Emotions in the digital world

Kelly Wong Jie Yin
School of Computer Science and Engineering

Asst. Prof Farhan Ali
National Institute of Education

Abstract - Microblogging has evolved into one of the most effective platforms for sharing daily life events and news, as well as expressing thoughts on those events. Because social media sites such as Facebook, Twitter, and Reddit are widely utilized, they are excellent sources of public sentiments regarding COVID-19 related events that happened in Singapore during the period from 1 January 2020 to 31 August 2021. Our study mainly consists of two parts.

First, we performed topic modelling using Latent Dirichlet Allocation (LDA) on Facebook posts. We collected Facebook posts from the Ministry of Health (MOH) and Ministry of Education (MOE) and comments made by netizens under each post using Facepager. We further performed sentiment analysis on Facebook comments using an IBM Watson engine.

Advancements in deep learning based language models have been promising for sentiment analysis using social network data. In the second part of the project, we employed multiple language models like long short-term memory (LSTM) recurrent neural networks with a global vector embedding and state-of-art BERT where we classified the text into positive, negative and neutral sentiment classes. Analyses were done on Twitter posts via Snscrape and Reddit submissions and comments via PushShiftAPI. We discovered that sentiment polarity varied from topic to topic.

Overall, our results indicate that there is a spike in the total number of posts when an important announcement related to COVID-19 arrangements is made and negative opinion played an important role in conditioning public sentiment.

Keywords – COVID-19, Singapore, Sentiment Analysis, LDA Topic Modelling, Facebook, Reddit, Twitter, Social Media Platforms.

1 INTRODUCTION

COVID-19 (Coronavirus Disease 2019) Timeline in Singapore

COVID-19 is an viral infection that first appeared in December 2019. The COVID-19 worldwide pandemic was a disastrous occurrence that had a tremendous impact on the global economy, resulting in an increase in unemployment, psychiatric difficulties, and depression. The effects of COVID-19 that resulted in abrupt changes in the economic, social, and travel aspects have prompted researches in a variety of fields focusing on specific countries, such as understanding outreach efforts during the period of COVID-19, performed by Singapore, England and United States Public Health Authorities (PHA), on Facebook, and the public response to the outreach efforts. (Raamkumar et al., 2020)

In Singapore, there were many key significant events happened due to the effects of COVID-19. On 23 January 2020, the first occurrence of COVID-19 in Singapore was reported. As a result, many Singaporeans have worn masks while they are out of home, and adopted social distancing practices. The first cases appearing in Singapore were mainly imported until spread of disease in the community occurred in February and March. Singapore managed to suppress the COVID-19 epidemic with minimal interruption to everyday life during the first several months of the pandemic, while keeping infection rates low in the early months of the outbreak. (Tan et al., 2020).

However, in late March and April, COVID-19 clusters were detected at multiple dormitories, which soon contributed to an overwhelming proportion of new cases in the country In combat to the spread of virus, Singapore implemented the "COVID-19 Control Order", declaring a strict set of preventative measures known as the “Circuit Breaker” on 3 April 2020 which is similar to lockdowns in other countries. Only essential services (supermarkets, delivery services) and enterprises in the global supply chain stay open during the lockdown, and all schools have shifted to home-based learning (HBL) (Ministry of Health, 2020).

On May 19, 2020, three phases of planned reopening were announced, which are "Safe Reopening", "Safe Transition", and lastly, "Safe Nation" (Gov.sg, 2020). The third phase of planned reopening will continue until an effective treatment or vaccine to stop the spread of COVID-19 is discovered, which was approved in December 2020 (Ministry of Health, 2020). “Safe Reopening” stage began on June 2, while “Safe Transition” began on June 19. “Safe Nation” began on December 28th, with preconditions such as the usage of TraceTogether (an application created for contact tracing activities during the period of COVID-19 pandemic), adherence to safety management standards, and enough testing capacity (Gov.SG, 2020).

Despite extensive testing by authorities, the number of cases each day among migrant workers living in dormitories slowly reduced but still remained in the hundreds until early August. While the total case numbers continued to fall, unrelated cases in the community began to rise, and some dormitories that had been cleared of infection were placed on lockdown again (Ang & Phua, 2020). On October 13, a crucial milestone was reached when it was announced that there had been no local instances of infection for the first time since March. However, the first case of the Alpha variant in Singapore appeared on December 24 (Aravindan, 2020).

Singapore tightened restrictions on 23 April 2021, in response to the increasing COVID-19 Delta wave (TTG Asia, 2021). Singapore initially planned to return to Phase 2 on 8 May 2021, however due to the rising Delta variation, it was pushed back to the stage of “Phase 2 Heightened Alert” from 16 May 2021 to 13 June 2021 (Ministry of Health, 2021).

Following a decline in the number of community cases, Singapore went back to "Phase 3 Heightened Alert," with 2 stages at the start of the re-opening phase. The first stage should start from 14 June, with a 5 person restriction for social events, and the second stage should start from 21 June, with the reintroduction of dining-in with a 5 person limit (Baker, 2021). However, a new cluster emerged, resulting in a new spike of community cases. As a result, the initial second step of re-opening was altered such that dining was only allowed for two people may dine (Ang, 2021).

Singapore declared on August 6, 2021, that because 66 percent of its people had been effectively vaccinated, limitations will be relaxed in two stages.

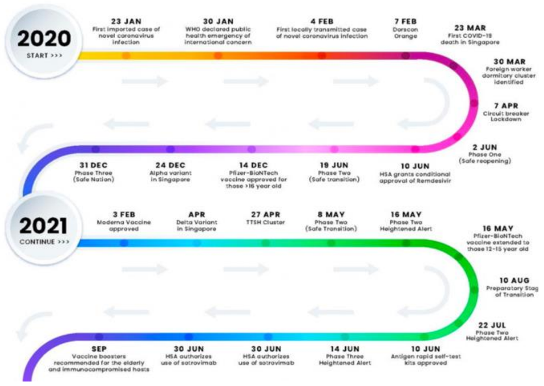

Figure 1: The Timeline of COVID-19 from Jan 2020 to Sep 2021. (Figure taken from Tambyah & Lum, 2022)

Social Media

With the rapid growth of social media platforms, more individuals are expressing their opinions on topics through social media platforms like Facebook, Twitter, and Reddit (Graciyal & Viswam, 2018).

In January 2020, the percentage of social media users in Singapore for Facebook, Twitter, and Reddit are 60.1%, 20.2%, and 0.57% respectively. In August 2021, the percentage of social media users in Singapore for Facebook, Twitter, and Reddit are 54.4%, 16.1%, and 2.74% respectively (StatCounter Global Stats, 2022). Based on studies conducted for year 2020, it was observed that a majority of local Facebook users are aged between 25 and 44 years old. This group of Facebook users is made up of people who were in their teenage or adolescent years during the rise of Facebook (Ng, 2021). In 2020, there were over 1.37 million Twitter users in Singapore, and a majority of local social media users are aged between 25 to 34 years old. The number of Twitter users in Singapore increased to over 1.75 million in 2021 (Statista, 2022). Also, the biggest part of Reddit’s audience skews young and male (Sattelberg, 2021).

As there is a variety of social media platforms each driven by its own set of conceptual frameworks and objectives, the demographics of the users differ for each social media platform. As such, exploring different platforms can lead to different insights, especially in times of crisis. A cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit revealed that users on the two different platforms show different sentiments towards the same earthquake, indicating the importance of investigating cross-platform social media platforms to show a more comprehensive overview of people's general public responses to a specific crisis. (Ruan et al., 2022)

Especially during the increase in the number of COVID-19 cases and stricter lockdown protocols, more people expressed their opinions and concerns about current situations on social media platforms. As a result, monitoring comments made by netizens through social media platforms has been proven to be very useful to understand how the public views public health-related issues. With more understanding of public sentiments, effective communication with the public can be made to further ease public concerns and promote the adoption of important risk-mitigating behaviours.

Topic Modelling

Topic modelling is an unsupervised machine learning technique used to detect abstract topics (a repeating pattern of co-occurring terms in a

corpus) in a collection of documents. It scans a set of documents, to detect word and phrase patterns within the documents, and automatically clusters word groups and similar expressions that best characterize a set of documents to classify each individual document to the detected topics.
During the early stages of COVID-19, topic modeling was combined with Twitter-based sentiment analysis, and feelings such as dread predominated. Tweets from the United States were utilized in region-specific studies to discover the patterns in dominating topics and emotions (Hung et al., 2020).

Sentiment Analysis

Sentiment analysis, also known as opinion mining or emotion artificial intelligence, is a natural language processing (NLP) technique used to determine whether text data is classified into positive, negative, or neutral sentiment classes.
The results acquired from performing sentiment analysis on text data show the sentiments and/or specific emotions of people who expressed their opinions through social media platforms.
Many research have been conducted using deep learning approaches to analyse sentiment on social media data. During the increase in the number of COVID-19 cases in India, a study of sentiment analysis using deep learning-based language models via long short-term memory (LSTM) recurrent neural networks was performed on tweets, and it was discovered that the majority of the tweets expressed positivity, worry, and lack of certainty. (Xue et al., 2020)
There are also other work done for example in China, a study utilised bi-directional encoder representations from transformers (BERT) model to analyse the trends, topics, and sentiments in their test dataset used (Wang et al., 2020). A cross-language sentiment analysis of Twitter comments in Europe during the first two months of the COVID-19 outbreak discovered that lockdown announcements correspond with mood worsening, which rebounds quickly. (Kruspe et al., 2020)
The previous sentiment analysis examples demonstrate how social media may help us understand human psychology and how people react to health-related issues. Furthermore, new breakthroughs in deep learning models as a tool for developing strong language models have offered further incentive in comprehending the attitudes during the initial phase of COVID-19 in Singapore.

Aim

Studies which performed sentiment analysis and topic modelling on COVID-19 related Tweets retrieved from Twitter and geolocated in Singapore have been well-studied (Ridhwan & Hargreaves, 2021).
However, there is a limited grasp of how public attitude and discussion regarding COVID-19 have varied over time across various social media platforms focusing on Singapore. Thus, we aim to perform topic modelling on the Facebook posts, and sentiment analysis on the Facebook comments related to COVID-19 in Singapore under Facebook pages (e.g. MOH and MOE), from 1 January 2020 to 31 August 2021.
We also hoped to determine if the sentiments of netizens on Facebook differ from other social media platforms like Twitter and Reddit. As a result, we performed sentiment analysis for Reddit submissions and comments and Twitter posts from Singapore related to covid19 keywords.

2 METHODOLOGY

In this study, we collected the relevant data from Facebook, filtered and performed text pre-processing before we performed topic modelling using Latent Dirichlet Allocation (LDA) to obtain the dominant topics of discussion on Facebook related to Covid-19. Thereafter, we performed sentiment analysis on the Facebook comments using several methods like IBM Watson Natural Language Understanding API tool. As an advanced step, we employed multiple sentiment analysis methods like long short-term memory (LSTM), which is an advanced recurrent neural networks (RNN) using a global vector embedding (GLoVe) and the state-of-art BERT where we classified the text into positive, negative and neutral sentiment classes. Procedures were approved by NTU Institutional Review Board (IRB-2022-229).

2.1 DATA COLLECTION

Firstly, for data collection, we used Facepager (Jünger, 2019), an application, to fetch public available data from Facebook. We collected Facebook posts created by Ministry of Education (MOE) and Ministry of Health (MOH) and the comments made by netizens under each respective post.
Secondly, we collected the Twitter posts via SNSCRAPE, which is a Python 3 library that can be used as a scraper for social networking services (SNS) (JustAnotherArchivist, 2021).
We queried the tweets using keywords like “covid”, “corona”, “wuhan”, “vaccine”, “nCoV”, “quarantine”, “lockdown” and “circuit breaker”. We also set the parameters of language “lang” to be “en” to query only English tweets and location “geoc” to be Singapore’s GPS coordinates in the format of latitude, longitude, radius to pick up tweets that are geolocated in Singapore.
Lastly, we utilised the Pushshift API, specifically a Python 3 wrapper for Pushshift API called PSAW (PushShift API Wrapper), to collect Reddit data.

To obtain Reddit data related to COVID-19, we collected Reddit submissions related to keywords like “Covid”, “Coronavirus” and “wuhan” under Singapore subreddit (reddit.com/r/Singapore) and the respective comments under each submission via PushShiftAPI (Baumgartner, 2020). Furthermore, we collected Reddit comments which specifically contains the specific keywords, but may not fall under a submission with the COVID-19 related keywords.

2.2 TOPIC MODELLING

Our framework for topic modelling involves a series of steps with major components that includes: 1) Data collection; 2) Text pre-processing; and 3) LDA Topic Modelling.

2.2.1 Text Pre-Processing

As the raw text data are often unstructured, informal, and noisy, the first step of topic modelling after collecting data is to filter and preprocess the data.

Firstly, we filtered the non-English texts using spacy-langdetect, a Python library for language detection. Furthermore, we removed URLs and lower-cased the text. Also, we removed emoticons, punctuations, and numbers, and performed word tokenization on the text using a regular expression tokenizer from Natural Language Toolkit (NLTK). Next, we lemmatized the tokens using the WordNet lemmatizer from NLTK, removed NLTK’s stopwords from the text and formed bigrams with the remaining tokens.

2.2.1 Topic Modelling Steps

Latent Dirichlet Allocation (LDA) is one of the most commonly used generative probabilistic topic modelling algorithms used to uncover hidden topics in a collection of text documents (Blei et al., 2003).

LDA has been used in many works that conducted topic modelling on different types of social media text data like COVID-19 tweets (Jia et al., 2021) and Facebook data (Amara et al., 2021).

LDA makes an assumption that each text document can be represented as a finite mixture over an underlying list of topics, where each topic is thus represented as an infinite mixture with a set of probabilities.

We need to select a value for the number of topics to be set for our LDA topic model, and then choose the number of topics that gives the highest coherence score in a plot of coherence values.

We followed the guidelines (Řehůřek, 2020) to choose values for the four main parameters for our topic modelling process using LDA: chunk size, passes, alpha, and eta. We set chunk size as 2000, which is more than the total number of text documents in the dataset, thus taking in all documents at once. We set our passes parameter to 20, picked an argument eval_every = 1, and then checked if most documents have converged by the final passes. Lastly, we set alpha parameter to be 'auto' and eta parameter to be 'auto', such that the model automatically learn these parameters. We performed training on the LDA models with different values for the number of topics ranging from 5 to 30 and compute the coherence score value of each topic model using the parameters, where we decided the upper bound estimate for the number of topics to be set for the LDA model to be 8.

We interactively visualized the LDA model's eight topics using a pyLDAvis visualization, in which eight topics are presented in a graphable 2D space and comparable topics are clustered together. This was accomplished by employing an approach to reduce the dimension of our feature set, which is t-SNE (Van der Maaten & Hinton, 2008), for topic clustering, as Medford et al. (2020) had done for their study on COVID-19 tweets. We also used Word Cloud to visualize the keywords for each topic (Oesper et al., 2011).

2.3 SENTIMENT ANALYSIS

Our framework for sentiment analysis involves a series of steps with major components that includes: 1) Data collection; 2) Pre-Processing of Text data; 3) Building Model and training LSTM and BERT; and 4) Making Predictions on selected COVID-19 related data.

2.3.1 Text Pre-Processing

As the raw text data that we are performing sentiment analysis on are often short and unstructured, the first step of sentiment analysis is to filter and pre-process the data. Similar to our filtering process performed for topic modelling, we filtered the non-English texts and removed URLs, hyperlinks, user mentions, punctuations, numbers, and excessive spaces. Furthermore, for tweets, we removed the phrase ‘RT’ (representing re-tweet) and user information, web links, and hashtags. Also, for Reddit posts and comments, we used BeautifulSoup to parse the texts and regular expression package (RE) to remove certain noises, like special characters. However, we did not filter out emoticons and emojis as they may possibly affect sentiment classification. Also, for sentiment analysis, we extracted the most commonly seen acronyms and Singlish slang in the text data and performed definition mapping on the acronyms and slang.

2.3.2 Sentiment Analysis Model

After performing text cleaning on the raw text data, we used IBM Watson Natural Language Understanding API to classify the sentiments of text data (IBM, 2017), which was used as our “ground truth” data for the training of models (LSTM and BERT).

The next step is to convert each word into its corresponding specific GLoVe embedding vector values and in our case, we converted each word to 300 dimensional vector. We chose GLoVe embedding since it has demonstrated strong results in the literature for sentiment analysis models (Pennington et al., 2014). Once the model architecture has been built, the aforementioned vector from each word thus retrieved is given on to the respective LSTM models for training. We evaluated our trained LSTM and BERT models with our Reddit dataset where the labels have been predicted using IBM Watson engine and then use them for sentiment analysis on the COVID-19 related datasets retrieved earlier, to classify the text into sentiments such as “Positive”, “Negative”, and “Neutral”.

The model hyper-parameters of the LSTM model are determined depending on how the model performs in our trial experiments. The GloVe embedding provides a word vector of size 300 to represent a word. For LSTM models with 300 input units, two bidirectional LSTM layers with 128 and 128 hidden units, and ultimately an output layer for sentiment classification, we employ a dropout regularisation probability of 0.2. (Hochreiter & Schmidhuber, 1997).

To evaluate the model, we used metrics like recall, precision, accuracy, and F1-score values which are defined using True Negative (TN), True Positive (TP), False Positive (FP), and False Negative (FN). The performance metrics are defined as follows.

Accuracy = (TN+TP)/(TN+TP+FN+FP)

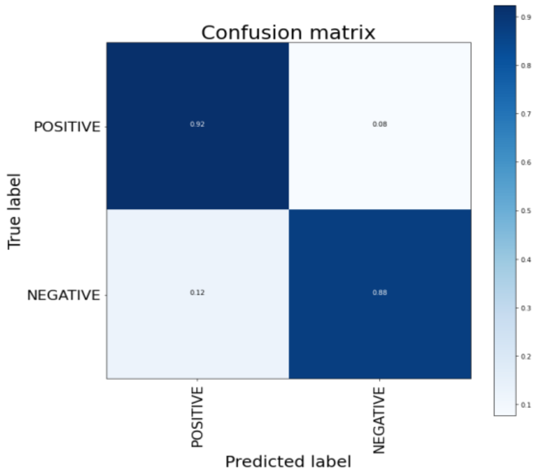

Figure 2: Confusion Matrix for probability of actual against predicted labels.

Recall = TP/(FN+TP)

Precision = TP/(FP+TP)

F1-Score = 2 * (Precision * Recall)/(Precision + Recall)

In the case of our BERT model, we merely modified the learning rate and epsilon value, and utilised the default hyper-parameters for our BERT-base-uncased model which has a hidden size of 768, 12 Transformer blocks (Vaswani et al., 2017), and 12 self-attention heads. Following after, we pre-trained BERT with a batch size of 64, a maximum sequence length of 128, a learning rate of 5e-5, and 4 epochs (Devlin et al., 2020).

We utilised accuracy, the Receiver Operating Characteristic curve (ROC curve), which is a useful tool when predicting the probability of a binary outcome, the Area Under the Curve (AUC), and F1-Score to evaluate our BERT model. It is a probability curve that shows the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values, thereby separating the signal from the noise. The AUC is the measure of a classifier’s ability to distinguish between classes and is used to summarise the ROC curve.

True Positive Rate (TPR) = TP/(TP + FN)

False Positive Rate (FPR) = FP/(TN + FP)

```
AUC: 0.9772
Accuracy: 92.15%
```

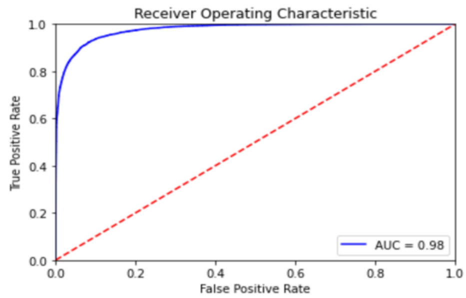

Figure 3: Plot for Receiver Operating Characteristic, AUC value and Accuracy value for pre-trained BERT model.

2.4 Analysing Data

We carefully identified a number of important positive and negative event dates related to COVID-19 in Singapore (Refer to Appendix A) and found the number of positive and negative posts from 10 days before to 20 days after the specific event day. After getting all the values for each specific event date, we used them to find the average of positive and negative posts for all events within the period of 31 days.

3 RESULTS AND DISCUSSION

We first present our results acquired from Topic Modelling performed on Facebook posts created by Ministry of Health (MOH) to get the dominant topics among the posts. We also performed topic modelling on the Facebook posts created by MOE, but due to space constraints, results are not plotted here. In our next step, we present our results acquired from Sentiment Analysis on the Facebook comments, Tweets and Reddit posts and comments related to COVID-19.

3.1 TOPIC MODELLING RESULTS

3.1.1 Facebook MOH Topics

A total of 8 topics were found using the LDA model. Within the pyLDAvis visualization of topics, the circles represent clusters of topics, and the size of the circle represents the frequency of the topic while hovering each circle allows us to examine the top words for each topic cluster (Carson & Kenneth, 2014).

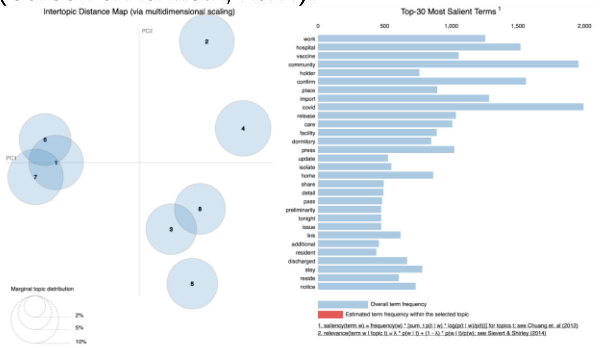

Figure 4: pyLDAvis visualisation of 8 topics for LDA model.

We manually assign labels to each of the eight subjects after obtaining a list of words with a high chance of belonging to each dominating topic. This was accomplished by manually reviewing the posts that weighed heavily under each topic and verifying the list of words that have a high chance of belonging to each topic to gain a better understanding of the content of MOH Posts with high topical content. Finally, we compiled our findings for all dominant topics and manually assigned a label to each topic. The full list of the 8 labelled topics is presented as follows.

Topic 1: Cases linked to specific events
Topic 2: Cases in Dormitory
Topic 3: Updates on Community Cases
Topic 4: Situation update of Measures implemented
Topic 5: Detection of Asymptotic cases
Topic 6: Situation update about cases in Hospitals and Community care facilities
Topic 7: Situation update about imported cases who have been placed on Stay-Home Notice (SHN)
Topic 8: Vaccination

We plotted each topic with their predicted topic proportions in terms of the percentage of total posts to investigate which topics were dominant in the MOH Posts dataset.

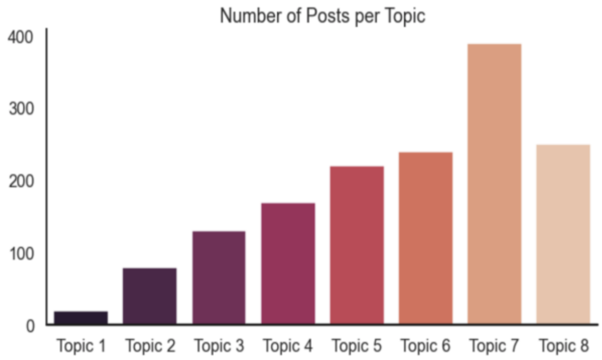

Figure 5: Number of Posts per Dominant Topic.

The most prevalent topic in MOH Posts is Topic 7 which is Imported cases who have been placed on Stay-Home Notice. Furthermore, based on Figure 4, we can tell that the least prevalent topic in MOH Posts is Topic 1 which is specifically about Cases linked to specific events (e.g. Church and SAFRA Jurong).

Figure 6: Word Cloud of Top 3 Dominant Topics with the highest number of posts, with the most right Word Cloud being for imported cases (Topic 7), the centre being for vaccination (Topic 8), and the most left being for situation update about cases in Hospitals and Community care facilities (Topic 6).

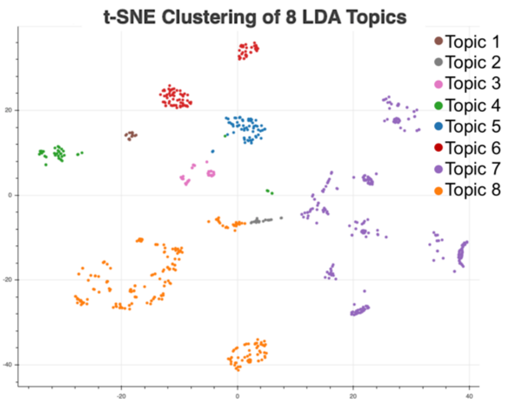

Figure 7: t-SNE plot to visualise the clusters of documents (MOH posts) in 2-dimensional space.

3.2 SENTIMENT ANALYSIS RESULTS

An overall sentiment analysis using various methods like IBM Watson Natural Language Understanding API, LSTM model and BERT model was performed on the text data collected from social media platforms. The sentiments of each text data acquired using BERT will be used for further data analysis.

3.2.1 Sentiment Analysis Models

An overview of the models we used to predict the sentiments for our test datasets like MOH comments, MOE comments, Tweets, and Reddit posts and comments.

Models	Dataset predicted
IBM Watson Engine	Facebook MOH, MOE and Reddit
LSTM	Facebook MOH, MOE, Reddit and Twitter
BERT	Facebook MOH, MOE, Reddit and Twitter

Plots in sections below are based on the results generated using BERT model for consistency and the validation accuracy for BERT (94%) is higher than that for LSTM (91%).

3.2.2 Facebook MOH

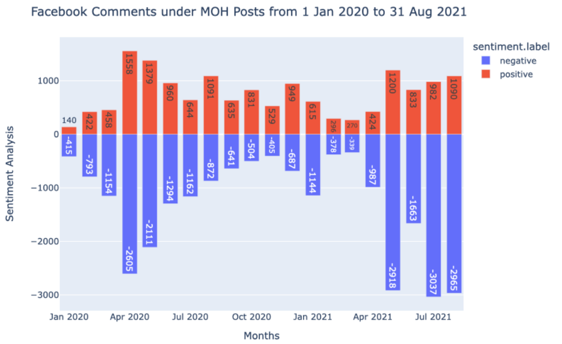

Figure 8: Number of Facebook Comments posted in each month under MOH Posts from 1 January 2020 to 31 August 2021.

Overall, there are more negative comments than positive comments to MOH posts. Based on Figure 8, we observed an increase in the number of Facebook comments to MOH posts from January 2020 to April 2020, followed by a decrease in the number of comments. However, the number of comments started to increase from November 2020 to January 2021, followed by a short decrease in comments and then a steeper increase in number of comments, especially in the number of negative comments, to August 2021.

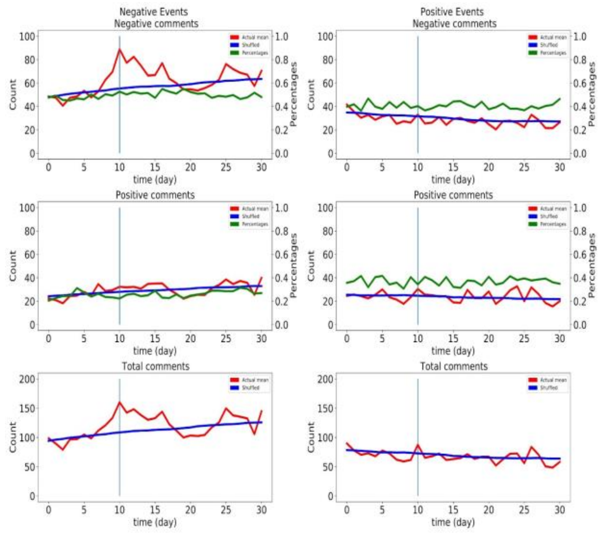

Figure 9: Average of the number of Negative, Positive and Total comments posted by netizens under MOH Posts based on Negative and Positive event days selected (Appendix A), with Day 10 as the day of announcement of measures implemented by the Singapore government.

Based on Figure 9, we observed a trend for Negative events where there is an increase in the number of negative and total comments from 10 days before the exact day of announcement, then followed by a decrease in the number of negative comments and the total number of comments. Also, some spikes in the number of comments could be observed on Day 16, which is 6 days after the event day, and Day 25, which is 15 days after the event day.

Furthermore, for Positive events, we observed different trends from trends in posting patterns for negative events. As for positive events, the number of negative comments posted were below the shuffled plot during the period of 10 days before and 20 days after the event day, except for Day 0, which is 10 days before the day of announcement, and Day 25, which is 15 days after the day of announcement.

3.2.3 Facebook MOE

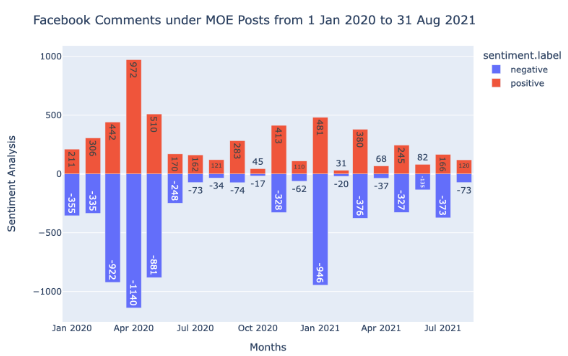

Figure 10: Number of Facebook Comments posted in each month under MOE Posts from 1 January 2020 to 31 August 2021.

Based on Figure 10, we observed a similar trend as in Figure 8, which is an rise in the number of Facebook comments under MOE posts from January 2020 to April 2020, followed by a decrease in the number of comments. The number of comments started to has a sharp increase in the number of negative comments in January 2021, followed by a short decrease in comments. However, as compared to Figure 8, the number of comments from April 2021 to August 2021 were relatively consistent.

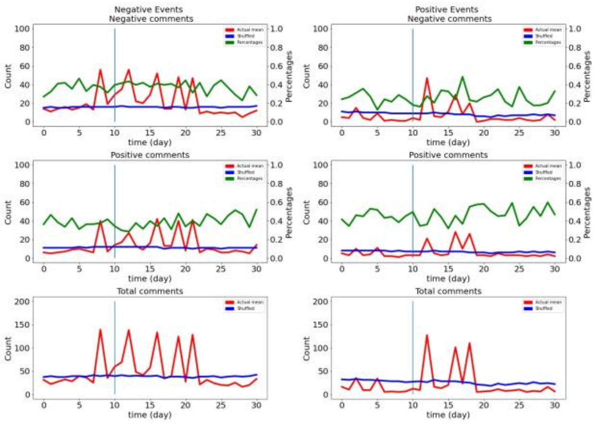

Figure 11: Average of the number of Negative, Positive, and Total comments posted by netizens under MOE Posts based on Negative and Positive event days selected (Appendix A), with Day 10 as the day of announcement of measures implemented by the Singapore government.
Overall, the total number of comments is much less as compared to other datasets retrieved from other social media platforms, including the comments under MOH Page during the study period which is likely due to MOE page having a more specific target audience group.

3.2.4 COVID-19 related Tweets

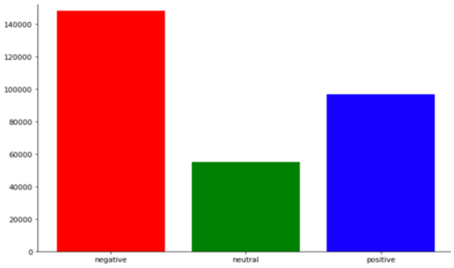

Figure 12: Distribution of sentiments in Tweets

Based on Figure 12, we observed that the proportion of negative sentiments which is about 50% in our Tweets dataset is higher than the proportion of positive sentiments which is about 30% and neutral sentiments which is about 20%. As the total number of negative sentiments takes such a large proportion in our Tweets dataset, we would like to find out the hidden trends in the sentiments results we retrieved during the study period. As such, we plotted Figure 13 to directly compare the trends in the number of positive and negative comments over time.

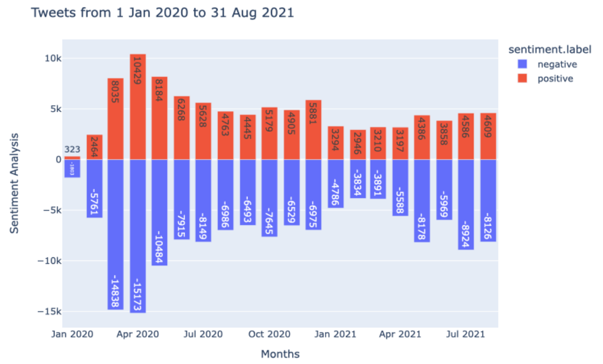

Figure 13: The number of COVID-19 related Tweets posted each month from 1 January 2020 to 31 August 2021.

Based on Figure 13, we discovered that there is a peak in the number of positive and negative Tweets in April 2020 which is the period of circuit breaker lockdown. Also, the trends in the number of Tweets followed Figure 8 (Distribution of comments under MOH Page).

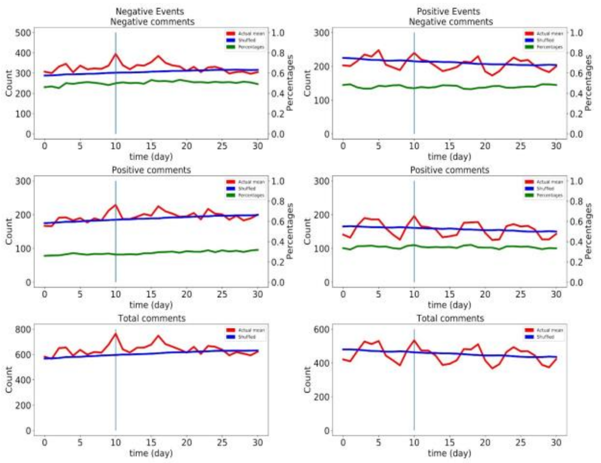

Figure 14: Average of the number of Negative, Positive and Total Tweets posted by netizens based on Negative and Positive event days selected (Appendix A), with Day 10 as the day of announcement of measures implemented by Singapore government.

Overall, the total number of Tweets is 174877, which is much more than the number of comments under the Facebook Pages (MOH and MOE) during the study period which is likely due to the pages having a specific target audience group. Based on Figure 14, we observed that the number of negative comments during the negative event period has a similar trend as the MOH comments (as shown in Figure 9), where there is a peak in the number of negative tweets on the announcement day, followed by another smaller spike on Day 16. However, we also noticed some differences for example, in the first 20 days of the study period of 31 days, the number of negative comments lie above the shuffled plot, which indicates that during the first 20 days of a specific negative event, there tends to be more negative comments than usual. A similar trend was noticed for positive tweets and all tweets during negative events.

As for the positive events, there is a fluctuation in the number of tweets in general during the 31 days. We also observed that before the event day, there is a fall in the number of tweets, and a rise in the tweets count on the event day, followed by a decrease in the number of tweets.

3.2.5 Reddit Posts and Comments related to COVID-19

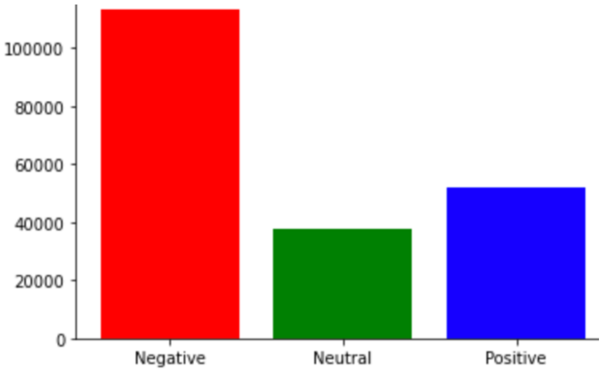

Figure 15: Distribution of sentiments in Reddit posts and comments.

The number of Reddit posts and comments collected is the most among all datasets, which is 203573. Based on Figure 15, we observed that the proportion of negative sentiments follows the Tweets dataset where the majority of sentiments is negative which has a proportion of more than 50%, followed by positive and lastly neutral sentiments.

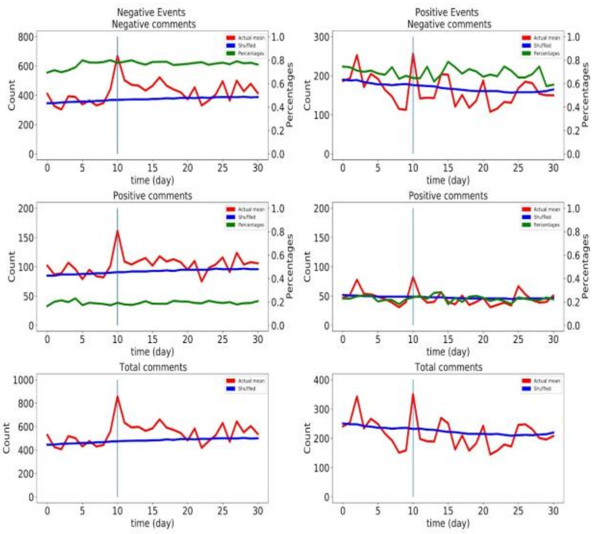

Figure 16: Average of the number of Negative, Positive and Total Reddit posts and comments based on Negative and Positive event days selected (Appendix A), with Day 10 as the day of announcement of measures implemented by the Singapore government.

Surprisingly, the posting pattern of netizens on Reddit is similar to netizens who posted under Facebook MOH page during negative events. However, for positive events, from Figure 16, we can observe obvious and sharper spikes in the number of comments on the event day, followed by fluctuations in the number of positive and negative comments. Also, we discovered a similar trend as compared to Figure 14 (Number of comments during positive event), which is the decrease in the number of comments before a sharp spike in the number of comments on the actual event day, which could possibly be because of the netizens observing improvements in situation which leads to less discussion on current situation required. However, as an important announcement on measures was given by the Singapore government, topics of discussion could be brought up by netizens, which results in a rise in the number of posts and comments.

4 CONCLUSION

Emotions in the digital world

In summary, for the first part of Topic Modelling, we explored the 8 dominant topics of MOH posts, which shows the main focus of issues related to COVID-19 was the imported cases and vaccination during our study period from 1 January 2020 to 31 August 2021. Furthermore, we performed sentiment analysis on our datasets acquired from 3 different social media platforms including Facebook, Twitter and Reddit using different methods like IBM Watson Natural Language Understanding API, LSTM model and BERT model. We believe that the most important takeaways from our results for sentiment analysis is that there is a spike in the total number of posts when an important announcement related to COVID-19 arrangements is made and negative opinion played an important part in the public sentiments for all 3 different social media platforms.

Future Directions

We believe that there are many other ways to study the text data retrieved from the social media platforms. Topic modelling could be performed on the Tweets and Reddit posts to determine if the topics of concern differ on different social media platforms to have more effective communication. Also, a comparison can be made for the sentiment polarity based on issues of concern between these social media platforms. Furthermore, a cross-country comparison between the posting patterns could bring interesting insights.

ACKNOWLEDGMENT

I wish to thank Assistant Professor Farhan Ali and his team in the Learning Sciences and Assessment Division at NIE for patiently guiding me and providing me with constructive feedback throughout my journey with them during this URECA term.

I would also like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project

REFERENCES

[1] Sesagiri Raamkumar, A., Tan, S. G., & Wee, H. L. (2020). Measuring the outreach efforts of public health authorities and the public response on Facebook during the covid-19 pandemic in early 2020: Cross-country comparison. *Journal of Medical Internet Research*, *22*(5). https://doi.org/10.2196/19334

[2] Tan, J. B., Cook, M. J., Logan, P., Rozanova, L., & Wilder-Smith, A. (2020, December 31). Singapore's pandemic preparedness: An overview of the first wave of covid-19. International journal of environmental research and public health. Retrieved June 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7795346/

[3] Ministry of Health. (2012). Circuit Breaker To Minimise Further Spread of COVID-19. Singapore. Retrieved June 8, 2022, from https://www.moh.gov.sg/news-highlights/details/circuit-breaker-to-minimise-further-spread-of-covid-19

[4] Graciyal, G., & Viswam , D. (2018). Freedom of expression in social media: A political perspective. Retrieved June 8, 2022, from https://www.researchgate.net/publication/324388685_Freedom_of_Expression_in_Social_Media_A_Political_Perspective

[5] Published by Statista Research Department. (2022, April 8). Singapore: Number of users on Twitter 2022. Retrieved June 8, 2022, from https://www.statista.com/statistics/490600/twitter-users-singapore/

[6] StatCounter Global Stats. (2022). Social Media Stats Singapore. Retrieved June 9, 2022, from https://gs.statcounter.com/social-media-stats/all/singapore/2021

[7] Ridhwan, K. M., & Hargreaves, C. A. (2021, June 11). Leveraging twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. International Journal of Information Management Data Insights. Retrieved June 9, 2022, from https://www.sciencedirect.com/science/article/pii/S2667096821000148

[8] Jünger, Jakob / Keyling, Till (2019). Facepager.

[9] JustAnotherArchivist. (2021). *snscrape: A social networking service scraper in Python*. GitHub. Retrieved June 30, 2022, from https://github.com/JustAnotherArchivist/snscrape

[10] Baumgartner J., Zannettou S., Keegan B., Squire M., Blackburn J. (2020). The Pushshift Reddit Dataset. arXiv:2001.08435.

[11] Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543.

[12] Gov.sg. (2020). Ending Circuit Breaker: Phased approach to resuming activities safely. MCI - Gov.SG. Retrieved June 26, 2022, from https://www.gov.sg/article/ending-circuit-breaker-phased-approach-to-resuming-activities-safely

[13] Ministry of Health. (2020). *News highlights*. Ministry of Health. Retrieved June 28, 2022, from https://www.moh.gov.sg/news-highlights/details/moving-into-phase-three-of-re-opening

[14] Aravindan, A. (2020, December 24). *Singapore confirms first case of new coronavirus variant found in UK*. Reuters. Retrieved June 28, 2022, from https://www.reuters.com/business/healthcare-pharmaceuticals/singapore-confirms-first-case-new-coronavirus-variant-found-uk-2020-12-24/

[15] Ang, H. M., & Phua, R. (2020). *Covid-19: 20,000 migrant workers to be discharged by end-May, but cases from dormitories likely to remain high*. CNA. Retrieved June 28, 2022, from https://www.channelnewsasia.com/singapore/covid-19-migrant-workers-dormitories-discharged-circuit-breaker-937606

[16] TTG Asia. (2021, April 23). *Singapore further tightens border measures for travellers from India*. TTG Asia. Retrieved June 28, 2022, from https://www.ttgasia.com/2021/04/23/singapore-further-tightens-border-measures-for-travellers-from-india/

[17] Ministry of Health. (2021). *Updates on local situation, border measures for Bangladesh, Nepal, Pakistan, Sri Lanka, and Thailand and precautionary measures to minimise transmission from Tan Tock Seng Hospital*

cluster. Ministry of Health. Retrieved June 28, 2022, from https://www.moh.gov.sg/news-highlights/details/updates-on-local-situation-border-measures-for-bangladesh-nepal-pakistan-and-sri-lanka-thailand-and-precautionary-measures-to-minimise-transmission-from-tan-tock-seng-hospital-cluster

[18] Baker, J. A. (2021, June 10). *Up to 5 in a group allowed from Jun 14; dining-in May resume on Jun 21 in phased easing of covid-19 curbs*. CNA. Retrieved June 28, 2022, from https://www.channelnewsasia.com/singapore/cap-of-5-people-social-gatherings-dining-in-phase-3-covid-19-1846791

[19] Ang, H. M. (2021). *Dining-in to resume from Jun 21 but in groups of two, not five*. CNA. Retrieved June 28, 2022, from https://www.channelnewsasia.com/singapore/covid-19-dining-in-resume-groups-of-2-people-jun-21-moh-1960036

[20] Ministry of Health. (2021). *Updates on Phase 3 (heightened-alert) measures*. Ministry of Health. Retrieved June 28, 2022, from https://www.moh.gov.sg/news-highlights/details/updates-on-phase-3-(heightened-alert)-measures-7Jul

[21] Tambyah, P., & Lum, L. (2022, February 10). Looking to 2022: What lies ahead in the covid-19 pandemic. Retrieved June 29, 2022, from https://news.nus.edu.sg/looking-to-2022-what-lies-ahead-in-the-covid-19-pandemic/

[22] R. Řehůřek. (November 4, 2020). Gensim: Topic Modelling for Humans. LDA Model - gensim (2020). Retrieved from https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html

[23] Oesper, L., Merico, D., Isserlin, R., & Bader, G. D. (2011). WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code for Biology and Medicine*, *6*(1), 7.

[24] IBM. (2017). IBM Watson

[25] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit

[26] Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social network analysis of COVID-19 Sentiments: application of artificial intelligence. Journal of medical Internet research. 2020;22(8):e22590. pmid:32750001

[27] Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PloS one. 2020;15(9):e0239441.

[28] Wang T, Lu K, Chow KP, Zhu Q. COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model. Ieee Access. 2020;8:138162–138169.

[29] Kruspe A, Häberle M, Kuhn I, Zhu XX. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. arXiv preprint arXiv:200812172. 2020.

[30] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[31] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[32] Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543.

[33] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

APPENDIX

Appendix A

The negative events selected for data analysis are as follows:

1. 23 Jan 2020: The first case in Singapore was reported to be a 66-year-old man from Wuhan.
2. 4 Feb 2020: Singapore reported its first case of infection in the local community, which was a 28-year-old Singaporean working at Yong Thai Hang, a Chinese health products shop in Lavender.
3. 7 Feb 2020: MOH upgrades the Dorscon level from yellow to orange, which resulted in a surge of frantic purchasing.
4. 21 March 2020: The virus caused the first 2 deaths in Singapore.
5. 23 March 2020: Peak of daily imported cases. From 11.59pm on 23 March, all short-term tourists are no longer permitted to enter or pass through Singapore.
6. 26 March 2020: All pubs, theatres, and entertainment venues are closed.
7. 3 April 2020: Announcement of Circuit Breaker Lockdown (3 April)
8. 7 April 2020: Circuit breaker measures go into effect on April 7, 2020. Most businesses are closed, and eating out is prohibited.
9. 14 April 2020: The wearing of a mask when one is out of home becomes compulsory.
10. 21 April 2020: According to PM Lee, the circuit breaker measures will be prolonged till 1 June 2020. The restrictions are also strengthened, with barbershops and bubble-tea shops shutting.

11. 19 May 2020: Circuit Breaker Termination, Announcing the phases to Resuming Activities Safely
12. 7 July 2020: Phase 2 re-opening
13. 23 Sep 2020: More Activities Resumed with Safety measures
14. 24 Dec 2020: The first occurrence of Alpha variant in Singapore
15. 28 December 2020: Entering Phase 3 of Re-opening
16. 23 April 2021: Singapore imposed restrictions due to the increase in the number of COVID-19 Delta variant cases.
17. 29 April 2021: Two new cluster were created from 7 cases in the community connected to a 38-year-old Singaporean ICA officer, and from 8 community cases connected to a 46-year-old Filipino nurse who worked at Tan Tock Seng Hospital.
18. 14 May 2021: Dine-in would be prohibited, and the maximum number of people allowed in a social gathering was lowered to 2.
19. 10 June 2021: "Phase 3 Heightened Alert", with a two-stage re-opening. The first stage would begin on 14 June, with a 5 person restriction for social events, while the second phase would start on 21 June, with a 5 person restriction on dining-in.
20. 15 July 2021: Another new cluster of cases was developed at numerous KTV lounges from three community instances related to a short-term travel permit holder from Vietnam who tested positive on 11 July, resulting in a large rise in number of cases.
21. 20 July 2021: Jurong Fishery Port restored to "Phase 2 Heightened Alert" status from 22 July to 10 August. On August 22nd and 24th, 2021, North Coast Lodge will open additional clusters.
22. 24 Aug 2021: A few new groups were observed at North Coast Lodge on 23 August, Bugis Junction on 24 August.

The positive events selected for data analysis are as follows:

1. 19 May 2020: Resumption of activities under the category of Travel Agencies and Tour Operators on May 19, 2020. The Multi-Ministry Taskforce declared on May 19, 2020, that Singapore would exit the Circuit Breaker when it expires on June 1, 2020, and will resume activities in three phases.
2. 2 June 2020: Phase 1 begins on June 2, 2020. (Circuit Breaker ends). Furthermore, beginning on June 2, 2020, Singapore will progressively resume economic activities that do not pose a high risk of transmission. High-risk social, economic, and recreational activities will stay closed.
3. 15 June 2020: Phase 2 begins on June 15, 2020. The Multi-Ministry Taskforce declared on June 15, 2020, that Singapore would enter Phase Two on June 18, 2359 hours (19 June 2020).
4. 23 September 2020: Resuming additional activities safely on September 23, 2020.
5. 28 September 2020: Announcement of COVID-19 Advisories for Visitors and Tourism Businesses.
6. 14 October 2020: Advisory for Phase 2 re-opening of Food & Beverage establishments
7. 27 October 2020: Arrangements for the Results Release for the 2020 National Examinations
8. 17 November 2020: Announcement on advisories for Phase 2 re-opening of retail establishments and lifestyle-related services
9. 14 December 2020: Moving into Phase Three of Re-opening
10. 28 December 2020: Phase 3 starts
11. 8 January 2021: Announcement for guideline for Phase 3 Re-opening of Food & Beverage Establishments
12. 27 February 2021: Over 26000 Home Team Frontline workers have received the Vaccination for COVID-19
13. 8 March 2021: COVID-19 vaccination is made available to all seniors, as well as vital service professionals and high-risk groups.
14. 24 March 2021: Extension of vaccination program, greater relaxation community measures.
15. 23 April 2021: Announcement of guidelines for re-opening of Food and Beverage in Phase 3.
16. 31 May 2021: COVID-19 Vaccination available for all students in schools and Institutes of Higher Learning (IHL).
17. 2 July 2021: Seven more clinics were authorised to deliver the SINOVAC-CORONAVAC COVID-19 vaccine.
18. 6 August 2021: MOM Resuming Entry Approval for Vaccinated Work Pass Holders and their Dependants.
19. 19 August 2021: The next steps in our journey towards Resilience during the COVID-19 period.