinions in some circumstances. Notwithstanding, emojis are not easy to be identified by the existing techniques, thus, emojis analysis would be the future objective of the next research.

References

Aisopos, F., Papadakis, G., & Varvarigou, T. (2011). Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd acm sigmm international workshop on social media* (pp. 9–14).

Alhajji, M., Al Khalifah, A., Aljubran, M., & Alkhalifah, M. (2020). Sentiment analysis of tweets in saudi arabia regarding governmental preventive measures to contain covid-19.

Anjaria, M., & Guddeti, R. M. R. (2014). A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, *4*(1), 181.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, pp. 2200–2204).

Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, *52*, 201–213.

Bhat, A. (2019). *Quantitative research: definition, methods, types and examples.* Retrieved from QuestionPro: https://www. questionpro. com/blog ....

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis. *arXiv preprint arXiv:2004.02566*.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, *10*(1), 1–14.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners.* sage.

Burel, G., & Alani, H. (2018). Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media.

Cambria, E., Havasi, C., & Hussain, A. (2012). Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Twenty-fifth international flairs conference.*

Candelieri, A., & Archetti, F. (2015). Detecting events and sentiment on twitter for improving urban mobility. In *Essem@ aamas* (pp. 106–115).

Carley, K. M., Malik, M., Landwehr, P. M., Pfeffer, J., & Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. *Safety science*, *90*, 48–61.

Castillo, C. (2016). *Big crisis data: social media in disasters and time-critical situations.* Cambridge University Press.

Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the united states and italy. *Social Science Computer Review*, *33*(1), 3–20.

Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2019). Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*.

Chatfield, A. T., Scholl, H. J. J., & Brajawidagda, U. (2013). Tsunami early warnings via twitter in government: Net-savvy citizens' co-production of time-critical public information services. *Government information quarterly*, *30*(4), 377–386.

Chen, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, *50*(4), 755–768.

Cheong, M., & Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, *13*(1), 45–59.

Chiu, C., Chiu, N.-H., Sung, R.-J., & Hsieh, P.-Y. (2015). Opinion mining of hotel customer-generated contents in chinese weblogs. *Current issues in tourism*, *18*(5), 477–495.

Cliche, M. (2017). Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dahal, B., Kumar, S. A., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, *9*(1), 1–20.

D'Andrea, E., Ducange, P., Bechini, A., Renda, A., & Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, *116*, 209–226.

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis. *IEEE transactions on intelligent transportation systems*, *16*(4), 2269–2283.

Dey, K., Shrivastava, R., & Kaushik, S. (2018). Topical stance detection for twitter: A two-phase lstm model using attention. In *European conference on information retrieval* (pp. 529–536).

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters*, *81*(2), 246–251.

Few, S. (2004). Eenie, meenie, minie, moe: selecting the right graph for your message. *Intelligent Enterprise*, *7*, 14–35.

Ghafarian, S. H., & Yazdi, H. S. (2020). Identifying crisis-related informative tweets using learning on distributions. *Information Processing & Management*, *57*(2), 102145.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., ... Siegel, E. (2020). Rapid

ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv preprint arXiv:2003.05037*.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, *26*(2), 91–108.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177).

Hu, Y.-H., Wu, F., Lo, C.-L., & Tai, C.-T. (2012). Predicting warfarin dosage from clinical data: a supervised learning approach. *Artificial intelligence in medicine*, *56*(1), 27–34.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1–38.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd international conference on world wide web* (pp. 159–162).

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). *Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions.* Elsevier.

Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on twitter. *Neurocomputing*, *315*, 496–511.

Jahanbin, K., Rahmanian, V., et al. (2020). Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, *13*(8), 378.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, *1*(2), 112–133.

Kalyanam, J., Quezada, M., Poblete, B., & Lanckriet, G. (2016). Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one*, *11*(12), e0166694.

Khan, S., Nabi, G., Han, G., Siddique, R., Lian, S., Shi, H., ... Shereen, M. A. (2020). Novel coronavirus: how things are in wuhan. *Clinical Microbiology and Infection*, *26*(4), 399.

Khedr, A. E., Yaseen, N., et al. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22.

Lamsal, R. (2020a). Coronavirus (covid-19) tweets dataset. *IEEE Dataport*, *10*.

Lamsal, R. (2020b). *Coronavirus (covid-19) tweets dataset.* Retrieved from https://doi.org/10.21227/

781w-ef42

Lamsal, R. (2020c). Design and analysis of a large-scale covid-19 tweets dataset. *Applied Intelligence*, 1–15.

Landwehr, P. M., Wei, W., Kowalchuck, M., & Carley, K. M. (2016). Using tweets to support disaster planning, warning and response. *Safety science*, *90*, 33–47.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature 521 (7553), 436-444. *Google Scholar Google Scholar Cross Ref Cross Ref*.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., ... others (2020). Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*.

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., ... Wang, F.-Y. (2020). Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems*, *7*(2), 556–562.

Li, Y.-M., & Li, T.-Y. (2013). Deriving market intelligence from microblogs. *Decision Support Systems*, *55*(1), 206–217.

Maes, M., Twisk, F. N., & Johnson, C. (2012). Myalgic encephalomyelitis (me), chronic fatigue syndrome (cfs), and chronic fatigue (cf) are distinguished accurately: results of supervised learning techniques applied on clinical and inflammatory data. *Psychiatry research*, *200*(2-3), 754–760.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093–1113.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mizumoto, K., Yanagimoto, H., & Yoshioka, M. (2012). Sentiment analysis of stock market news with semi-supervised learning. In *2012 ieee/acis 11th international conference on computer and information science* (pp. 325–328).

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, *17*(3), 1–23.

Mohsan, F., Nawaz, M. M., Khan, M. S., Shaukat, Z., & Aslam, N. (2011). Impact of customer satisfaction on customer loyalty and intentions to switch: Evidence from banking sector of pakistan. *International journal of business and social science*, *2*(16).

Muijs, D. (2010). *Doing quantitative research in education with spss*. Sage.

Naudé, W. (2020). Artificial intelligence against covid-19: An early review.

Nguyen, D., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the international aaai conference on web and social media* (Vol. 11).

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement

prediction. *Expert Systems with Applications*, *42*(24), 9603–9611.

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international aaai conference on web and social media* (Vol. 4).

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research.

Olariu, A. (2014). Efficient online summarization of microblogging streams. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics, volume 2: Short papers* (pp. 236–240).

Organization, W. H., et al. (2020). Coronavirus disease (covid-19): situation report, 200.

Oxholm, T., Rivera, C., Schirrman, K., & Hoverd, W. J. (2021). New zealand religious community responses to covid-19 while under level 4 lockdown. *Journal of religion and health*, *60*(1), 16–33.

Paddeu, D., Fancello, G., & Fadda, P. (2017). An experimental customer satisfaction index to evaluate the performance of city logistics services. *Transport*, *32*(3), 262–271.

Patro, S. P., Padhy, N., & Chiranjevi, D. (2020). Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. *Evolutionary Intelligence*, 1–29.

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of# conversation is twitter? mining# psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, *29*(6), 2438–2447.

Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), 760–770.

Rosenthal, S., Farra, N., & Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 502–518).

Rotmans, J., Kemp, R., & Van Asselt, M. (2001). More evolution than revolution: transition management in public policy. *foresight*.

Rudra, K., Goyal, P., Ganguly, N., Imran, M., & Mitra, P. (2019). Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. *IEEE Transactions on Computational Social Systems*, *6*(5), 981–993.

Samuel, J., Ali, G., Rahman, M., Esawi, E., Samuel, Y., et al. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314.

Shannon, S., & Kent, N. (2020). charts on internet use around the world as countries grapple with covid-19 internet. *Pew Research Center*.

Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., & Liu, Y. (2020). Coronavirus on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*.

Shou, L., Wang, Z., Chen, K., & Chen, G. (2013). Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 533–542).

Silver, L., Huang, C., & Taylor, K. (2019). In emerging economies, smartphone and social media users have broader social networks. *Pew Research Center*.

Singhal, K., Agrawal, B., & Mittal, N. (2015). Modeling indian general elections: sentiment analysis of political twitter data. In *Information systems design and intelligent applications* (pp. 469–477). Springer.

Tacconelli, E. (2010). Systematic reviews: Crd's guidance for undertaking reviews in health care. *The Lancet Infectious Diseases*, *10*(4), 226.

Takahashi, B., Tandoc Jr, E. C., & Carmichael, C. (2015). Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in human behavior*, *50*, 392–398.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1555–1565).

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Valdivia, A., Luzíón, M. V., & Herrera, F. (2017). Neutrality in the sentiment analysis problem based on fuzzy majority. In *2017 ieee international conference on fuzzy systems (fuzz-ieee)* (pp. 1–6).

Wang, B., & Zhuang, J. (2018). Rumor response, debunking response, and decision makings of misinformed twitter users during disasters. *Natural Hazards*, *93*(3), 1145–1162.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the acl 2012 system demonstrations* (pp. 115–120).

Wang, Z., Shou, L., Chen, K., Chen, G., & Mehrotra, S. (2014). On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, *27*(5), 1301–1315.

Wang, Z., Ye, X., & Tsou, M.-H. (2016). Spatial, temporal, and content analysis of twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523–540.

Worldmeter. (2020). *Covid-19 coronavirus pandemic.* Retrieved from https://www.worldometers.info/coronavirus/

Xiong, S., Lv, H., Zhao, W., & Ji, D. (2018). Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*, *275*, 2459–2466.

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., ... others (2020). Clinical course and outcomes of

critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, *8*(5), 475–481.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, *57*(1), 102107.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Zou, L., Lam, N. S., Cai, H., & Qiang, Y. (2018). Mining twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers*, *108*(5), 1422–1441.

APPENDIX

The Python code of this research is attached below:

```python
import sentiment_mod as s
import matplotlib.pyplot as plt
import matplotlib.animation as animation
import matplotlib.ticker as ticker
from matplotlib import style
import time
import pandas as pd
import glob
from textblob import TextBlob
import re
import string
import nltk
import seaborn as sns
from collections import Counter
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from nltk.corpus import stopwords
import os
import csv
from sklearn.feature_extraction.text import CountVectorizer

'''
Delete the column of sentiment score and hydrate the tweet IDs
'''
path = '/Volumes/Vincent/COVID19 Tweets/Tweets with text/CSV file/first wave of covid19/'
path2 = '/Volumes/Vincent/COVID19 Tweets/Tweets with text/CSV file/second wave of covid19/'

files = []

for file in os.listdir(path):
    if file.endswith(".csv"):
        files.append(path+file)

for file in files:
    dataFrame = pd.read_csv(file, header=None)
    dataFrame = dataFrame[0]
    dataFrame.to_csv(file, index=False, header=None)
```

```
37
38  for file in os.listdir(path2):
39      if file.endswith(".csv"):
40          files.append(path2+file)
41
42  for file in files:
43      dataFrame = pd.read_csv(file, header=None)
44      dataFrame = dataFrame[0]
45      dataFrame.to_csv(file, index=False, header=None)
46
47
48  '''
49  Read the daily tweets of the first wave of covid19 (03.20 ~ 06.30)
50  '''
51  tweet0320_p1 = pd.read_csv(path + 'corona_tweets_0320_p1.csv')
52  tweet0320_p2 = pd.read_csv(path + 'corona_tweets_0320_p2.csv')
53  tweet0320_p3 = pd.read_csv(path + 'corona_tweets_0320_p3.csv')
54  tweet0321 = pd.read_csv(path + 'corona_tweets_0321.csv')
55  tweet0322 = pd.read_csv(path + 'corona_tweets_0322.csv')
56  tweet0323 = pd.read_csv(path + 'corona_tweets_0323.csv')
57  tweet0324 = pd.read_csv(path + 'corona_tweets_0324.csv')
58  tweet0325 = pd.read_csv(path + 'corona_tweets_0325.csv')
59  tweet0326 = pd.read_csv(path + 'corona_tweets_0326.csv')
60  tweet0327 = pd.read_csv(path + 'corona_tweets_0327.csv')
61  tweet0328 = pd.read_csv(path + 'corona_tweets_0328.csv')
62  tweet0330 = pd.read_csv(path + 'corona_tweets_0330.csv')
63  tweet0331 = pd.read_csv(path + 'corona_tweets_0331.csv')
64  tweet0401 = pd.read_csv(path + 'corona_tweets_0401.csv')
65  tweet0402 = pd.read_csv(path + 'corona_tweets_0402.csv')
66  tweet0403 = pd.read_csv(path + 'corona_tweets_0403.csv')
67  tweet0404 = pd.read_csv(path + 'corona_tweets_0404.csv')
68  tweet0405 = pd.read_csv(path + 'corona_tweets_0405.csv')
69  tweet0406 = pd.read_csv(path + 'corona_tweets_0406.csv')
70  tweet0407 = pd.read_csv(path + 'corona_tweets_0407.csv')
71  tweet0408 = pd.read_csv(path + 'corona_tweets_0408.csv')
72  tweet0409 = pd.read_csv(path + 'corona_tweets_0409.csv')
73  tweet0410 = pd.read_csv(path + 'corona_tweets_0410.csv')
```

```
74  tweet0411 = pd.read_csv(path + 'corona_tweets_0411.csv')
75  tweet0412 = pd.read_csv(path + 'corona_tweets_0412.csv')
76  tweet0413 = pd.read_csv(path + 'corona_tweets_0413.csv')
77  tweet0414 = pd.read_csv(path + 'corona_tweets_0414.csv')
78  tweet0415 = pd.read_csv(path + 'corona_tweets_0415.csv')
79  tweet0416 = pd.read_csv(path + 'corona_tweets_0416.csv')
80  tweet0417 = pd.read_csv(path + 'corona_tweets_0417.csv')
81  tweet0418 = pd.read_csv(path + 'corona_tweets_0418.csv')
82  tweet0419 = pd.read_csv(path + 'corona_tweets_0419.csv')
83  tweet0420 = pd.read_csv(path + 'corona_tweets_0420.csv')
84  tweet0421 = pd.read_csv(path + 'corona_tweets_0421.csv')
85  tweet0422 = pd.read_csv(path + 'corona_tweets_0422.csv')
86  tweet0423 = pd.read_csv(path + 'corona_tweets_0423.csv')
87  tweet0424 = pd.read_csv(path + 'corona_tweets_0424.csv')
88  tweet0425 = pd.read_csv(path + 'corona_tweets_0425.csv')
89  tweet0426 = pd.read_csv(path + 'corona_tweets_0426.csv')
90  tweet0427 = pd.read_csv(path + 'corona_tweets_0427.csv')
91  tweet0428 = pd.read_csv(path + 'corona_tweets_0428.csv')
92  tweet0429 = pd.read_csv(path + 'corona_tweets_0429.csv')
93  tweet0430 = pd.read_csv(path + 'corona_tweets_0430.csv')
94  tweet0501 = pd.read_csv(path + 'corona_tweets_0501.csv')
95  tweet0502 = pd.read_csv(path + 'corona_tweets_0502.csv')
96  tweet0503 = pd.read_csv(path + 'corona_tweets_0503.csv')
97  tweet0504 = pd.read_csv(path + 'corona_tweets_0504.csv')
98  tweet0505 = pd.read_csv(path + 'corona_tweets_0505.csv')
99  tweet0506 = pd.read_csv(path + 'corona_tweets_0506.csv')
100 tweet0507 = pd.read_csv(path + 'corona_tweets_0507.csv')
101 tweet0508 = pd.read_csv(path + 'corona_tweets_0508.csv')
102 tweet0509 = pd.read_csv(path + 'corona_tweets_0509.csv')
103 tweet0510 = pd.read_csv(path + 'corona_tweets_0510.csv')
104 tweet0511 = pd.read_csv(path + 'corona_tweets_0511.csv')
105 tweet0512 = pd.read_csv(path + 'corona_tweets_0512.csv')
106 tweet0513 = pd.read_csv(path + 'corona_tweets_0513.csv')
107 tweet0514 = pd.read_csv(path + 'corona_tweets_0514.csv')
108 tweet0515 = pd.read_csv(path + 'corona_tweets_0515.csv')
109 tweet0516 = pd.read_csv(path + 'corona_tweets_0516.csv')
110 tweet0517 = pd.read_csv(path + 'corona_tweets_0517.csv')
```

```
tweet0518 = pd.read_csv(path + 'corona_tweets_0518.csv')
tweet0519 = pd.read_csv(path + 'corona_tweets_0519.csv')
tweet0520 = pd.read_csv(path + 'corona_tweets_0520.csv')
tweet0521 = pd.read_csv(path + 'corona_tweets_0521.csv')
tweet0522 = pd.read_csv(path + 'corona_tweets_0522.csv')
tweet0523 = pd.read_csv(path + 'corona_tweets_0523.csv')
tweet0524 = pd.read_csv(path + 'corona_tweets_0524.csv')
tweet0525 = pd.read_csv(path + 'corona_tweets_0525.csv')
tweet0526 = pd.read_csv(path + 'corona_tweets_0526.csv')
tweet0527 = pd.read_csv(path + 'corona_tweets_0527.csv')
tweet0528 = pd.read_csv(path + 'corona_tweets_0528.csv')
tweet0529 = pd.read_csv(path + 'corona_tweets_0529.csv')
tweet0530 = pd.read_csv(path + 'corona_tweets_0530.csv')
tweet0531 = pd.read_csv(path + 'corona_tweets_0531.csv')
tweet0601 = pd.read_csv(path + 'corona_tweets_0601.csv')
tweet0602 = pd.read_csv(path + 'corona_tweets_0602.csv')
tweet0603 = pd.read_csv(path + 'corona_tweets_0603.csv')
tweet0604 = pd.read_csv(path + 'corona_tweets_0604.csv')
tweet0605 = pd.read_csv(path + 'corona_tweets_0605.csv')
tweet0606 = pd.read_csv(path + 'corona_tweets_0606.csv')
tweet0607 = pd.read_csv(path + 'corona_tweets_0607.csv')
tweet0608 = pd.read_csv(path + 'corona_tweets_0608.csv')
tweet0609 = pd.read_csv(path + 'corona_tweets_0609.csv')
tweet0610 = pd.read_csv(path + 'corona_tweets_0610.csv')
tweet0611 = pd.read_csv(path + 'corona_tweets_0611.csv')
tweet0612 = pd.read_csv(path + 'corona_tweets_0612.csv')
tweet0613 = pd.read_csv(path + 'corona_tweets_0613.csv')
tweet0614 = pd.read_csv(path + 'corona_tweets_0614.csv')
tweet0615 = pd.read_csv(path + 'corona_tweets_0615.csv')
tweet0616 = pd.read_csv(path + 'corona_tweets_0616.csv')
tweet0617 = pd.read_csv(path + 'corona_tweets_0617.csv')
tweet0618 = pd.read_csv(path + 'corona_tweets_0618.csv')
tweet0619 = pd.read_csv(path + 'corona_tweets_0619.csv')
tweet0620 = pd.read_csv(path + 'corona_tweets_0620.csv')
tweet0621 = pd.read_csv(path + 'corona_tweets_0621.csv')
tweet0622 = pd.read_csv(path + 'corona_tweets_0622.csv')
tweet0623 = pd.read_csv(path + 'corona_tweets_0623.csv')
```

```
tweet0624 = pd.read_csv(path + 'corona_tweets_0624.csv')
tweet0625 = pd.read_csv(path + 'corona_tweets_0625.csv')
tweet0626 = pd.read_csv(path + 'corona_tweets_0626.csv')
tweet0627 = pd.read_csv(path + 'corona_tweets_0627.csv')
tweet0628 = pd.read_csv(path + 'corona_tweets_0628.csv')
tweet0629 = pd.read_csv(path + 'corona_tweets_0629.csv')
tweet0630 = pd.read_csv(path + 'corona_tweets_0630.csv')

'''
Put all the tweets together and remove the duplicated tweets
'''
first_wave_tweets = pd.concat([tweet0320_p1, tweet0320_p2, tweet0320_p3, tweet0321, tweet0322, tweet0323, tweet0324,
                               tweet0325, tweet0326, tweet0327, tweet0328, tweet0330, tweet0331, tweet0401, tweet0402,
                               tweet0403, tweet0404, tweet0405, tweet0406, tweet0407, tweet0408, tweet0409, tweet0410,
                               tweet0411, tweet0412, tweet0413, tweet0414, tweet0415, tweet0416, tweet0417, tweet0418,
                               tweet0419, tweet0420, tweet0421, tweet0422, tweet0423, tweet0424, tweet0425, tweet0426,
                               tweet0427, tweet0428, tweet0429, tweet0430, tweet0501, tweet0502, tweet0503, tweet0504,
                               tweet0505, tweet0506, tweet0507, tweet0508, tweet0509, tweet0510, tweet0511, tweet0512,
                               tweet0513, tweet0514, tweet0515, tweet0516, tweet0517, tweet0518, tweet0519, tweet0520,
                               tweet0521, tweet0522, tweet0523, tweet0524, tweet0525, tweet0526, tweet0527, tweet0528,
                               tweet0529, tweet0530, tweet0531, tweet0601, tweet0602, tweet0603, tweet0604, tweet0605,
                               tweet0606, tweet0607, tweet0608, tweet0609, tweet0610, tweet0611, tweet0612, tweet0613,
                               tweet0614, tweet0615, tweet0616, tweet0617, tweet0618, tweet0619, tweet0620, tweet0621,
                               tweet0622, tweet0623, tweet0624, tweet0625, tweet0626, tweet0627, tweet0628, tweet0629,
                               tweet0630], axis=0, ignore_index=True).drop_duplicates()

print(first_wave_tweets.shape)

'''
Choose the English tweets
'''
first_wave_tweets = first_wave_tweets.loc[first_wave_tweets["lang"] == "en"]
print(first_wave_tweets.shape)
# print(first_wave_tweets)
```

```
'''
Reset the index of the tweets
'''
first_wave_tweets = first_wave_tweets.reset_index(drop=True)
# print(first_wave_tweets)

'''
Extract the text of tweets
'''
first_wave_tweets = first_wave_tweets["text"]
# print(first_wave_tweets)

'''
Convert the tweets format to tabular data
'''
first_wave_tweets = pd.DataFrame(first_wave_tweets)
# print(first_wave_tweets)


def cleanTweetsAttribute(text):
    # Removing urls
    text = re.sub(r'http\S+', '', text)  # remove http links
    text = re.sub(r'bit.ly/\S+', '', text)  # remove bitly links
    text = text.strip('[link]')  # remove [links]
    text = text.strip('RT ')  # remove retweet sign 'RT'

    # Removing user mentions
    text = re.sub('(@[A-Za-z]+[A-Za-z0-9-_]+)', '', text)

    # Removing hashtags
    text = re.sub(r"#(\w+)", '', text, flags=re.MULTILINE)

    return text
```

```
def tweetsPreprocessing(text):
    text = cleanTweetsAttribute(text)

    # Converting to lowercase
    text = text.lower()

    # Removing punctuations
    text = text.translate(str.maketrans('', '', string.punctuation))

    return text


def sentiment_analyzer(input_text):
    score = TextBlob(input_text).sentiment.polarity
    return score

first_wave_tweets["text"] = first_wave_tweets["text"].apply(tweetsPreprocessing)
print(first_wave_tweets)

'''
Sentiment analysis of tweets(positive, negative or neutral) by using TextBlob library
TextBlob's output for a polarity task is a float within the range [-1.0, 1.0]
where -1.0 is a negative polarity and 1.0 is positive. This score can also be equal to 0,
which stands for a neutral evaluation
'''

first_wave_tweets['sentiment'] = first_wave_tweets["text"].apply(sentiment_analyzer)
print(first_wave_tweets)

positive_count = 0
negative_count = 0
neutral_count = 0
output = open("datasets/twitter_sentiment.txt", "w")

```

```
for i in first_wave_tweets["sentiment"]:
    if i < 0:
        negative_count += 1
        output.write("neg")
        output.write('\n')
    if i == 0:
        neutral_count += 1
    else:
        positive_count += 1
        output.write("pos")
        output.write('\n')

x = [positive_count, negative_count, neutral_count]
tot = positive_count + negative_count + neutral_count
positive_count_per = round((positive_count / tot) * 100, 1)
negative_count_per = round((negative_count / tot) * 100, 1)
neutral_count_per = 100 - positive_count_per - negative_count_per
print(positive_count_per, negative_count_per, neutral_count_per)

labels = "positive", "negative", "neutral"
sizes = [positive_count_per, negative_count_per, neutral_count_per]
explode = (0, 0.1, 0.1)
fig, ax = plt.subplots()
ax.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%', shadow=False, startangle=90)
ax.axis("equal")
plt.title('First wave of COVID19')
plt.show()
```

```


'''
Trend of public sentiment towards COVID-19
'''
def tweet_sentiment(tweets):
    output = open("datasets/twitter_sentiment.txt", "w")

    for tweet in tweets["text"]:
        sentiment_value, confidence = s.sentiment(tweet)
        print(tweet, sentiment_value, confidence)

        if confidence * 100 >= 80:
            output.write(sentiment_value)
            output.write('\n')

    output.close()
    return True

tweet_sentiment(first_wave_tweets)


fig = plt.figure()
ax1 = fig.add_subplot(1, 1, 1)

def animate(i):
    pullData = open("datasets/twitter_sentiment.txt").read()
    lines = pullData.split("\n")

    xar = []
    yar = []

    x = 0
    y = 0

```

```
    for l in lines:
        x += 1
        if "pos" in l:
            y += 1
        elif "neg" in l:
            y -= 1

        xar.append(x)
        yar.append(y)

    ax1.clear()
    ax1.plot(xar, yar)

ani = animation.FuncAnimation(fig, animate, interval=1000)
plt.show()


'''
Word cloud of 50 most frequent words
'''
stopwords = set(stopwords.words('english'))
stopwords.update(["40", "jan", "ppl", "amp", "may"])

cv = CountVectorizer(stop_words=stopwords)
words = cv.fit_transform(first_wave_tweets["text"])

sum_words = words.sum(axis=0)

words_freq = [(word, sum_words[0, i]) for word, i in cv.vocabulary_.items()]
words_freq = sorted(words_freq, key=lambda x: x[1], reverse=True)

frequency = pd.DataFrame(words_freq, columns=['word', 'freq'])

frequency.head(50).plot(x='word', y='freq', kind='bar', figsize=(15, 7), color='blue')
plt.title("Top 50 most frequently mentioned words")
plt.show()
```

```


wordcloud = WordCloud(max_words=50, width=1500, height=1250,
                      background_color="black").generate_from_frequencies(dict(words_freq))

# Display the generated image:
plt.figure(1, figsize=(12, 10))
plt.imshow(wordcloud, interpolation='bilinear')
# plt.imshow(wordcloud)
plt.axis("off")
plt.show()


'''
The trend of the volume of COVID-19-related tweets (first wave)
'''
labels = ["20/3", "21/3", "22/3", "23/3", "24/3", "25/3", "26/3", "27/3", "28/3", "30/3", "31/3",
          "01/4", "02/4", "03/4", "04/4", "05/4", "06/4", "07/4", "08/4", "09/4", "10/4", "11/4",
          "12/4", "13/4", "14/4", "15/4", "16/4", "17/4", "18/4", "19/4", "20/4", "21/4", "22/4",
          "23/4", "24/4", "25/4", "26/4", "27/4", "28/4", "29/4", "30/4", "01/5", "02/5", "03/5",
          "04/5", "05/5", "06/5", "07/5", "08/5", "09/5", "10/5", "11/5", "12/5", "13/5", "14/5",
          "15/5", "16/5", "17/5", "18/5", "19/5", "20/5", "21/5", "22/5", "23/5", "24/5", "25/5",
          "26/5", "27/5", "28/5", "29/5", "30/5", "31/5", "01/6", "02/6", "03/6", "04/6", "05/6",
          "06/6", "07/6", "08/6", "09/6", "10/6", "11/6", "12/6", "13/6", "14/6", "15/6", "16/6",
          "17/6", "18/6", "19/6", "20/6", "21/6", "22/6", "23/6", "24/6", "25/6", "26/6", "27/6",
          "28/6", "29/6", "30/6"]

fig, ax = plt.subplots(1, 1)
plt.xticks(rotation=60)
tick_spacing = 3
ax.xaxis.set_major_locator(ticker.MultipleLocator(tick_spacing))
plt.bar(labels, tweet_volume, align="center", alpha=1, color="blue")
plt.ylabel('Volume')
plt.xlabel("Date")
plt.title('First wave of COVID19')
plt.show()

```

```
'''
Read the daily tweets of the second wave of covid19 (09.15 ~ 12.31)
'''
tweet0915 = pd.read_csv(path2 + 'corona_tweets_0915.csv')
tweet0916 = pd.read_csv(path2 + 'corona_tweets_0916.csv')
tweet0917 = pd.read_csv(path2 + 'corona_tweets_0917.csv')
tweet0918 = pd.read_csv(path2 + 'corona_tweets_0918.csv')
tweet0919 = pd.read_csv(path2 + 'corona_tweets_0919.csv')
tweet0920 = pd.read_csv(path2 + 'corona_tweets_0920.csv')
tweet0921 = pd.read_csv(path2 + 'corona_tweets_0921.csv')
tweet0922 = pd.read_csv(path2 + 'corona_tweets_0922.csv')
tweet0923 = pd.read_csv(path2 + 'corona_tweets_0923.csv')
tweet0924 = pd.read_csv(path2 + 'corona_tweets_0924.csv')
tweet0925 = pd.read_csv(path2 + 'corona_tweets_0925.csv')
tweet0926 = pd.read_csv(path2 + 'corona_tweets_0926.csv')
tweet0927 = pd.read_csv(path2 + 'corona_tweets_0927.csv')
tweet0928 = pd.read_csv(path2 + 'corona_tweets_0928.csv')
tweet0929 = pd.read_csv(path2 + 'corona_tweets_0929.csv')
tweet0930 = pd.read_csv(path2 + 'corona_tweets_0930.csv')
tweet1001 = pd.read_csv(path2 + 'corona_tweets_1001.csv')
tweet1002 = pd.read_csv(path2 + 'corona_tweets_1002.csv')
tweet1003 = pd.read_csv(path2 + 'corona_tweets_1003.csv')
tweet1004 = pd.read_csv(path2 + 'corona_tweets_1004.csv')
tweet1005 = pd.read_csv(path2 + 'corona_tweets_1005.csv')
tweet1006 = pd.read_csv(path2 + 'corona_tweets_1006.csv')
tweet1007 = pd.read_csv(path2 + 'corona_tweets_1007.csv')
tweet1008 = pd.read_csv(path2 + 'corona_tweets_1008.csv')
tweet1009 = pd.read_csv(path2 + 'corona_tweets_1009.csv')
tweet1010 = pd.read_csv(path2 + 'corona_tweets_1010.csv')
tweet1011 = pd.read_csv(path2 + 'corona_tweets_1011.csv')
tweet1012 = pd.read_csv(path2 + 'corona_tweets_1012.csv')
tweet1013 = pd.read_csv(path2 + 'corona_tweets_1013.csv')
tweet1014 = pd.read_csv(path2 + 'corona_tweets_1014.csv')
tweet1015 = pd.read_csv(path2 + 'corona_tweets_1015.csv')
tweet1016 = pd.read_csv(path2 + 'corona_tweets_1016.csv')
tweet1017 = pd.read_csv(path2 + 'corona_tweets_1017.csv')
```

```
tweet1018 = pd.read_csv(path2 + 'corona_tweets_1018.csv')
tweet1019 = pd.read_csv(path2 + 'corona_tweets_1019.csv')
tweet1020 = pd.read_csv(path2 + 'corona_tweets_1020.csv')
tweet1021 = pd.read_csv(path2 + 'corona_tweets_1021.csv')
tweet1022 = pd.read_csv(path2 + 'corona_tweets_1022.csv')
tweet1023 = pd.read_csv(path2 + 'corona_tweets_1023.csv')
tweet1024 = pd.read_csv(path2 + 'corona_tweets_1024.csv')
tweet1025 = pd.read_csv(path2 + 'corona_tweets_1025.csv')
tweet1026 = pd.read_csv(path2 + 'corona_tweets_1026.csv')
tweet1027 = pd.read_csv(path2 + 'corona_tweets_1027.csv')
tweet1028 = pd.read_csv(path2 + 'corona_tweets_1028.csv')
tweet1029 = pd.read_csv(path2 + 'corona_tweets_1029.csv')
tweet1030 = pd.read_csv(path2 + 'corona_tweets_1030.csv')
tweet1031 = pd.read_csv(path2 + 'corona_tweets_1031.csv')
tweet1101 = pd.read_csv(path2 + 'corona_tweets_1101.csv')
tweet1102 = pd.read_csv(path2 + 'corona_tweets_1102.csv')
tweet1103 = pd.read_csv(path2 + 'corona_tweets_1103.csv')
tweet1104 = pd.read_csv(path2 + 'corona_tweets_1104.csv')
tweet1105 = pd.read_csv(path2 + 'corona_tweets_1105.csv')
tweet1106 = pd.read_csv(path2 + 'corona_tweets_1106.csv')
tweet1107 = pd.read_csv(path2 + 'corona_tweets_1107.csv')
tweet1108 = pd.read_csv(path2 + 'corona_tweets_1108.csv')
tweet1109 = pd.read_csv(path2 + 'corona_tweets_1109.csv')
tweet1110 = pd.read_csv(path2 + 'corona_tweets_1110.csv')
tweet1111 = pd.read_csv(path2 + 'corona_tweets_1111.csv')
tweet1112 = pd.read_csv(path2 + 'corona_tweets_1112.csv')
tweet1113 = pd.read_csv(path2 + 'corona_tweets_1113.csv')
tweet1114 = pd.read_csv(path2 + 'corona_tweets_1114.csv')
tweet1115 = pd.read_csv(path2 + 'corona_tweets_1115.csv')
tweet1116 = pd.read_csv(path2 + 'corona_tweets_1116.csv')
tweet1117 = pd.read_csv(path2 + 'corona_tweets_1117.csv')
tweet1118 = pd.read_csv(path2 + 'corona_tweets_1118.csv')
tweet1119 = pd.read_csv(path2 + 'corona_tweets_1119.csv')
tweet1120 = pd.read_csv(path2 + 'corona_tweets_1120.csv')
tweet1121 = pd.read_csv(path2 + 'corona_tweets_1121.csv')
tweet1122 = pd.read_csv(path2 + 'corona_tweets_1122.csv')
tweet1123 = pd.read_csv(path2 + 'corona_tweets_1123.csv')
```

```
tweet1124 = pd.read_csv(path2 + 'corona_tweets_1124.csv')
tweet1125 = pd.read_csv(path2 + 'corona_tweets_1125.csv')
tweet1126 = pd.read_csv(path2 + 'corona_tweets_1126.csv')
tweet1127 = pd.read_csv(path2 + 'corona_tweets_1127.csv')
tweet1128 = pd.read_csv(path2 + 'corona_tweets_1128.csv')
tweet1129 = pd.read_csv(path2 + 'corona_tweets_1129.csv')
tweet1130 = pd.read_csv(path2 + 'corona_tweets_1130.csv')
tweet1201 = pd.read_csv(path2 + 'corona_tweets_1201.csv')
tweet1202 = pd.read_csv(path2 + 'corona_tweets_1202.csv')
tweet1203 = pd.read_csv(path2 + 'corona_tweets_1203.csv')
tweet1204 = pd.read_csv(path2 + 'corona_tweets_1204.csv')
tweet1205 = pd.read_csv(path2 + 'corona_tweets_1205.csv')
tweet1206 = pd.read_csv(path2 + 'corona_tweets_1206.csv')
tweet1207 = pd.read_csv(path2 + 'corona_tweets_1207.csv')
tweet1208 = pd.read_csv(path2 + 'corona_tweets_1208.csv')
tweet1209 = pd.read_csv(path2 + 'corona_tweets_1209.csv')
tweet1210 = pd.read_csv(path2 + 'corona_tweets_1210.csv')
tweet1211 = pd.read_csv(path2 + 'corona_tweets_1211.csv')
tweet1212 = pd.read_csv(path2 + 'corona_tweets_1212.csv')
tweet1213 = pd.read_csv(path2 + 'corona_tweets_1213.csv')
tweet1214 = pd.read_csv(path2 + 'corona_tweets_1214.csv')
tweet1215 = pd.read_csv(path2 + 'corona_tweets_1215.csv')
tweet1216 = pd.read_csv(path2 + 'corona_tweets_1216.csv')
tweet1217 = pd.read_csv(path2 + 'corona_tweets_1217.csv')
tweet1218 = pd.read_csv(path2 + 'corona_tweets_1218.csv')
tweet1219 = pd.read_csv(path2 + 'corona_tweets_1219.csv')
tweet1220 = pd.read_csv(path2 + 'corona_tweets_1220.csv')
tweet1221 = pd.read_csv(path2 + 'corona_tweets_1221.csv')
tweet1222 = pd.read_csv(path2 + 'corona_tweets_1222.csv')
tweet1223 = pd.read_csv(path2 + 'corona_tweets_1223.csv')
tweet1224 = pd.read_csv(path2 + 'corona_tweets_1224.csv')
tweet1225 = pd.read_csv(path2 + 'corona_tweets_1225.csv')
tweet1226 = pd.read_csv(path2 + 'corona_tweets_1226.csv')
tweet1227 = pd.read_csv(path2 + 'corona_tweets_1227.csv')
tweet1228 = pd.read_csv(path2 + 'corona_tweets_1228.csv')
tweet1229 = pd.read_csv(path2 + 'corona_tweets_1229.csv')
tweet1230 = pd.read_csv(path2 + 'corona_tweets_1230.csv')
```

```
tweet1231 = pd.read_csv(path2 + 'corona_tweets_1231.csv')

'''
The trend of the volume of COVID-19-related tweets (second wave)
'''
labels = ["15/9", "16/9", "17/9", "18/9", "19/9", "20/9", "21/9", "22/9", "23/9", "24/9", "25/9",
          "26/9", "27/9", "28/9", "29/9", "30/9", "01/10", "02/10", "03/10", "04/10", "05/10", "06/10",
          "07/10", "08/10", "09/10", "10/10", "11/10", "12/10", "13/10", "14/10", "15/10", "16/10", "17/10",
          "18/10", "19/10", "20/10", "21/10", "22/10", "23/10", "24/10", "25/10", "26/10", "27/10", "28/10",
          "29/10", "30/10", "31/10", "01/11", "02/11", "03/11", "04/11", "05/11", "06/11", "07/11", "08/11",
          "09/11", "10/11", "11/11", "12/11", "13/11", "14/11", "15/11", "16/11", "17/11", "18/11", "19/11",
          "20/11", "21/11", "22/11", "23/11", "24/11", "25/11", "26/11", "27/11", "28/11", "29/11", "30/11",
          "01/12", "02/12", "03/12", "04/12", "05/12", "06/12", "07/12", "08/12", "09/12", "10/12", "11/12",
          "12/12", "13/12", "14/12", "15/12", "16/12", "17/12", "18/12", "19/12", "20/12", "21/12", "22/12",
          "23/12", "24/12", "25/12", "26/12", "27/12", "28/12", "29/12", "30/12", "31/12"]

fig, ax = plt.subplots(1, 1)
plt.xticks(rotation = 60)
tick_spacing = 3
ax.xaxis.set_major_locator(ticker.MultipleLocator(tick_spacing))
plt.bar(labels, tweet_volume, align="center", alpha=1, color="blue")
plt.ylabel('Volume')
plt.xlabel("Date")
plt.title('Second wave of COVID19')
plt.show()

second_wave_tweets = pd.concat([tweet0915, tweet0916, tweet0917, tweet0918, tweet0919, tweet0920, tweet0921, tweet0922,
                                tweet0923, tweet0924, tweet0925, tweet0926, tweet0927, tweet0928, tweet0929, tweet0930,
                                tweet1001, tweet1002, tweet1003, tweet1004, tweet1005, tweet1006, tweet1007, tweet1008,
                                tweet1009, tweet1010, tweet1011, tweet1012, tweet1013, tweet1014, tweet1015, tweet1016,
                                tweet1017, tweet1018, tweet1019, tweet1020, tweet1021, tweet1022, tweet1023, tweet1024,
                                tweet1025, tweet1026, tweet1027, tweet1028, tweet1029, tweet1030, tweet1031, tweet1101,
                                tweet1102, tweet1103, tweet1104, tweet1105, tweet1106, tweet1107, tweet1108, tweet1109,
                                tweet1110, tweet1111, tweet1112, tweet1113, tweet1114, tweet1115, tweet1116, tweet1117,
                                tweet1118, tweet1119, tweet1120, tweet1121, tweet1122, tweet1123, tweet1124, tweet1125,
                                tweet1126, tweet1127, tweet1128, tweet1129, tweet1130, tweet1201, tweet1202, tweet1203,
                                tweet1204, tweet1205, tweet1206, tweet1207, tweet1208, tweet1209, tweet1210, tweet1211,
```

```
                                tweet1212, tweet1213, tweet1214, tweet1215, tweet1216, tweet1217, tweet1218, tweet1219,
                                tweet1220, tweet1221, tweet1222, tweet1223, tweet1224, tweet1225, tweet1226, tweet1227,
                                tweet1228, tweet1229, tweet1230, tweet1231], axis=0, ignore_index=True).drop_duplicates()

print(second_wave_tweets.shape)

'''
Choose the English tweets
'''
second_wave_tweets = second_wave_tweets.loc[second_wave_tweets["lang"] == "en"]
print(second_wave_tweets.shape)
# print(second_wave_tweets)

'''
Reset the index of the tweets
'''
second_wave_tweets = second_wave_tweets.reset_index(drop=True)
# print(second_wave_tweets)

'''
Extract the text of tweets
'''
second_wave_tweets = second_wave_tweets["text"]
# print(second_wave_tweets)

'''
Convert the tweets format to tabular data
'''
second_wave_tweets = pd.DataFrame(second_wave_tweets)
# print(second_wave_tweets)

second_wave_tweets["text"] = second_wave_tweets["text"].apply(tweetsPreprocessing)
print(second_wave_tweets)

second_wave_tweets['sentiment'] = second_wave_tweets["text"].apply(sentiment_analyzer)
print(second_wave_tweets)
```

```

positive_count = 0
negative_count = 0
neutral_count = 0
output = open("datasets/twitter_sentiment.txt", "w")

for i in second_wave_tweets["sentiment"]:
    if i < 0:
        negative_count += 1
        output.write("neg")
        output.write('\n')
    if i == 0:
        neutral_count += 1
    else:
        positive_count += 1
        output.write("pos")
        output.write('\n')

x = [positive_count, negative_count, neutral_count]
tot = positive_count + negative_count + neutral_count
positive_count_per = round((positive_count / tot) * 100, 1)
negative_count_per = round((negative_count / tot) * 100, 1)
neutral_count_per = 100 - positive_count_per - negative_count_per
print(positive_count_per, negative_count_per, neutral_count_per)


labels = "positive", "negative", "neutral"
sizes = [positive_count_per, negative_count_per, neutral_count_per]
explode = (0, 0.1, 0.1)
fig, ax = plt.subplots()
ax.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%', shadow=False, startangle=90)
ax.axis("equal")
plt.title('Second wave of COVID19')
plt.show()

'''
Word cloud of 50 most frequent words
```

```
'''
stopwords = set(stopwords.words('english'))
stopwords.update(["40", "jan", "ppl", "amp", "may"])

cv = CountVectorizer(stop_words = stopwords)
words = cv.fit_transform(second_wave_tweets["text"])

sum_words = words.sum(axis=0)

words_freq = [(word, sum_words[0, i]) for word, i in cv.vocabulary_.items()]
words_freq = sorted(words_freq, key = lambda x: x[1], reverse = True)

frequency = pd.DataFrame(words_freq, columns=['word', 'freq'])

frequency.head(50).plot(x='word', y='freq', kind='bar', figsize=(15, 7), color='blue')
plt.title("Top 50 most frequently mentioned words")
plt.show()


wordcloud = WordCloud(max_words=50, width=1500, height=1250,
                      background_color="black").generate_from_frequencies(dict(words_freq))

# Display the generated image:
plt.figure(1, figsize=(12, 10))
plt.imshow(wordcloud, interpolation='bilinear')
# plt.imshow(wordcloud)
plt.axis("off")
plt.show()

```