



NYU

**TANDON SCHOOL
OF ENGINEERING**

COVID-19 Live Sentiment Analysis



Sabarni Kundu, sk7970

Saqib Patel, sip257

Arpit Ranka, aar703

Abstract

On 11th March 2020, the World Health Organization announced COVID19 outbreak as a pandemic. Starting from China, this virus has infected and killed thousands of people across Italy, Spain, USA, Iran and other European countries as well. While this pandemic has continued to affect the lives of millions, many countries have resorted to complete lockdown. During this lockdown, people have taken to social networks to express themselves. In this research, we perform live sentiment analysis of tweets with hashtags '#Covid19', '#Coronavirus', '#Lockdown', '#StaySafeStayHome', '#safehands', '#socialdistancing', '#FlattenTheCurve', and '#WorkingFromHome'. These tweets have been gathered from 10th May 2020 to 14th May 2020, and are all related to COVID19 in some or the other way. We analyse how the citizens of different countries are dealing with the situation. The tweets have been collected, pre-processed, and then used for frequency and sentiment analysis. The results of the study conclude that while the majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and angst exhibited worldwide.

Keywords : '#Covid19' | '#Coronavirus' | '#Lockdown' | '#StaySafeStayHome' | '#safehands' | '#socialdistancing' | '#FlattenTheCurve' | '#WorkingFromHome'

Introduction

The origin of COVID19 is said to be in the starting of December 2019, when several patients from Wuhan, Hubei Province reported severe respiratory infections. These patients had a background of working in the wholesale fish and seafood market, also known as wet markets [1]. In January 2020, the markets were completely closed down and disinfectants were used to sanitize them. On 7th January 2020, the researchers isolated a novel coronavirus which was referred as SARS-CoV-2 or 2019-nCoV. Initially the World Health Organization denied the possibilities of human-to-human transmission 2019-nCoV on 11th January 2020 [2]. However, the confirmed cases continued to soar and on 30th January 2020, World Health Organization declared this COVID19 a Public Health Emergency of International Concern (PHEIC) and an epidemic [3].

By the end of January, the novel coronavirus had already started spreading out to other countries steadily. The number of patients affected by this virus globally was 11,950 on 31st January, 69,197 on 15th February and 86,604 on 28th February. By the time when WHO finally decided to declare COVID19 on 11th March 2020, the number of patients had increased to 126,214, with a total of 4,628 casualties. However, there was an exponential increase in the number in the month of March. On 20th March, there were 275,550 coronavirus affected patients, which meant a rise of more than 100% within 9 days. In the next 11 days, the number of COVID19 patients increased to 858,361, showing a rise of more than 211% and a total of 47,192 deaths were reported till 31st March 2020.

COVID19 has affected more than 166 countries till 31st March 2020. While the case fatality rate of the virus was reported to be 2.5% on 16th February 2020 [4], recent studies show that the CFR for COVID19 can range up to 9.26% [5]. The

countries which have been severely affected by COVID19 includes USA (188,530 patients on 31st March), Italy (105,792 patients on 31st March), Spain (68,200 patients on 31st March), Germany (71,808 patients on 31st March), China (81,554 patients on 31st March), France (52,128 patients on 31st March), UK (23,226 patients on 31st March), Switzerland (16,605 patients on 31st March), Belgium (12,775 patients on 31st March), Netherland (12,595 Patients on 31st March) and Australia (4,763 patients on 31st March) [6].

World Health Organization has suggested that isolation and self-quarantine is one of the major ways to stop this pandemic from spreading at such an alarming rate. China has witnessed benefits of the one of the largest lockdowns at the start of this pandemic, where it locked down 20 provinces and regions. On March 18th, China reported no new local cases for the first time since this pandemic began. This has also encouraged the Chinese government to decide that the lockdown will be lifted on 8th April [7]. Following the suit, countries like Jordan, Argentina, Israel, Belgium and other countries have locked down their countries as well. Similarly, on 25th March, the Indian government took a major decision of locking down the whole nation for 21 days. This can easily be the biggest lockdown the world has ever seen, with 1.3 billion citizens of India being locked down for three weeks [8].

One of the most famous micro blogging site, Twitter has been one of the major ways for information sharing and self-documentation [9]. As the world is fighting with COVID19 since the last four months and the majority of the people are under lockdown, the importance of Twitter has increased more than ever. Even in the past, people have been using twitter to communicate, express and disseminate information related to the crisis, be it cyclones [10], ebola [11], floods [12] or Zika [13]. Twitter has been one of the platforms for millions to express their emotions regarding different issues.

With over 300 million monthly users, the micro-blogging platform Twitter is used increasingly to disseminate public health information and obtain real-time health data using crowdsourcing methods . Researchers analyzed Twitter data to project the spread of influenza and other infectious outbreaks in real time [2]. In 2009, investigators measured the evolving interest in an Influenza A outbreak by analyzing tweet keywords and estimating real-time disease activity and disease prevention efforts . During the Ebola virus (EV) outbreak in 2014, Twitter users publicized pertinent health information from media sources with peak Twitter activity within 24 hours following news events . Tweet content analysis following the EV epidemic discovered that Ebola-related tweets revolved mainly around risk factors, prevention, disease trends, and compassion . Similarly, the 2015 Middle Eastern Respiratory Syndrome (MERS) outbreak, disease spread was found to be correlated with Twitter activity, promoting Twitter as a potential surveillance tool for emerging infectious diseases [6]. During the Zika virus epidemic, Twitter was used to study significant changes in travel behavior due to mounting public concerns . Recognizing Twitter's potential to inform and educate the public, governmental agencies such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) have adopted the use of Twitter and other social media. In the first 12 weeks of the Zika outbreak in late 2015, the WHO Twitter account was retweeted over 20,000 times, demonstrating its widespread impact on disseminating health information.

We postulate that analysis of the content and sentiments expressed on Twitter in the early stages of the coronavirus disease 2019 (COVID-19) pandemic will parallel the spread of the disease and can aid understanding of the effect of the outbreak on the sentiments, beliefs, and thoughts of the general public. Such understanding would enable large-scale opportunities for education and appropriate information dissemination about public health recommendations.

Technique

Catastrophic global circumstances have a pronounced effect on the lives of humans across the world. The ramifications of such a scenario are experienced in diverse and multiplicative ways spanning routine tasks, media and news reports, detrimental physical and mental health, and also routine conversations. A similar footprint has been left by the global pandemic Coronavirus particularly since February 2020. The outbreak has not only created havoc in the economic conditions, physical health, working conditions, and manufacturing sector to name a few but has also created a niche in the minds of the people worldwide. It has had serious repercussions on the psychological state of the humans that is most evident now.

One of the best possible mechanisms of capturing human emotions is to analyze the content they post on social media websites like Twitter and Facebook. Not to be surprised, social media is ablaze with myriad content on Coronavirus reflecting facts, fears, numbers, and the overall thoughts dominating people's minds at this time.

Two techniques have been employed to undertake statistical interpretation of text messages posted on twitter; first being word frequency analysis and second sentiment analysis. A well known and profoundly researched as well as used statistical tool for quantitative linguistics is word frequency analysis. Determining word frequencies in any document gives a strong idea about the patterns of word used and the sentimental content of the text. The analysis can be carried out in computational as well as statistical settings. An investigation of the probability distribution of word frequencies extracted from

the Twitter text messages posted by different users during the coronavirus outbreak in 2020 has been presented below.

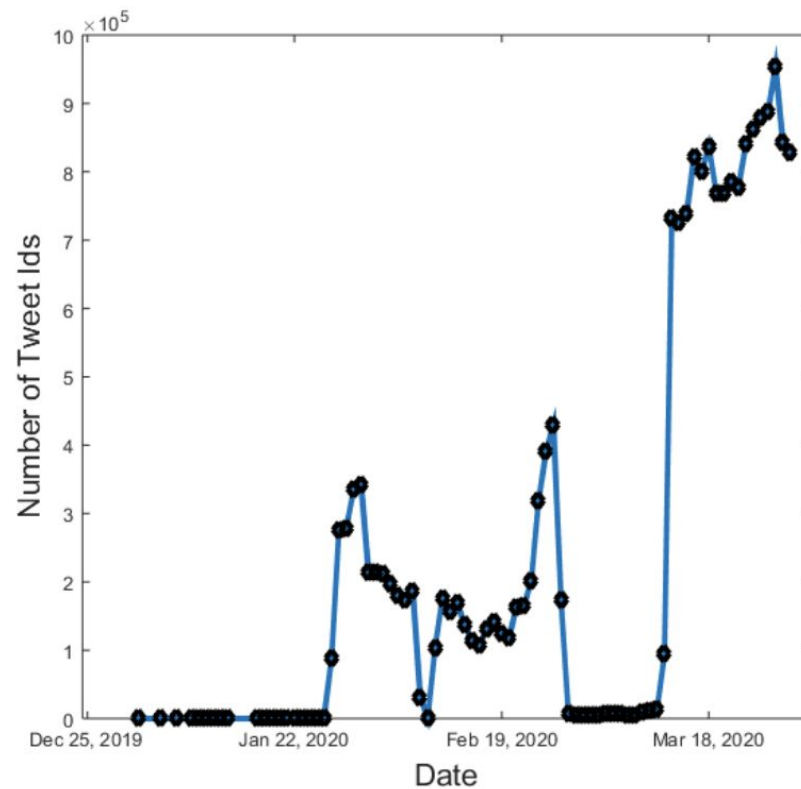


Figure 1: Evolution of number of Twitter ids involved in covid-19 posts

Secondly, Sentiment analysis is a technique to gauge the sentimental content of a writing. It can help understand attitudes in a text related to a particular subject. Sentiment analysis is a highly intriguing field of research that can assist in inferring the emotional content of a text. The sentimental quotient from the tweets has been deduced by computing the positive and negative polarities from the messages.

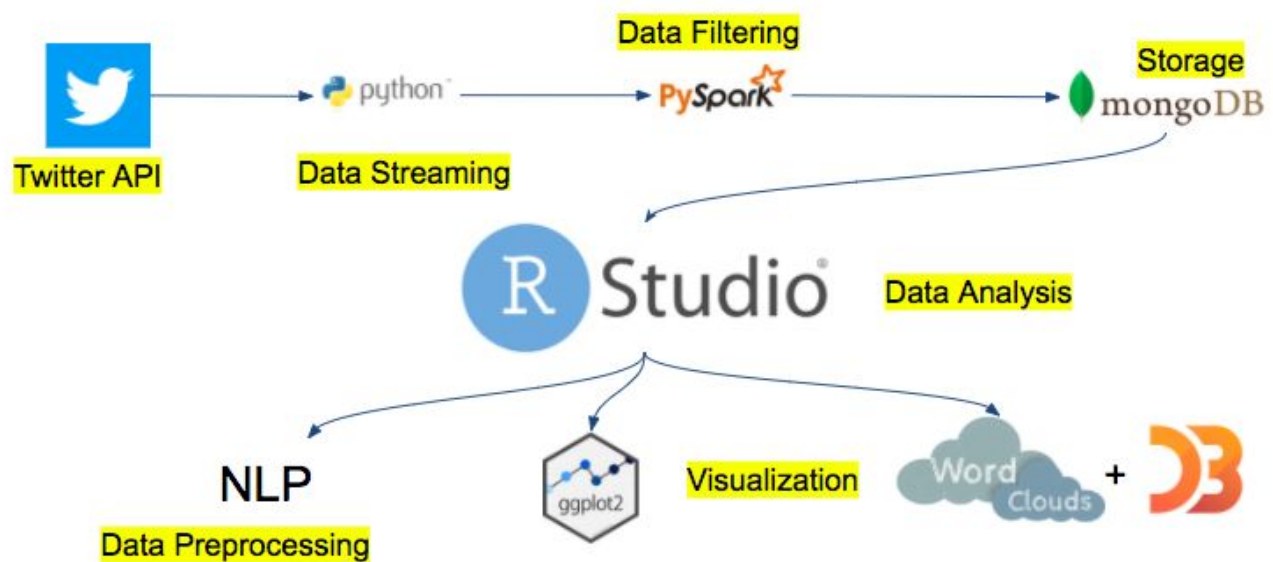
Implementation

A number of big data related technologies have been employed in this project.

They are

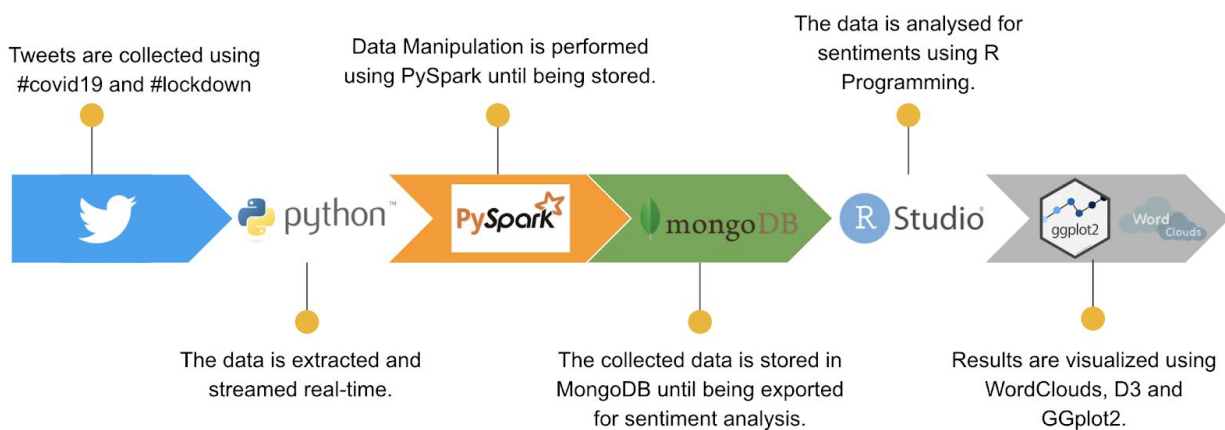
- Twitter API
- Python
- PySpark
- MongoDB
- R Studio
- Ggplot2
- Wordclouds
- D3.js

A flow diagram depicting the architecture of the project can be seen below.



Process :

The overview of the entire implementation is depicted below. Following that , is the detailed description of each step in the process.



Data Collection :

We begin by collecting the real-time tweets relevant to Covid-19 and lockdown. Tweets with hashtags '#Covid19', '#Coronavirus', '#Lockdown', '#StaySafeStayHome', '#safehands', '#socialdistancing', '#FlattenTheCurve', and '#WorkingFromHome' are used for this purpose. In Order to access the Twitter API, we create an app on the Twitter developer site to get the application and consumer tokens and secret keys required for authentication. Then we follow several steps in our python code to stream the data which are listed as follows,

- Authenticate our app to set a connection with the Twitter api.
- Create a StreamListener object which monitors and fetches real time tweets for us.

- Create a stream object which is used to filter out all the relevant tweets based on the track words we provide.
- These track words are the top hashtags trending for covid-19 mentioned above.
- The json we get from the stream object is then directly stored in a Pyspark Dataframe.
- The data in a tweet may contain fields which are futile for our purpose, like usernames. So, we filter out all the other fields and just store the text and the location of the tweets.
- In the final step of data collection, we just transfer our data from the dataframe to a Mongo database using the Pymongo API for python.

For this analysis, we collected more than 1 million tweets in a span of 5 days from 10th May 2020 to 14th May 2020.

Data Analysis :

After obtaining the filtered data, we need to clean the data so that it's fit for our later purposes performing frequency and sentiment analysis.

WordClouds : For word cloud formation, we use a NLP library called `tm` library. For cleaning data in R studio ,we performed the following steps,

- Vectorising the data
- Converting everything in lowercase
- Removing Punctuation,Numbers
- Removing Stopwords
- Removing URLs starting with "https//...."
- Removing Usernames starting with"@...."
- Converting the clean corpus into Matrix which stores the count of each words

- Extracting Top 10 words for test
- Feeding the corpus in WordCloud function where we have set the Random.Order as FALSE, used "Dark" Palette, Given the maximum no. of Words

Sentiment Distribution Bar Graph : For this visualization we have used ggplot2 and have used get_sentiments_nrc()

Below you can see some of the result of get_nrc_sentiments

```
> Test <- "Fed Chairman Jerome Powell said, almost 40% of those in households making less than $40,000 a year lost a job in March during a Wednesday morning briefing. He also called for more fiscal spending by the federal government.."
> get_nrc_sentiment(Test)
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     0              0      0   1   0          1         0       1         2         3
> |
```

```
> Test <- "California hasn't seen the huge death toll from the coronavirus like New York and other hot spots, but the state is still struggling with a growing number of fatalities and confirmed cases."
> get_nrc_sentiment(Test)
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     2              1      1   1   0          1         1       1         1         1
>
```

In the similar manner we used the tweets' text, fed that into get_nrc_sentiments(), obtained the sentiment scores and finally plotted the distribution with the ggplot2 bar graph.

World Map Visualization:

For world map visualization we have used D3.js to show the distribution and comparison of negative and positive sentiments across the world.

D3.js Code Link --- > <https://codepen.io/sabby2928/pen/rNOrZPL?editors=1100>

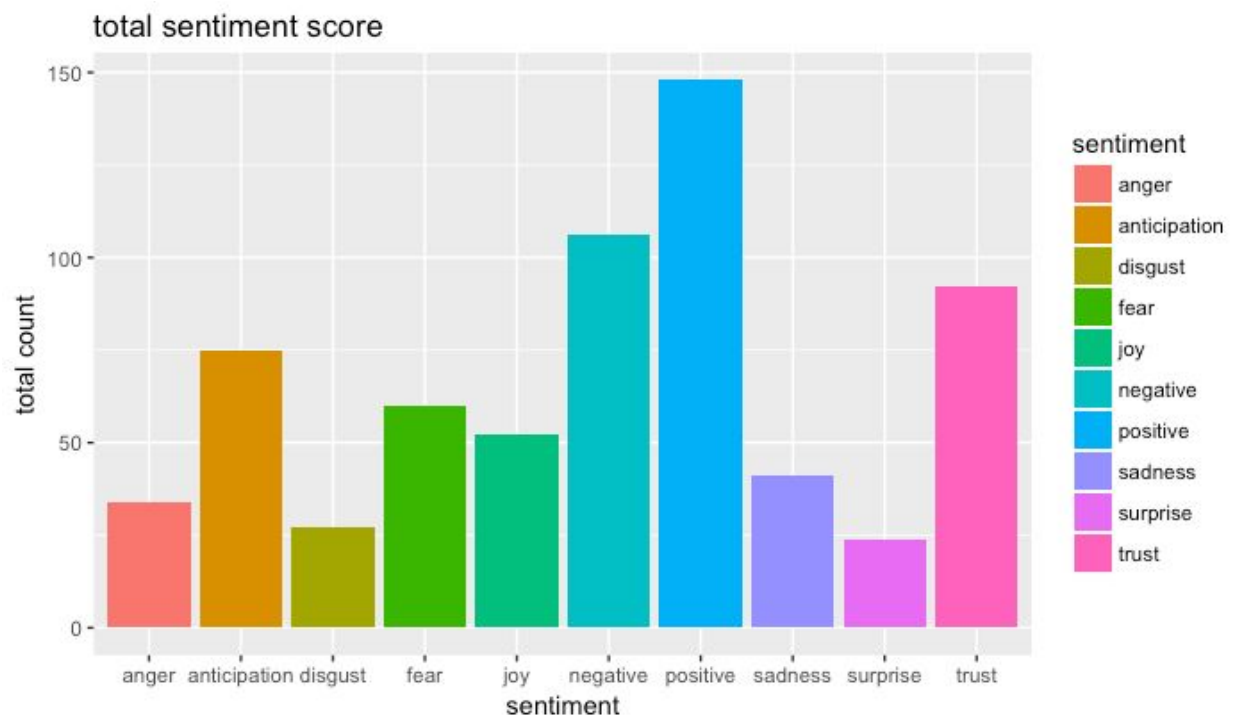
Results



Seen above is the Wordcloud for the collected data. Wordclouds are a great tool in visualizing the results of a frequency analysis. The size of a word in a wordcloud is proportional to their occurrence in the document.

It can be deduced from our wordcloud that the most frequently used are 'people', 'lockdown' and 'social distancing'. Occurrence of these words with such frequency is quite plausible considering that the human race is the most ravaged by this pandemic , leading to many countries imposing lockdown.

And social distancing is by far, the only way WHO has suggested to curb the spread of the virus.



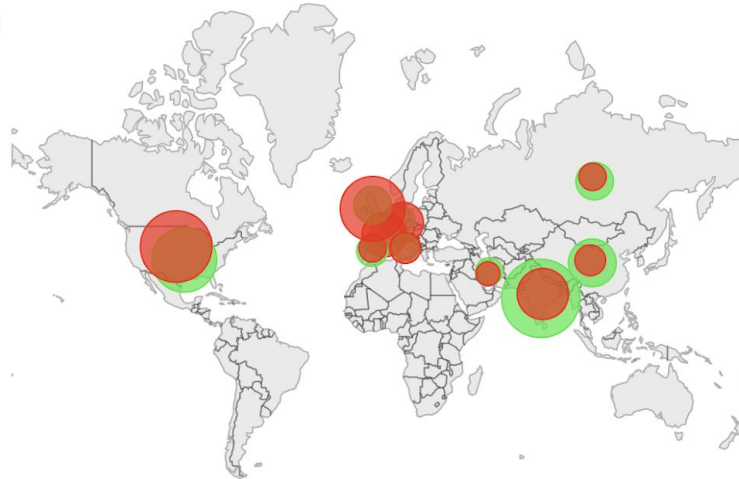
Next is the bar graph depicting the distribution of sentiments extracted from sentiment analysis. On the x-axis are the '6 basic human emotions' by Paul Ekman plus a mix of positive and negative sentiment. Y-axis measures the magnitude of the occurrence of these emotions.

It can be deduced from the bar graph that there's a wave of positivity and optimism around despite the pandemic and that people are placing a lot of trust in their governments and the healthcare workers. However, an equally turbulent wave of negative emotions like fear , anticipation and anger can be observed which can be understood by the fact that this pandemic has left millions unemployed and helpless across the world.

Distribution of COVID-19 sentiments across the world

✓ Positive Sentiments ✓ Negative Sentiments

● Negative Sentiments
● Positive Sentiments



The final result is the distribution of negative and positive sentiments plotted on a world map. The size of the bubbles is proportional to the amount of tweet data available and collected from the respective country.

It can also be extrapolated that there is an excessive negative sentiment than positive in Europe; Asia, per contra, is more optimistic than pessimistic.

Implications and Ethics

There have been some ethical concerns about the way in which Twitter data has been used for research and by practitioners - numerous potential issues have been identified, including the use of Tweets made by vulnerable persons in crisis situations . It is also important to recognize the deviation from researcher obligations to human subjects, to researcher obligations to "data subjects", and this approach does not compromise on ethics, but rather acknowledges the value of publicly available data as voluntary contributions to public space by Twitter users. Past research has also identified the use of Twitter data analytics for pandemics, including the 2009 Swine Flu, indicating a mature stream of thought towards using social media data to help understand and manage contagions and crisis scenarios.

As a global pandemic COVID-19 is adversely affecting people and countries. Besides necessary healthcare and medical treatments, it is critical to protect people and societies from psychological shocks (e.g., distress, anxiety, fear, mental illness). In this context, automated machine learning driven sentiment analysis could help health professionals, policymakers, and state and federal governments to understand and identify rapidly changing psychological risks in the population. Consequently, timely responses and initiatives (e.g., counseling, internet-based psychological support mechanisms) taken by the agencies to mitigate and prevent adverse emotional and psychological consequences will significantly improve public health and well being during crisis phenomena. Sentiment analysis using social media data will thus provide valuable insights on attitudes, perceptions, and behaviors for critical decision making for business and political leaders, and societal representatives.

Conclusion and Future Work

The spread of the disease has created an environment of threat, risks and uncertainty among the population globally. While tweets with misinformation and societal prejudice were present, tweets were also significantly used to disseminate valuable public health information. Twitter offers novel opportunities to public health and governmental agencies to not only track public perception of infectious outbreaks, but also to target messages of a public health nature based on user interest and emotion.


The results of the study conclude that while the majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, anticipation and anger exhibited worldwide.

The tweets collected for this study were in English language which might serve as a limitation for the study. Also, while NRC Lexicon used for this study analyzed the tweets for eight different emotions, it does not count the emotions of sarcasm and irony.

Our work here can aid in identifying a sustainable pathway to recovery post-COVID-19. It will enable policy makers to cater to public needs more specifically and also design sentiment specific communication strategies. Corporations and small businesses can also benefit through this to better understand consumer sentiment and expectations and deliver accordingly.

For the future works, this study can be used to analyze the changing emotions and sentiments of people from these countries and check whether there are major shifts in them over the period of time. It is expected that as the spread of this pandemic will increase, the sentiments and emotions in the tweets may change on the lines of what was seen in the case of China.

References

1. Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." *The Lancet* 395, no. 10223 (2020): 497-506.
2. (WHO), World Health Organization. "Preliminary Investigations Conducted by the Chinese Authorities Have Found No Clear Evidence of Human-to-Human Transmission of the Novel #Coronavirus (2019- NCoV) Identified in #Wuhan, #China . Pic.twitter.com/Fnl5P877VG." Twitter, Twitter, 14 Jan. 2020, twitter.com/WHO/status/1217043229427761152.
3. Jiang, Fang, Liehua Deng, Liangqing Zhang, Yin Cai, Chi Wai Cheung, and Zhengyuan Xia. "Review of the clinical characteristics of coronavirus disease 2019 (COVID-19)." *Journal of General Internal Medicine* (2020): 1-5.
4. "Chinese Center For Disease Control And Prevention". 2020. Chinacdc.Cn. <http://www.chinacdc.cn/en/>.
5. Khafaie, Morteza Abdullatif, and Fakher Rahim. "Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2." *Osong Public Health and Research Perspectives* 11, no. 2 (2020): 74.
6. "Coronavirus Update (Live): 1,410,095 Cases And 81,010 Deaths From COVID-19 Virus Pandemic - Worldometer". 2020. Worldometers.Info. <https://www.worldometers.info/coronavirus/>.
7. Nectar Gan, CNN. 2020. "China To Lift Lockdown On Wuhan, Ground Zero Of Coronavirus Pandemic". CNN. <https://edition.cnn.com/2020/03/24/asia/coronavirus-wuhan-lockdown-lifted-intl-hnk/index.html>.
8. "India's 1.3Bn Population Told To Stay At Home". 2020. BBC News. <https://www.bbc.com/news/worldasia-india-52024239>.
9. Liu, Ivy LB, Christy MK Cheung, and Matthew KO Lee. "Understanding Twitter Usage: What Drive People Continue to Tweet." *Pacis* 92 (2010): 928-939.
10. Soriano, Cheryl Ruth, Ma Divina Gracia Roldan, Charibeth Cheng, and Nathaniel Oco. "Social media and civic engagement during calamities: the case of Twitter use during typhoon Yolanda." *Philippine Political Science Journal* 37, no. 1 (2016): 6-25.

11. Van Lent, Liza GG, Hande Sungur, Florian A. Kunneman, Bob Van De Velde, and Enny Das. "Too far to care? Measuring public attention and fear for Ebola using Twitter." Journal of medical Internet research 19, no. 6 (2017): e193.
12. Nair, Meera R., G. R. Ramya, and P. Bagavathi Sivakumar. "Usage and analysis of Twitter during 2015 Chennai flood towards disaster management." Procedia computer science 115 (2017): 350-358.
13. Fu, King-Wa, Hai Liang, Nitin Saroha, Zion Tsz Ho Tse, Patrick Ip, and Isaac Chun-Hai Fung. "How people react to Zika virus outbreaks on Twitter? A computational content analysis." American journal of infection control 44, no. 12 (2016): 1700-1702.
14. "India Could Be Next Coronavirus Hotspot With 'Avalanche' Of Cases". 2020. South China Morning Post.
<https://www.scmp.com/news/asia/south-asia/article/3075662/india-could-be-next-coronavirushotspot-avalanche-cases>.
15. Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic
By Nikhil Kumar Rajput, Bhavya Ahuja Grover, Vipin Kumar Rathi
16. Twitter Sentiment Analysis on Coronavirus using Textblob by Chhinder Kaur and Anand Sharma
17. An "Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak
18. Sentiment Analysis of Nationwide Lockdown due to COVID 19 Outbreak: Evidence from India by Gopalkrishna Barkur, Vibha, and Giridhar B. Kamath
19. Twitter Sentiment Analysis during COVID19 Outbreak by Dr. Akash D Dubey
20. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification by Jim Samuel1 , G. G. Md. Nawaz Ali1 , Md. Mokhlesur Rahman , Ek Esawi1, and Yana Samuel