Analyzing Risk Communication During COVID-19 Pandemic

IT 7993 Capstone Report

Kennesaw State University

College of Computing and Software Engineering

Project Site

https://sites.google.com/view/capstoneproject1fall2020

Team

Elaine Harris, Team Lead

Michael Farris

Maryline Kwa

Sang Nguyen

Gabriel Oyebanji

Raviteja Pasumarthi

Chelsey West

Contents

[Executive Summary 3](#_Toc55841208)

[Project Presentation 3](#_Toc55841209)

[Introduction 3](#_Toc55841210)

[Background 4](#_Toc55841211)

[Project Scope 4](#_Toc55841212)

[Objectives 4](#_Toc55841213)

[Deliverables 5](#_Toc55841214)

[Related Studies 5](#_Toc55841215)

[Research Method 6](#_Toc55841216)

[Design and Implementation 7](#_Toc55841217)

[Data Collection 7](#_Toc55841218)

[Data Extraction 8](#_Toc55841219)

[Data Shaping and Cleanup 10](#_Toc55841220)

[Topic Modeling and Text Classification 11](#_Toc55841221)

[Latent Dirichlet Allocation (LDA) 11](#_Toc55841222)

[Multiclass Classification with Deep Learning 13](#_Toc55841223)

[Behavior Keyword Identification 18](#_Toc55841224)

[Statistical Modeling 19](#_Toc55841225)

[Time Series Analysis 19](#_Toc55841226)

[Findings 19](#_Toc55841227)

[Is Social Media an effective way to communicate mitigations strategies? 19](#_Toc55841228)

[Social Media Engagement Statistics 19](#_Toc55841229)

[User Engagement by Topic 21](#_Toc55841230)

[Average User Response by Topic - Facebook 22](#_Toc55841231)

[Average User Response by Topic – Twitter 25](#_Toc55841232)

[Is there a correlation between public health messaging and confirmed positivity rates? 29](#_Toc55841233)

[Topic Frequency vs Infections Trends 29](#_Toc55841234)

[Behavior Frequency vs Infections Trends 32](#_Toc55841235)

[Are there daily downward trends in positivity rates when mitigation mandates are issued? 35](#_Toc55841236)

[Stop importation of new cases - Border Control measures 35](#_Toc55841237)

[Limit spread among elderly population 36](#_Toc55841238)

[Limit spread in the local population 36](#_Toc55841239)

[Timeline of Behavior Messaging in Social Media 36](#_Toc55841240)

[Conclusion 37](#_Toc55841241)

[Project Summary 37](#_Toc55841242)

[Limitations and Future Direction 38](#_Toc55841243)

[References 39](#_Toc55841244)

[Appendix A – Text Classification and Deep Learning Resources 40](#_Toc55841245)

[Introduction to Text Classification 40](#_Toc55841246)

[Introduction to Deep Learning and Word Embeddings 40](#_Toc55841247)

[Appendix B – SPSS and Statistical Modeling Resources 42](#_Toc55841248)

[Appendix C – Additional Resources for Data Context Interpretation 43](#_Toc55841249)

[Acknowledgements 44](#_Toc55841250)

# Executive Summary

Social media channels including Facebook and Twitter are popular platforms for the dissemination of news alerts. These mechanisms of information distribution can reach millions of people worldwide and daily messaging through these channels is imperative for organizations to connect with citizens locally, regionally, and globally.

These communications have been magnified during the COVID-19 pandemic. Official government channels use Twitter and Facebook to communicate important updates, alerts, and public health warnings about current health recommendations and trends, as well as local mandates such as face-mask requirements in public settings.

This research project seeks to assess and analyze the relationship between public health advisory statements during the COVID-19 Pandemic and the trends on new infections in Singapore. Social Media posts from the Singapore Ministry of Health (<https://www.moh.gov.sg/>) Facebook and Twitter accounts are extracted and cross-referenced against the World Health Organization data on daily new infections reported in Singapore.

# Project Presentation

Our presentation can be viewed at <https://sites.google.com/view/capstoneproject1fall2020/home>.

Direct links to presentation and supporting materials:

Recorded Project Presentation:

<https://youtu.be/qcsLEZi4WEU>

Short Video:

<https://youtu.be/DGaHAFEDcFg>

Power BI Workbook:

<https://app.powerbi.com/view?r=eyJrIjoiNWU4YWM2NDEtODhlZS00MGExLWFiYjYtYWE4ZTE5OGUxOGIyIiwidCI6IjA4ZjQ2OTg4LTMxMWMtNGRlYS05YWRiLTRjMjQwNmU3MGExOSIsImMiOjF9>

GitHub link to download all supporting files:

<https://github.com/eharr147/IT7993-capstone.git>

# Introduction

The COVID-19 pandemic has fundamentally altered our daily routines. Public places close intermittently, social gatherings are limited by size and location, and facial coverings are now required in many countries in attempt to slow the spread of the virus. From the onset, recommendations and guidelines have been issued through public health departments and government agencies. At times, however, information can be conflicting depending on what channel it is found on, which agency posted the material, and simply because with a novel virus, science is constantly changing as more research is available.

## Background

This project is sponsored by the Office of International Health & Biodefense and Bureau of Oceans and International Environmental and Scientific Affairs, both agencies within the United States Department of State Diplomacy Lab. The geographical area of focus is Singapore. The first case of COVID -19 in Singapore was confirmed of January 23, 2020 (Wei et al., 2020). Throughout the pandemic mitigation strategies such as frequent handwashing, social distancing and wearing face coverings in public spaces where social distancing is not feasible have been broadcast on media platforms globally. Our research will focus on the messaging from Singapore health officials and will seek to identify trends between messaging and daily case totals. We are interested in the following questions:

Our research will identify topics related to COVID-19 that the Ministry of Health posts and compare those topics to case counts. Any correlations will be reported to our sponsors.

## Project Scope

The scope of this project includes:

* Historical data extracted from Singapore Ministry of Health official Twitter and Facebook feeds between February 01, 2020 and August 01, 2020.
* Daily COVID-19 infection logs published by the World Health Organization between January 23, 2020 and August 28, 2020.

## Objectives

* To identify any trends between official messaging from government agencies in Singapore regarding COVID-19 and daily positive case counts.
* To identify the topics of official messages, classify the posts according to identified categories and perform basic analytics on social media user responses to different topic categories.
* To identify certain behaviors contained in the messages (wearing masks, washing hands, avoiding crowds, etc.), count the frequency of these messages in the social media feed, and identify if changes in the frequency of such messages are related to changes in the daily positivity case counts.

## Deliverables

1. Research report with findings on the relationship between the social media data collected and new COVID-19 cases
2. Power BI Workbook (including data and visualizations)
3. SPSS Workbook(s) with data and statistical analysis results
4. All data files utilized
5. All source code used during extraction and preparation

# Related Studies

The scope and objectives of this project are specific to the novel Coronavirus outbreak in Singapore and the Singapore Ministry of Health’s presence on social media platforms Twitter and Facebook. Since the first cases of COVID-19 were identified in December 2019, there is not longitudinal research regarding responsiveness to official messaging. There are, however, a few studies that have been published in the short time that the virus has been circulating, and studies from previous crises and the public response to the dissemination of information.

The most pertinent study is from Spain: “Covid-19 communication management in Spain: Exploring the effect of information-seeking behavior and message reception in public’s evaluation” (Moreno et al., 2020). The researchers created a questionnaire that was sent by invitation through WhatsApp, Twitter, Facebook, Instagram, and LinkedIn that allowed invitees to not only take the quiz but encouraged them to send the survey link to their contacts (Moreno et al., 2020). The study sought to assess how “information forms and sources influence the public’s information-seeking behaviors, and the perception of government’s crisis response strategies during the pandemic” (Moreno et al., 2020). The survey was active between March 14, 2020 and April 14, 2020.

The sample included 546 completed questionnaires, which is considered a representative sample based on the adult population in Spain. Topics of the survey included which channels citizens used for information, perception of the government’s response, and levels of trust in information sources. Researchers found that many people were unable to “attribute correctly the information provided by public authorities” (Moreno et al., 2020). This is important because regardless of the accuracy of information disseminated by officials, the constant barrage of information from many different sources overshadows the messaging. Additionally, the results show the decline in public trust of authorities, specifically the World Health Organization (WHO). Specific to our study, researchers found that:

“People who were mainly informed through *Twitter* (53.4%, p ≤ 0.01) and *Facebook* (52.5%, p ≤ 0.01) strongly believed that the government’s communication caused social alarm, and confused the population (50.7 and 49.5%, respectively). However, most audiences for all media agree with the statement “The government has not revealed the whole truth,” especially *Twitter* users (57.1%, p ≤ 0.01) and print press readers (56.7%, p ≤ 0.01)” (Moreno et al., 2020).

The research presented in this report focuses on Twitter and Facebook messaging from official government channels and examines the response to that messaging and the effect it has on daily case counts. An aspect that is not further researched due to the data utilized) is if users in Singapore feel similarly to citizens in Spain about the government’s communications.

# Research Method

To complete this study, a variety of data extraction, data analyses methods and machine learning techniques are utilized.

Data Collection: Singapore official messages are collected directly from publicly available posts intended to raise awareness to the COVID-19 situation. Social Media text extraction is one of the many techniques used under Big Data Analytics, and it is applicable to this project because we are not directly sponsored by the Singapore Ministry of Health and must rely on public sources.

Text Analytics: it is not possible to analyze raw text in any useful way. In order to extract analytical context from the social media posts, a variety of techniques are applied. Depending on the objective, Machine Learning algorithms and/or manual observation are applied.

1. Topic Modeling (Identification): to answer the question “What are the posts about?”, we will start with a Machine Learning approach for Topic Modeling known as LDA (Latent Dirichlet Allocation). The intent is to allow the algorithm to identify topics based on word frequency distribution. After implementation, we assessed that the topics were not clear enough to perform analytics and required a different approach: Supervised Text Classification.
2. Text Classification: using the broad topics generated by Topic Modeling, create a set of categories by which the social media posts can be analyzed. These categories will be used to perform statistical analyses on user reaction and correlation analysis to infection trends. We chose to use a Machine Learning approach to create a Deep Learning Model on a sample of the social media posts that is then applied to the entire dataset for automatic text categorization.
3. Behavior Identification: we are interested in measuring how much of the social media contents focuses on individual behavior modifications, such as wearing masks, hand washing, avoiding social gatherings, telecommuting and other safe distance measures. Any combination of behaviors may show in any post, regardless of the main topic. We will identify these behavior clues by pattern match to the original post text.
4. Government Mandates Timeline: by visual observation, we identify social media posts related to government announcement of new or lifted restrictions on travel, social gatherings and economic activities. This is done to help us interpret drastic changes in the Infections Trends over time.

Statistical Analysis: median and outlier measurements are used to measure user response to each topic identified in the posts. The objective is to identify which topics in general and which posts in particular elicited the most responses from the public, as indicated by the number of Likes an Shares. SPSS software is used to calculate descriptives and grapg boxplots of user responses by text category.

Correlation Analysis: Pearson Correlation is calculated in SPSS and aims at identifying the existence of a relationship between the frequency of posts for a given topic or behavior and trends in new infections. For each post, we calculate the number of infections on that day and 14 days later. We then aggregate the posts by month and topic or behavior and calculate:

* + Frequency of topic/behavior to total posts that month
  + Average number of daily infections for that month
  + Average number of daily infections 14 days after the post for that month
  + Bi-variate Pearson Correlation between Topic/Behavior Frequency and Average Number of Daily Infections
  + Bi-variate Pearson Correlation between Topic/Behavior Frequency and Average Number of Daily Infections after 14 Days

***Disclaimer:*** *we don’t have any datasets that clearly identify actual behavior changes in response to government messages. We will use the average number of infections trend as a surrogate, assuming that changes to individual behavior might have some level of influence in the number of new infections.*

Regression Models: the COVID-19 Pandemic falls in the scope of Infectious Diseases and Epidemiology Modeling. This project team doesn’t have the background or the data required to create the complex models required to predict trends in contagion. It is naïve to assume that a Linear Regression Model on the number of future infections can be created based on government educational efforts alone, without taking into account the many variables utilized by professional epidemiologists when creating specialized infectious disease spreading models. As such, this project shall limit its scope to the analysis of possible correlations between the choice and frequency of government official messages and the recorded trends of new infections.

Data Visualization: Data Visualization is both popular and useful in conveying otherwise complex messages to a broad audience. In this project, findings and conclusions are based on statistical analysis, but we also provide engaging and broad-reaching visualizations created in Power BI that are meant to provide a compelling and engaging message to all audiences.

# Design and Implementation

## Data Collection

This project utilizes three data sources: Social Media extract from Twitter and Facebook, and the Daily COVID-19 infections log published by the World Health Organization (WHO).

Facebook and Twitter posts from the official ***Singapore Ministry of Health*** pages were collected from **02/01/20** to **08/01/20**. Public replies or comments in either social media platform were not extracted. Additionally, daily case counts for the period were obtained from the World Health Organization (WHO).

### Data Extraction

The number of daily cases and fatalities is published by the World Health Organization (WHO). The Project Sponsor supplied a data feed for Singapore called ***Singapore.xlsx***.

The WHO data contains 8 fields (Table 1) and was merged with the social media data using Excel Power Query.

Table 1:Layout of 'Cases' file

|  |  |
| --- | --- |
| Field Name | Description |
| Date\_reported | Calendar Date cases were reported |
| Country\_code | Literal “**SG”** |
| Country | Literal “Singapore” |
| WHO\_region | Literal “WPRO” |
| New\_cases | Reported New Cases on Date\_reported |
| Cumulative\_cases | Running Total of New Cases from first date reported (01/23/2020) |
| New\_deaths | Reported New Deaths on Date\_reported |
| Cumulative\_deaths | Running Total of New Deaths from first date reported (01/23/2020) |

#### Twitter Data Collection

Twitter is the social media platform used for text-based social network analysis. Created in 2006, Twitter has seen a total of 1.3 billion accounts created in its history and over 500 million tweets sent each day. The official Singapore Ministry of Health page currently has 66,600 followers.

With Python being the programming language of choice for most Data Scientists, libraries such as Python-Twitter and Tweepy are generally used as Python interfaces to the Twitter Search API in order to access the information on Twitter. However, this technique poses a problem. Twitter Search, and by extension its API, are not meant to be an exhaustive source of tweets. The Twitter Streaming API places a limit of just one week on how far back tweets can be extracted from that match the input parameters. So, in order to extract all historical tweets relevant to a set of search parameters for analysis, the Twitter Official API needs to be bypassed and custom libraries that mimic the Twitter Search Engine need to be used. One of these, courtesy Jefferson Henrique & Dmitry Mottl is called GetOldTweets3 (GOT3).

To extract the data an open source distribution of Python and R programming, Anaconda, is used. Anaconda can be downloaded and installed for free and provides a great graphical interface that is user friendly. Anaconda can be downloaded at <https://www.anaconda.com/products/individual>.

To install the GetOldTweets-python program in the command line:

python3.6 -m pip install GetOldTweets3

The metadata provides information about the id of the tweet, username the tweet was made from, the date and time it was tweeted, the number of retweets, favorites and mentions the tweet has received, the textual and hashtag data that the tweets contains and the permalink for the tweet. All tweet ids are 64bit integers and each tweet has an URL associated with it that is referred to as the permalink.

Table 2: Metadata

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Available Python Output Fields | | | | |
| *id (string)* | *permalink (string)* | *username (string)* | *text (string)* | *date (date)* |
| *retweets (integer)* | *favorites (integer)* | *mentions (string)* | *hashtags (string)* | *geo (string)* |
| <https://github.com/Jefferson-Henrique/GetOldTweets-python> | | | | |

In the below code we extracted the data from the specified user “@sporeMOH” from 2020-02-01 to 2020-08-02. The data was then pre-processed and formatted to an Excel workbook.

Text

Description automatically generated

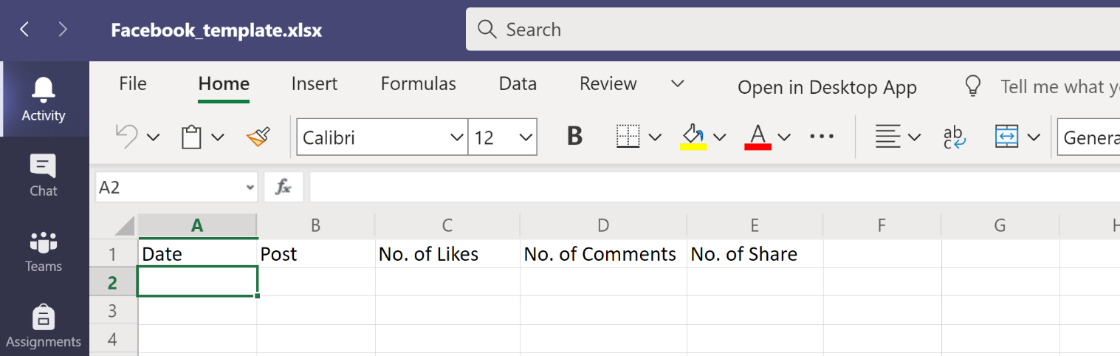
Figure 1: Sample Code

#### Facebook Data Collection

The Facebook data gathering process was manual in nature due to the increased restrictions on accessing Facebook via a script or API. Each data collector scrolled down the page until finding the month they were tasked with collecting data from. The date, post content, total reactions (likes), number of comments, and number of shares for each post during the assigned month were extracted. Captured data is saved in multiple Excel files, one for each team member involved in the collection process.



Figure 2: Sample Facebook Post



*Figure 3: Facebook Collection Template*

### Data Shaping and Cleanup

The data from each social media channel is basically the same, however, the raw extracts are shaped slightly differently. To facilitate future analysis, the raw data was transformed to standardize field names, clean up text data and consolidate the Facebook files.

All data shaping operations were performed in Excel, using Power Query as a personal ETL tool. Power Query is easy to use and performs very well on small data volumes, such as the extracts used in this project.

The final product of the Power Query transformations was published as regular Excel Sheets that can be easily imported into SSPS and Power Query.

#### Data Shaping Technique

1. Create an Excel Workbook.
2. Import data from raw Twitter Extract.
3. Place all Facebook extracts in the same folder. Import the Folder. Power Query will generate the steps to combine all files located in the folder. New files can be added to the folder later, and the Power Query steps will automatically detect and process the new file.
4. Combine both extracts in a standard layout. Create Source field as 1 for Facebook and 2 for Twitter.
5. Use the following key to rename and reorder the columns into standard layout:

|  |  |  |  |
| --- | --- | --- | --- |
| **Standard Field** | **Facebook** | **Twitter** | **Description** |
| **Source** | Literal 1 | Literal 2 | Data Source Id |
| **Date** | Date | Date | Post or Tweet Date |
| **Text** | Post | Text | Text |
| **Shares** | No. of Share | retweets | Number of times post was shared by users |
| **Likes** | No. of Likes | Favorites | Number of times post was liked by users |
| **Comments** | No. of Comments | N/A | Number of comments (Facebook only) |

1. Apply Text clean transformation to remove all non-readable characters.
2. Remove all commas from text, in case data needs to be exported as CSV.
3. Change all text to lowercase, for future Text Topic Modeling.
4. Publish Shaped Data as Excel Sheets in the same workbook (Load to Table option in query).

## Topic Modeling and Text Classification

Topic Modeling is defined as a technique to automatically identify topics in a corpus (collection of texts). The technique requires the use of some Unsupervised Machine Learning algorithm to read the text and identify topics and keywords without human intervention.

### Latent Dirichlet Allocation (LDA)

A popular algorithm for Topic Modeling is Latent Dirichlet Allocation (LDA). The algorithm is offered in several popular open source Python Libraries, including Gensim and Keras.

Topic Modeling with LDA “builds a topic per document model and words per topic model, modeled as Dirichlet distributions 1” (Li, 2020), and the algorithm works in two steps:

First, the algorithm detects a set of topics across all documents. Each topic consists of a set of words selected based on the calculated probability of those words appearing together. In a large corpus, the top *x* words are selected to represent a topic.

Second, the algorithm evaluates each *document* (in our case, each post is a document) against the list of topics. A probability distribution is calculated that results in a match score to each topic. For example, if we had 3 topics, document A might be scored as 60% topic 1, 35% topic 2 and 5% topic 3.

Text

Description automatically generatedInitially, topic modeling was completed using LDA. Execution of the Gensim implementation of LDA is as simple as cleaning up the text and instantiating the model in Python:

Figure 4: LDA Topic Modeling

The performance of the model is measured by its *Coherence Score*. Generally speaking, the higher the score, the better. Conveniently, Gensim also provides a Coherence Model to measure the performance of the LDA model.

The final model was implemented using the Gensim library wrapper of the Mallet LDA Algorithm. A total of 15 topics were created by the algorithm, and the Dominant Topic for each post was identified by selecting the topic with the highest percentage contribution.

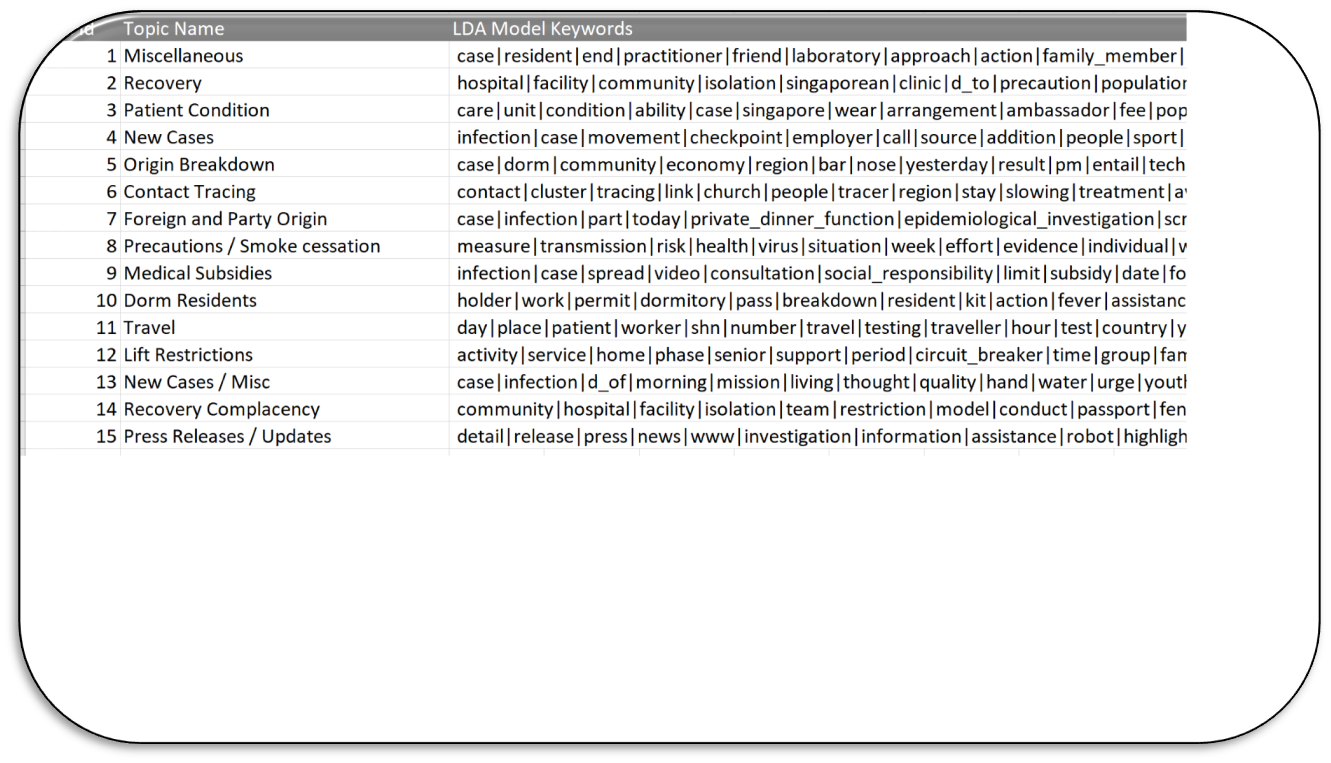


Figure 5: LDA Topics Identified

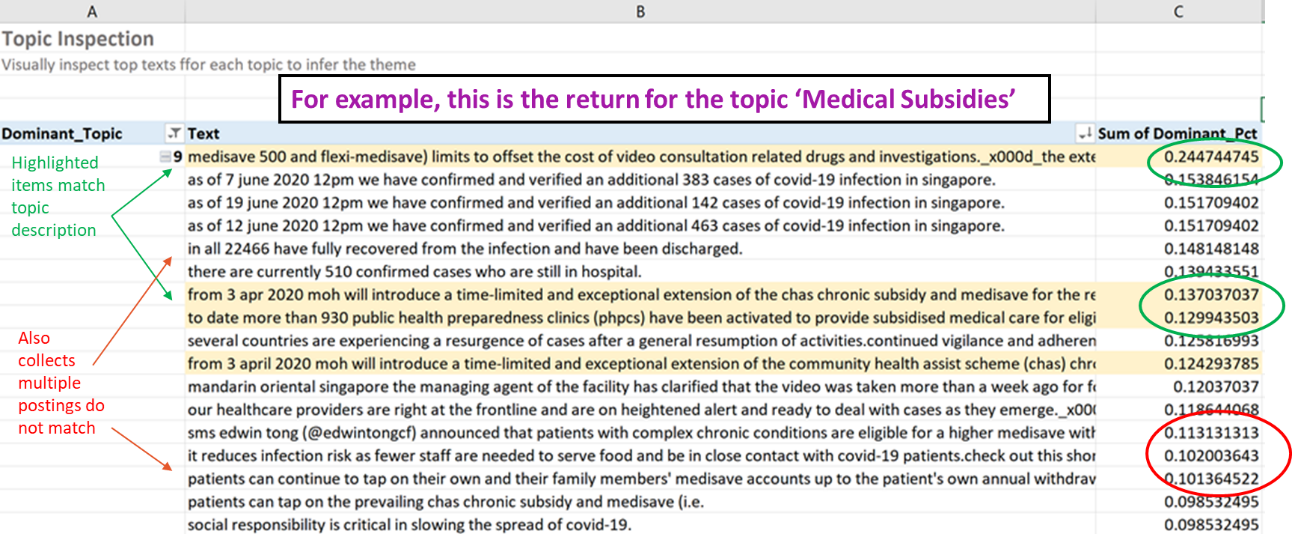
After topics were identified and given a collective description, visual observation easily determines that the posts that more closely match the dominant topic appear to be representative of the subject. However, many posts match all topics at low probabilities, and even the highest match (the dominant topic) does not appear to fit with the other posts. *Figure 6* shows an example for *Topic 9 – Medical Subsidies*. The highlighted posts match the topic description, but there are multiple other posts that are not a good match. This same issue appears on most of the topics identified by the LDA Analysis.

Figure 6: Topic 9-Medical Subsidies

The resulting topics derived from LDA do not provide a clear distinction between the identified topics. Furthermore, LDA failed to single out particular subjects such as social distancing, facial coverings, etc. In order to provide an accurate analysis, the project team decided to switch approach from Unsupervised Topic Modeling to a Supervised Learning Classification method.

The social media dataset represents a *Multiclass Classification* problem, meaning that the algorithm will be trained to classify each sentence into one of multiple pre-identified categories. The categories are based on the topics identified by the LDA algorithm but refined to discard low frequency topics and consolidate similar ones. Furthermore, because this is a Supervised Learning approach, a subset of the social media dataset must be manually labeled.

### Multiclass Classification with Deep Learning

Deep Learning algorithms are implemented based on guidance found in the following articles: Multi-*Class Text Classification with LSTM* (Li, 2020) and *Using Deep Learning for End to End Multiclass Text Classification* (Agarwal, 2020). At this point, we prefer to use Multiclass classification and obtain a single label per post, but we are aware that Multi-label classification could also be a valid solution.  Part of the dataset was labeled (to be used as the training data) and a Python implementation of Deep Learning Multiclass Classification Model was used to label the rest of the data.

The classifier was implemented in a local Jupyter Notebook, using the Keras library for Python.

The Deep Learning Network uses one LSTM layer with one Feed Forward Dense Layer, with Softmax activation function to perform the assignment of each observation to one of the pre-defined label categories.

Please note that a comprehensive explanation of Deep Learning methods and terminology is too broad a subject and not in scope for this project report. However, several resources on both Deep Learning and Text Classification were used in this project. A list of resources and helpful articles is provided in Appendix A – Text Classification and Deep Learning Resources.

The training data was labeled manually and loosely based on the topics identified in LDA. The output of the LDA analysis is already broken down into sentences. The dataset was run through additional text cleanup and word embeddings were created in Keras.

Based on the guidelines in the articles mentioned earlier, the model has the following basic layers:

#### Text Classifier Training Set

1. Identify categories based on the topics identified by LDA algorithm.

Topics are ranked from most to least interesting: Precautions is the most interesting topic, while all OTHER messages are the least interesting.

Figure 7: Ranked Topics for Text Categorization

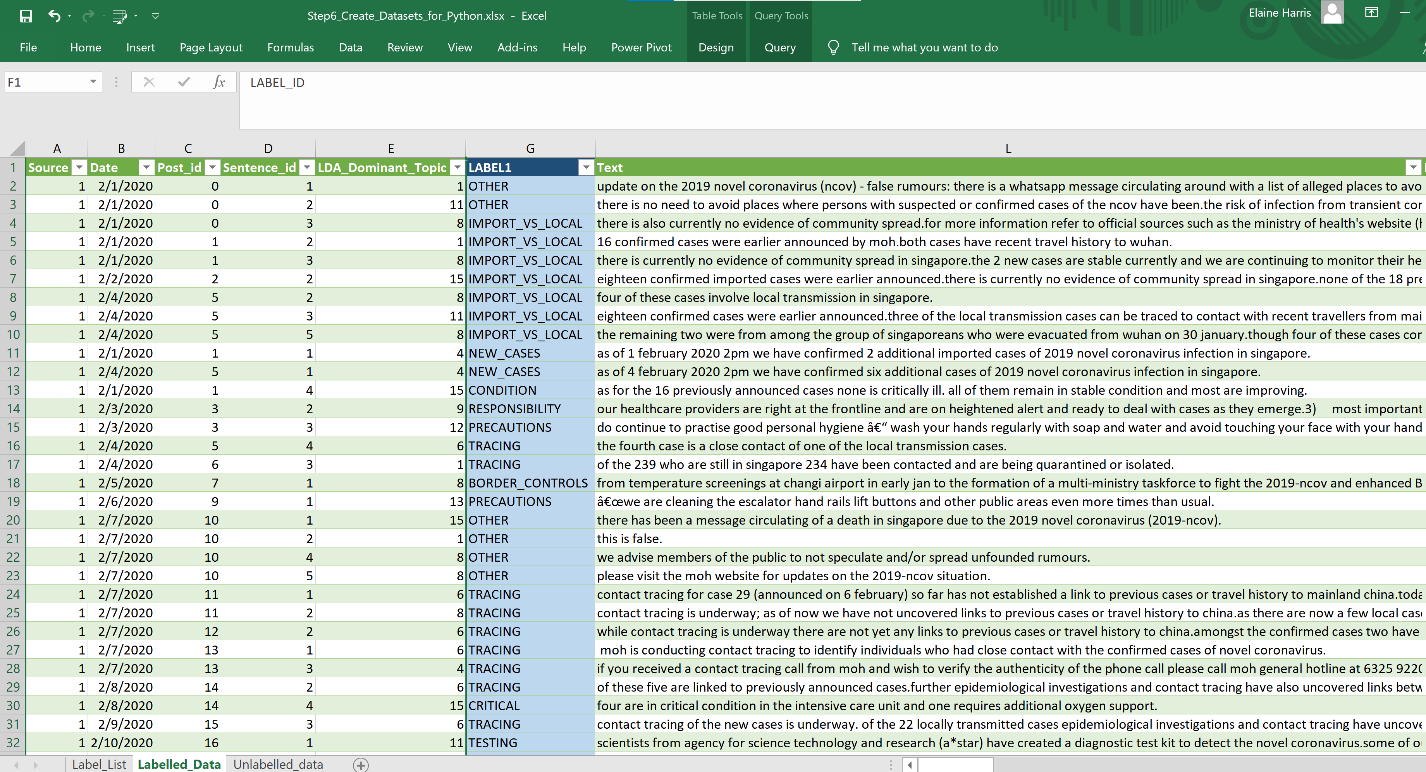
2) Label several posts, leaving other similar posts un-labeled. Split the dataset into Labeled and Unlabeled. The labeled dataset will be used to train the model, and the unlabeled dataset will be fed to the model to classify remaining posts. Split is roughly 30% training, 70% unlabeled. Datasets are saved to file Step6\_Create\_Datasets\_for\_Python.xlsx. Examples of Labelled and Unlabeled datasets below:

Figure 8:Examples of labeled and unlabeled datasets

#### Text Classifier Output and Corrections

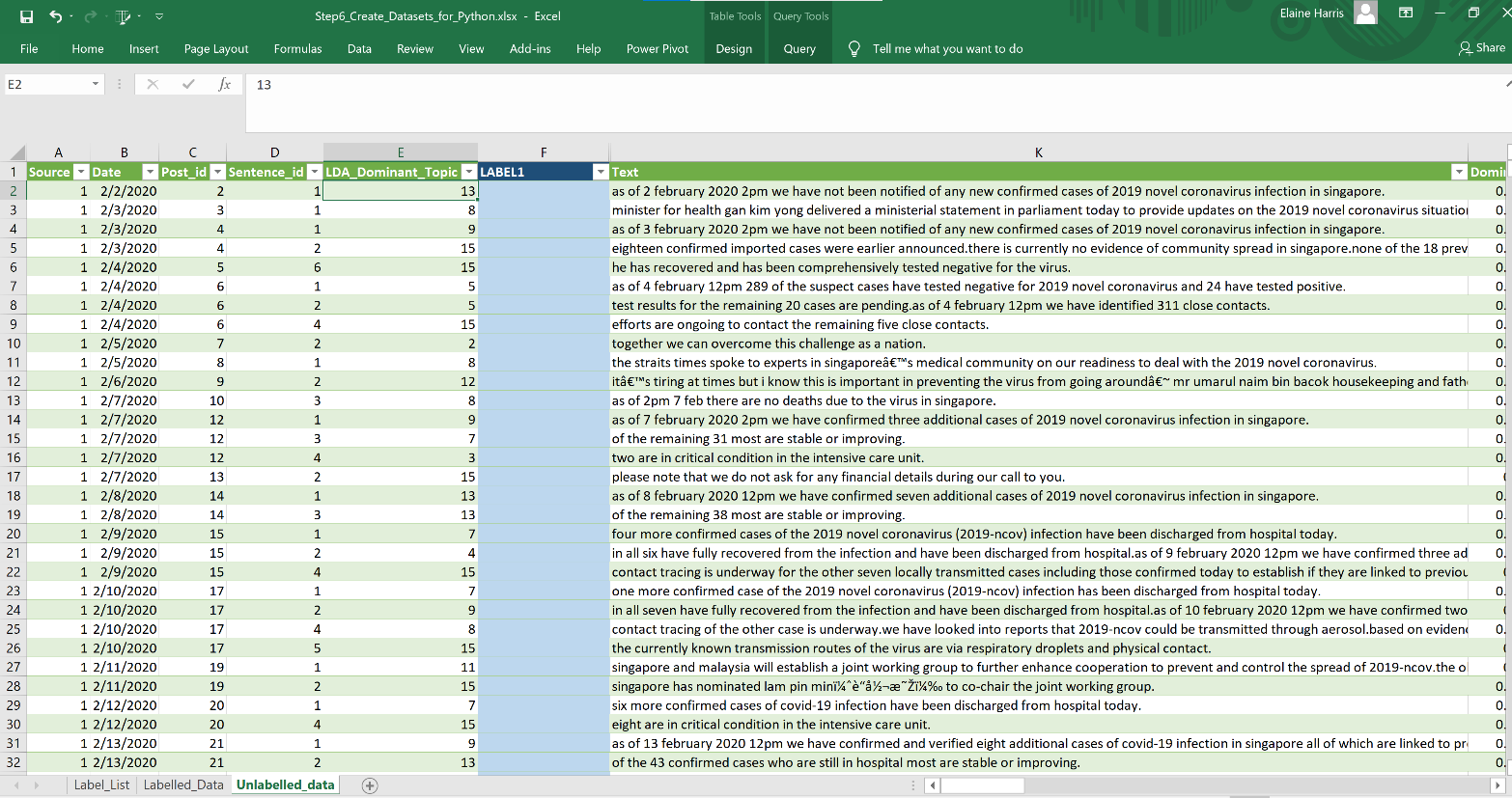
Out Text Classifier was implemented in Python, using the Keras Library for Deep Learning. It is worth noticing that the classifications produced by the algorithm are not 100% accurate. For this project, we visually examined the predicted classes and manually corrected the worst predicted results.

Figure 9: Un-labeled Dataset for Text Classification

Figure 10: Keras imports for Text Classifier

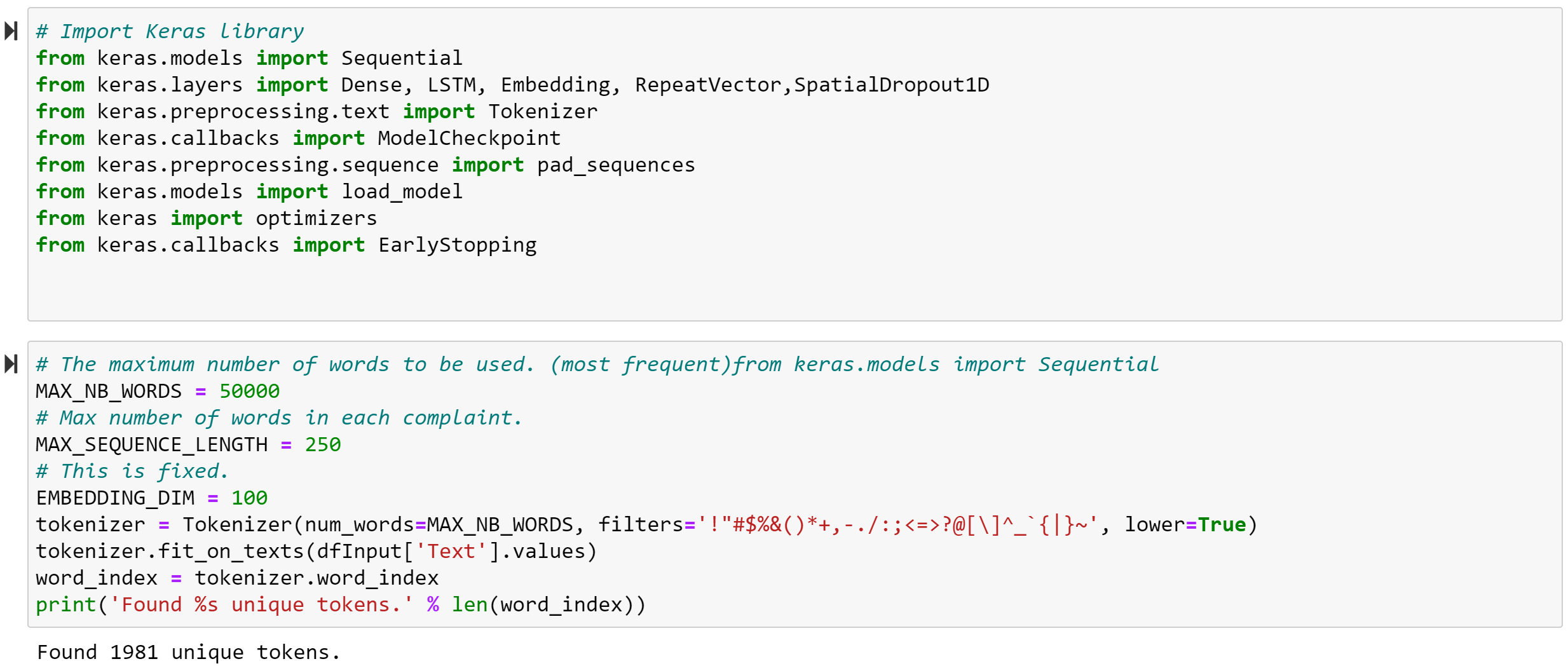


Figure 11: Text Classifier Model Training



Resulting predictions were exported to CSV format and incorporated into the project analysis workflow using Excel. The worst predictions were manually corrected, as exemplified below:

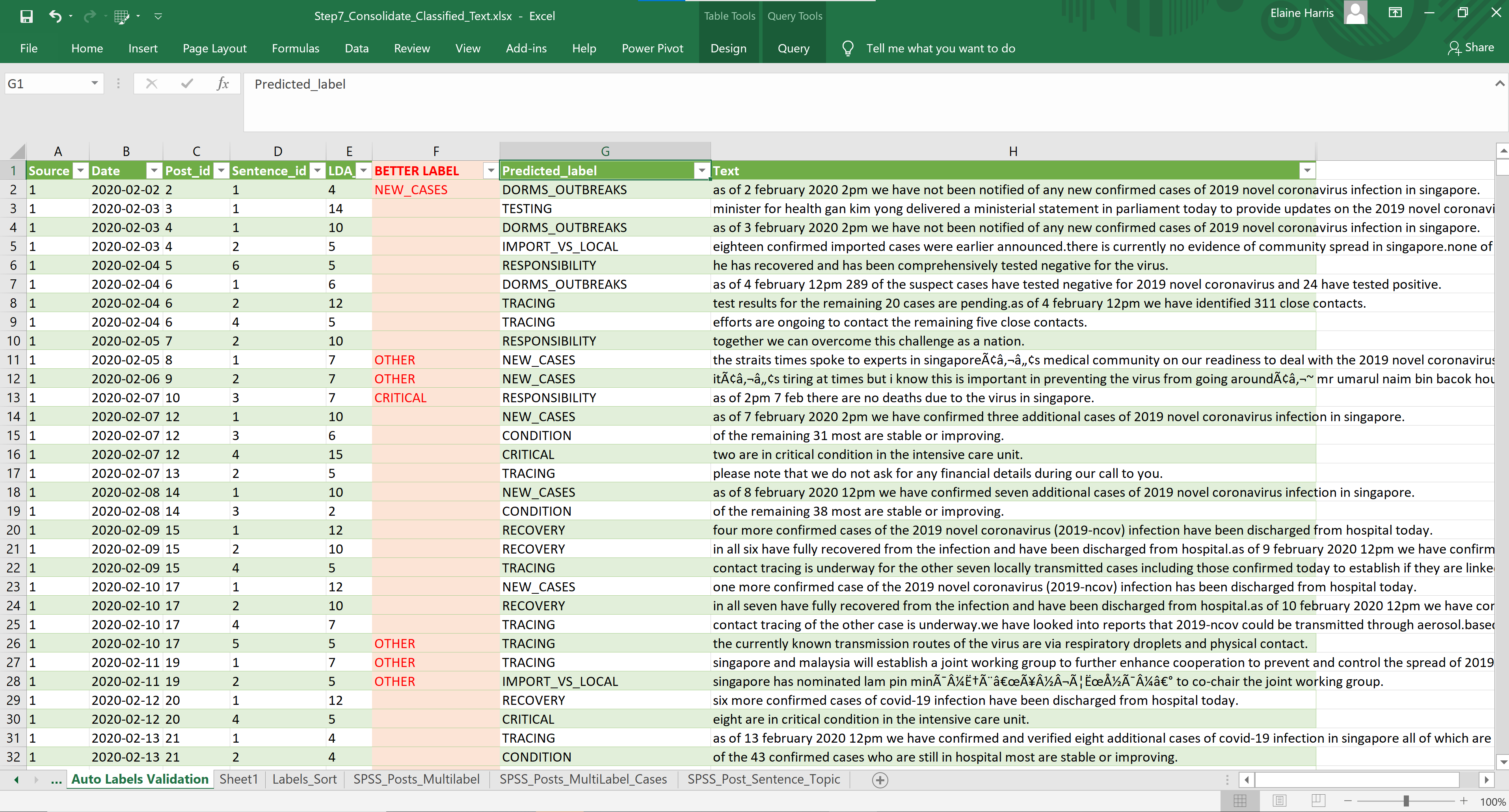


Figure 12: Manual corrections to Text Classifier Output

#### Final Text Classification by Post

|  |  |
| --- | --- |
| **Topic Id** | **Topic Sort** |
| **PRECAUTIONS** | **1** |
| **BORDER\_CONTROLS** | **2** |
| **SHN** | **3** |
| **DORMS\_OUTBREAKS** | **4** |
| **IMPORT\_VS\_LOCAL** | **5** |
| **LIFT\_RESTRICTIONS** | **6** |
| **CIRCUIT** | **7** |
| **RESPONSIBILITY** | **8** |
| **TESTING** | **9** |
| **CRITICAL** | **10** |
| **NEW\_CASES** | **11** |
| **CONDITION** | **12** |
| **RECOVERY** | **13** |
| **TRACING** | **14** |
| **OTHER** | **15** |

Table 3: Text Categories Ranking

Each post may have several sentences. The Text Classifier assigned a label to each sentence. In order to simplify a categorical analysis, we assigned a single Dominant Topic to each post based in the fixed ranking on the labels of each of its sentences. For instance, if the same topic contains a sentence labeled as “Precautions” and another one labelled as “Recovery”, the dominant topic for the entire post is “Precautions”. This is done so that analysis of Shares and Likes by Topic produce a closer representation of the reason why the post received feedback. Otherwise, we would credit feedback to less interesting topics (such as New Cases or Condition, which are very common), on account of them appearing in the same post with a more interesting topic such as *Precautions* or *Border Control*.

### Behavior Keyword Identification

Natural Language analysis is a tricky subject, because so many different topics may be implied in the same sentence or paragraph. In out dataset, we find that messages related to wearing masks, washing hands, keeping safe distance, etc, may appear in several different broader topics, though they may tend to appear more under Precautions and Border Controls. In order to capture as much of their frequency as possible, we decoupled Behavior Frequency from the post classification, making it possible to account for each behavior regardless of the broader context of the message. Behavior-based messages are identified by keyword match and accounted as One Hot Encoding (1 = behavior present, 0 = not present in the message).

The following behaviors were identified by visual observation:

Table 4: Behavior and Keyword Match Pattern

|  |  |
| --- | --- |
| Theme | Keywords |
| Hygiene | hand, wash, soap, hygiene |
| Masks | mask, shield, mouth, nose |
| Stay Home | stay home, stay at home, go out, leave home |
| Elderly | elderly, seniors, grand-parent, grandparent,nursing, senior-centric, elder |
| Work Remote | student, workplace, office, employer, telecommut, remote, from home, interact, teleconsult, socialise |
| Avoid Crowds | crowd, gathering, centre, outdoor, social contact, socialise |
| Activity caps | safe-entry, total number, maximum, visitors per day |

### Statistical Modeling

IBM SPSS Software was used to analyze the distributions of Likes and Shares by post Topic (category) and to calculate the Pearson Correlation between Topic Frequency and New Infection Trends and Behavior Frequency and New Infection Trends.

Output from SPSS is provided in the Findings section of this report, along with graphs for easy visualization.

The SPSS files and saved outputs are included in the Project Solution Package.

### Time Series Analysis

As previously explained, Regression and Forecast Models are out of scope in this project due to their extremely complex domain. Instead, we performed a Time Series Analysis on New Infections Trend attempting to provide explanations to drastic changes based on the time elapsed since different government mandates were enacted: Border Controls and Travel Advisories, start and end of restrictions under the Circuit Breaker period between April 3 and June 1st, self-isolation requirements (Stay-Home Notice) and increased testing and quarantine of Foreign Worker Dormitories.

Results are provided in the Findings section of this report, in the form of a Timeline plot annotated with the dates of the enactment of main government mandates, as identified in the Ministry of Health Facebook feed.

# Findings

This project attempts to answer the following questions:

1) Is Social Media an effective way to communicate mitigations strategies?

2) Is there a correlation between public health messaging and confirmed positivity rates?

3) Are there daily downward trends in positivity rates when mitigation mandates are issued?

We utilized statistical analysis on the datasets provided or required by the project sponsor, Dr Tian, and augmented the context with other sources that are also publicly available. Then project findings are organized as answers to the questions proposed above.

## Is Social Media an effective way to communicate mitigations strategies?

### Social Media Engagement Statistics

We looked at the statistics for Facebook and Twitter published by *Hootsuite.com* (https://hootsuite.com/pages/digital-in-2019#accordion-148291)

Facebook does not provide statistics on how many people actually *read* a post. Marketers use approximations based on the number of friends (or followers) an account has.

HootSuite.com places user engagement on Facebook in Singapore posts at **4.03%.**

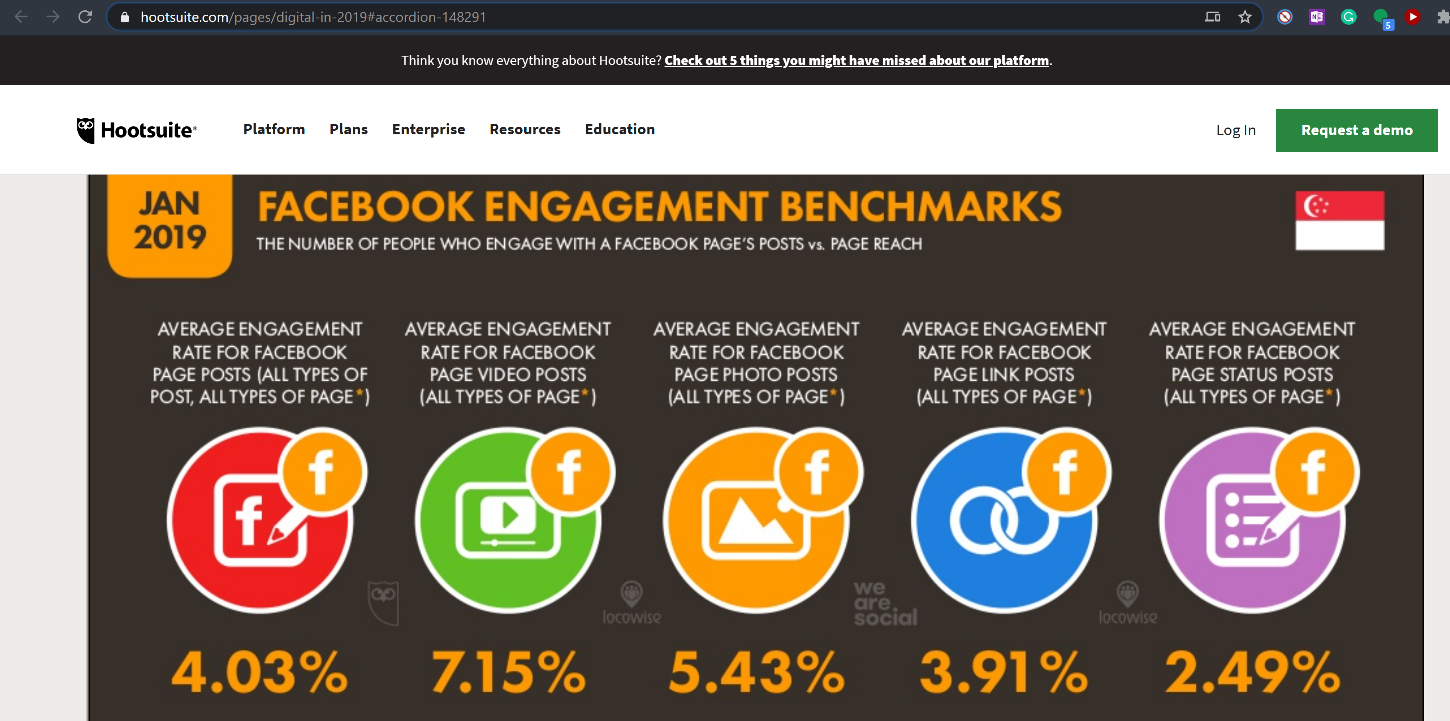


Figure 13: Hootsuite Facebook Engagement Benchmarks for Singapore in as of January 2019

The same report publishes the totals reachable audience by Social Media site, based on the monthly number of active users.

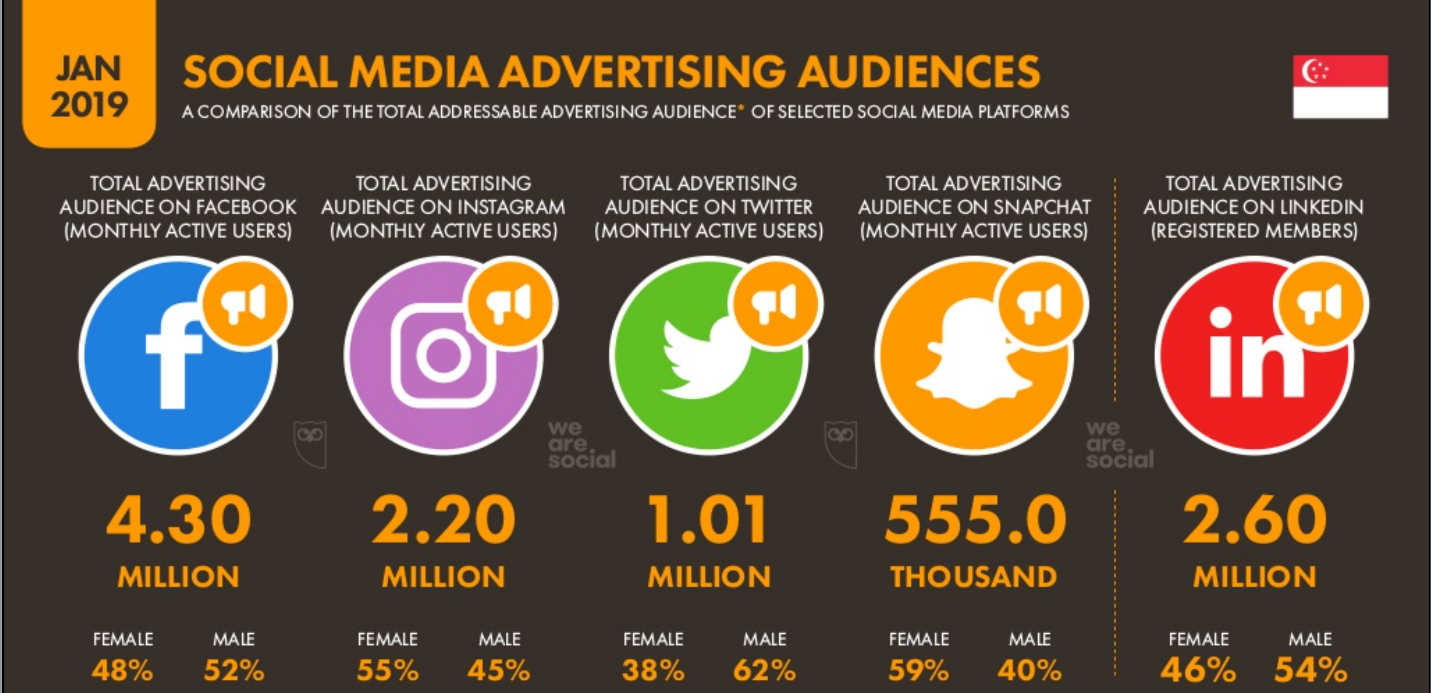


Figure 14 Hootsuite Social Media Reachable Audience Benchmarks for Singapore as of January 2019

The Ministry of Health Facebook page had 386,733 followers on Nov 8, 2020 (https://www.facebook.com/sghealthministry)

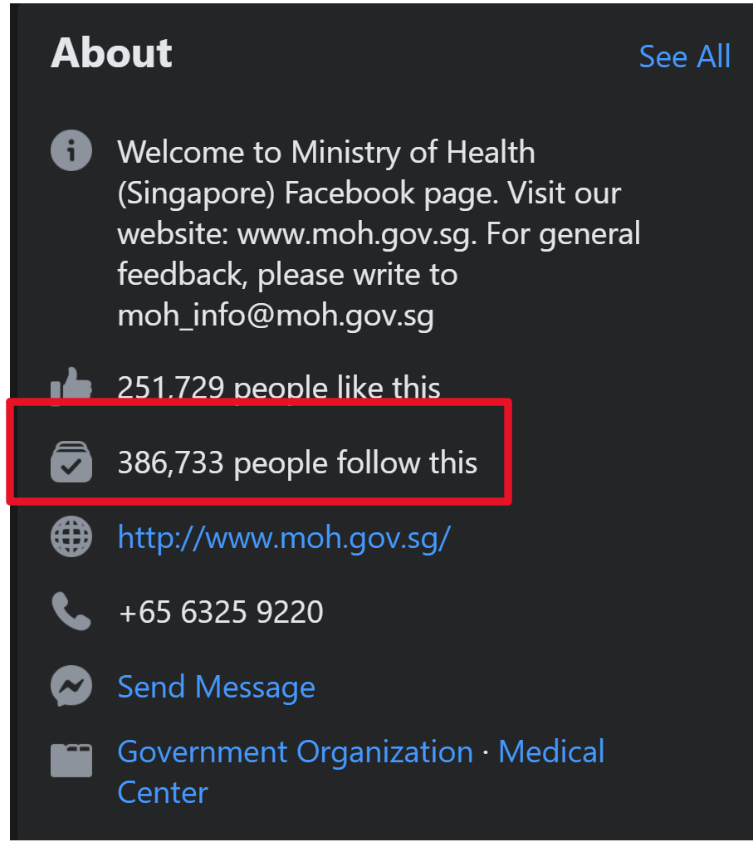


Figure 15 MOH Facebook page followers as of Nov 8, 2020 - Source: https://www.facebook.com/sghealthministry

The population of Singapore is about 5,850,342 people (Source: https://www.singstat.gov.sg/find-data/search-by-theme/households/households/latest-data)

Engagement based on total Facebook market: 4.03% x 4.3 million = 173k = **2.96%** of population

Engagement based on MOH Facebook page followers: 4.03% x 386.7K = 15.5k = **0.27%** of population



Figure 16 Facebook Potential User Engagement vs Population

*Hootsuite.com* does not provide engagement numbers for Twitter users. It does provide a market of 1.01 million Twitter users, compared to 4.3 million users on Facebook.

### User Engagement by Topic

Most Facebook posts (57%) focuses on recapping the number of daily infections, mostly on foreign worker dormitories. This topic ranks 2nd on user Share preferences.

While keeping the population informed on the containment measures in dense housing, these messages have little educational value.

Posts related to education and behavior management (Safe Distancing, Masks, Hygiene, etc), rank 2nd in the posts frequency rates, at a low 14%. It ranks 3rd in the users Shares preferences.

While users do take notice of such educational posts, their presence is drowned by to prevalence of posts that recap the number of daily infections.

### Average User Response by Topic - Facebook

We analyzed which topics are most favored by the government officials (in terms of frequency by month) and by the social media users (in terms of average Likes and Shares). Frequency tables created in Excel:

Figure 17: Facebook Topic Frequency, from highest to lowest post frequency count



Figure 18:Facebook User Response Averages by Topic, from highest to lowest



The **most frequent topic** for Facebook is *Dorms Outbreaks*, which is also the second most shared topic.

The average number of Likes is consistently higher than the average number of Shares, at a 14:1 ratio

Out of the top 5 topics by Average Shares, 3 match the Top 5 topics by frequency post: *Dorms Outbreaks*, *Precautions* and *Imported vs Local Cases*.

The top shared topic was ***Tracing***, which has a low frequency of 4 posts.

#### Shares by Topic - What do users find worthy sharing with others?

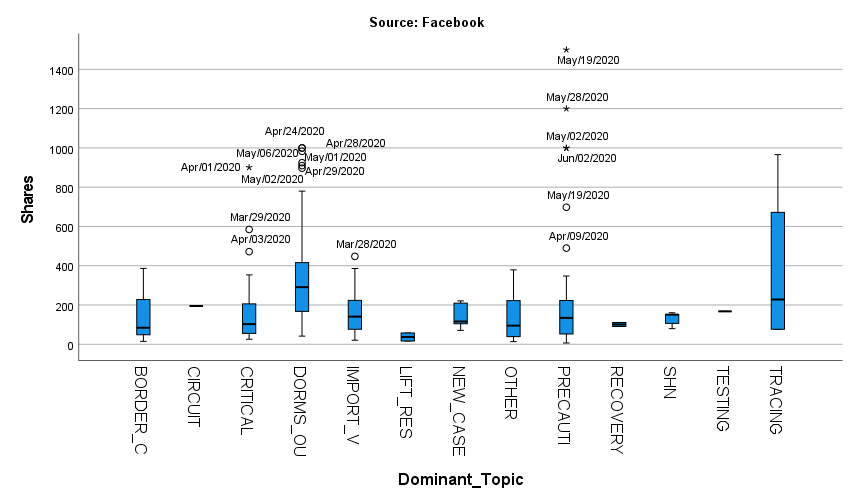
Facebook peaks public interest on 05/19 with 1,500 Shares: Announcement of end of Circuit Breaker restrictions. Topic: *Precautions*

The second most Shared post in from 05/28, at 1,200 Shares, and it is also about the end of Circuit Breaker: Singapore will be exiting the circuit breaker and embarking on phase one of re-opening on 2 June 2020. we expect more than three quarters of the economy to resume operations by then.

**Facebook users interest stems from imminent changes to daily life restrictions, and the constant follow up on source of new infections (outbreaks in foreign worker dormitories and tracing to known clusters or imported cases)**

Based on Median, the most shared topics are *Dorms Outbreak* (291), *Tracing* (228) and *Precautions* (134).

Figure 19: Facebook Shares by Topic Boxplot



#### Likes by Topic - What do users find worthy showing any reaction at all?

Facebook gets the most Likes on 05/28 at 4,800 Likes, topic Precautions: post confirms exit of circuit Breaker starting on June 1.

On 05/24, also at 4,800 likes: we may be celebrating the hari raya festivities a little differently this year with the covid-19 pandemic but it remains just as meaningful.

**Facebook users interest stems from imminent changes to daily life restrictions.**

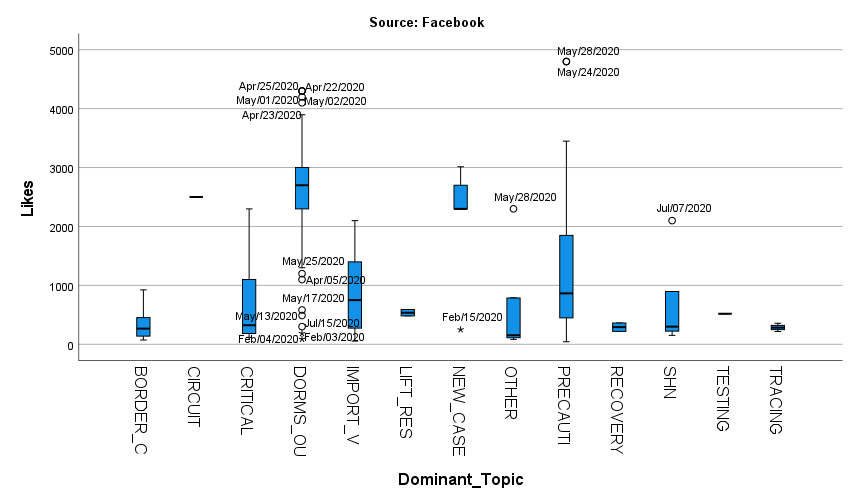


Figure 20: Facebook Likes by Topic Boxplot

### Average User Response by Topic – Twitter



Figure 21: Twitter Posts by Topic, from highest to lowest frequency



Figure 22: Twitter Average User Response by Topic, from highest to lowest average of Shares

Twitter posts are more granular, with a more even distribution of the top 5 topics.

#### Shares by Topic - What do users find worthy sharing with others?

Twitter peaks public interest on 04/18 with 125 Shares, topic **Dorms Outbreak**.

Text: as of 18 Apr 2020 12pm we have preliminarily confirmed an additional 942 cases of covid-19 infection in Singapore the vast majority of whom are work permit holders residing in foreign worker dormitories.

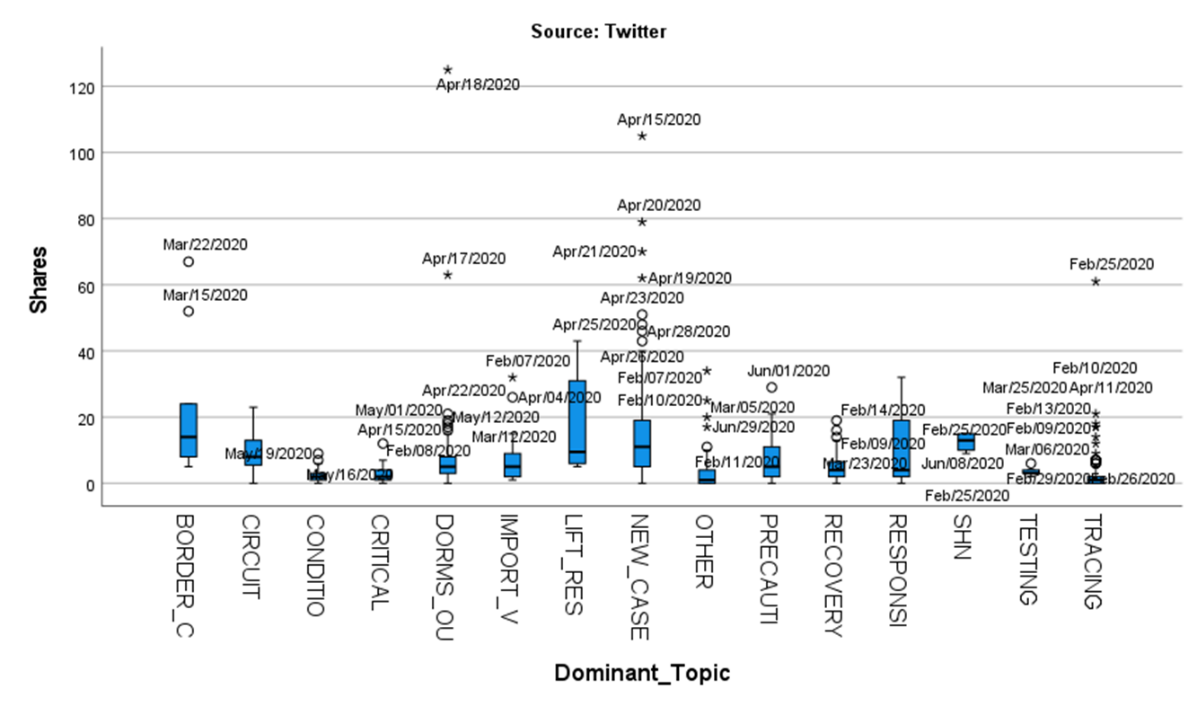
Three days earlier (04/15) is the second Twitter peak at 105 Shares, under topic **New Cases**.

Text: as of 15 Apr 2020 12pm we have confirmed and verified an additional 447 cases of covid-19 infection in Singapore.

**Twitter user interest appears to stem from the doubling of the number of confirmed cases over a period of 3 days.**

The median number of shares in Twitter is too low to warrant further analysis.

Figure 23:Twitter Shares Boxplot



#### Likes by Topic - What do users find worthy showing any reaction at all?

Twitter Likes peak on the same days as Twitter Shares: 04/15 and 04/18.

Twitter peaks public interest on 04/18, at 141 Likes, topic Dorms Outbreak: as of 18 Apr 2020 12pm we have preliminarily confirmed an additional 942 cases of covid-19 infection in Singapore the vast majority of whom are work permit holders residing in foreign worker dormitories.

Three days earlier (04/15) is the second Twitter peak, at 137 Likes, under topic New Cases: as of 15 Apr 2020 12pm we have confirmed and verified an additional 447 cases of covid-19 infection in Singapore.

**Twitter user interest appears to stem from the doubling of the number of confirmed cases over a period of 3 days.**

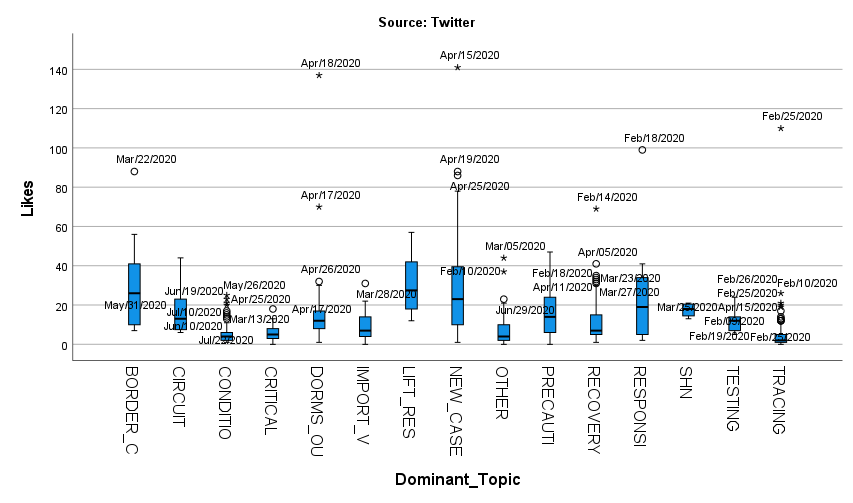


Figure 24:Twitter Likes by Topic

## Is there a correlation between public health messaging and confirmed positivity rates?

The objective of this question is to find a statistical relationship between the government officials choice of public messaging content and a (hopefully) downtrend in new infections within 14 days.

We have executed two Pearson Correlation analysis based on information extracted from the posts: *Topic Frequency vs Infection Trends* and *Behavior Frequency vs Infection Trends*.

### Topic Frequency vs Infections Trends

Let’s first recall the list of topics identified in the dataset:



Figure 25: Post Topics Identified in the Social Media Dataset

#### Methodology - Investigating Relationships Between Post Topics and Infection Trends

Facebook and Twitter Posts have been categorized by Topic. Case trends are a time series.

We will use two metrics for Case Trends: ***Daily Average of New Cases*** and the ***Average of Cases after 14 Days.*** The second metric is the actual number of infections 14 days after the date of the post, not a forecasted value. Both the daily cases and the 14-day offset are then averaged *by month*.

We also calculate the frequency of each Topic in the posts as the Percentage of each topic to the Total Posts for each month.

We then calculate the *Pearson Correlation* between the Cases and the Topic frequency rates by month in IBM SPSS. The Pearson R indicates the existence of a possible relationship between the variables, but not a cause-effect direction.

We will only note relationships where the absolute correlation is >= 0.70. We don’t expect SSPS to flag any correlations as significant, due to the small sample size (6 months/observations).

#### A word on Regression Modeling

We will not run any Regression Modeling or Forecasting Models for this dataset.

Forecasting infection rates is a complex subject in the domain of Epidemiology and relies on specialized multi-variable predictive models outside the scope of this project.

For our purposes, we will look at simple correlations and how government officials used social media to educate the public, and whether or not their choices correlate to the infection trends in the subsequent 14 days.

#### Correlation Analysis - Avg Daily Cases vs Topic Percentages

The high number of Twitter posts dilutes the topic frequencies across both platforms.

Correlation Analysis was executed for Facebook posts only. Twitter Posts are not as detailed and have negligible response rates.



Figure 26: Input variables for Topics vs Cases Correlation Analysis

SPSS software did not find any strong correlations (>= 0.70) between any of the topics and the average number of infections after 14 days.

The software found two strong correlations between *Avg Daily Cases* and *Precautions* (R = 0.712) and *Avg Daily Cases* and *Lifting Restrictions* (0.70). Scatter plots are provided for illustration.

**These correlations most likely indicate that the MOH officials increase the frequency of *Precautions* messages as the cases are climbing, and decrease them as infections rates drop.**

Figure 27: Topics vs Cases Correlation - SPSS Output



Figure 28: Avg Daily Cases vs Precautions Topic Post Frequency

Figure 29: Avg Daily Cases vs Lift Restrictions Topic Post Frequency

The graphic below plots the topic frequency every month vs the new infection trends.

* Most emphasized topic is *Dorm Outbreaks*. Despite the mathematical correlation, *Precautions* and *Lift Restrictions* frequencies are too low to draw any conclusions.

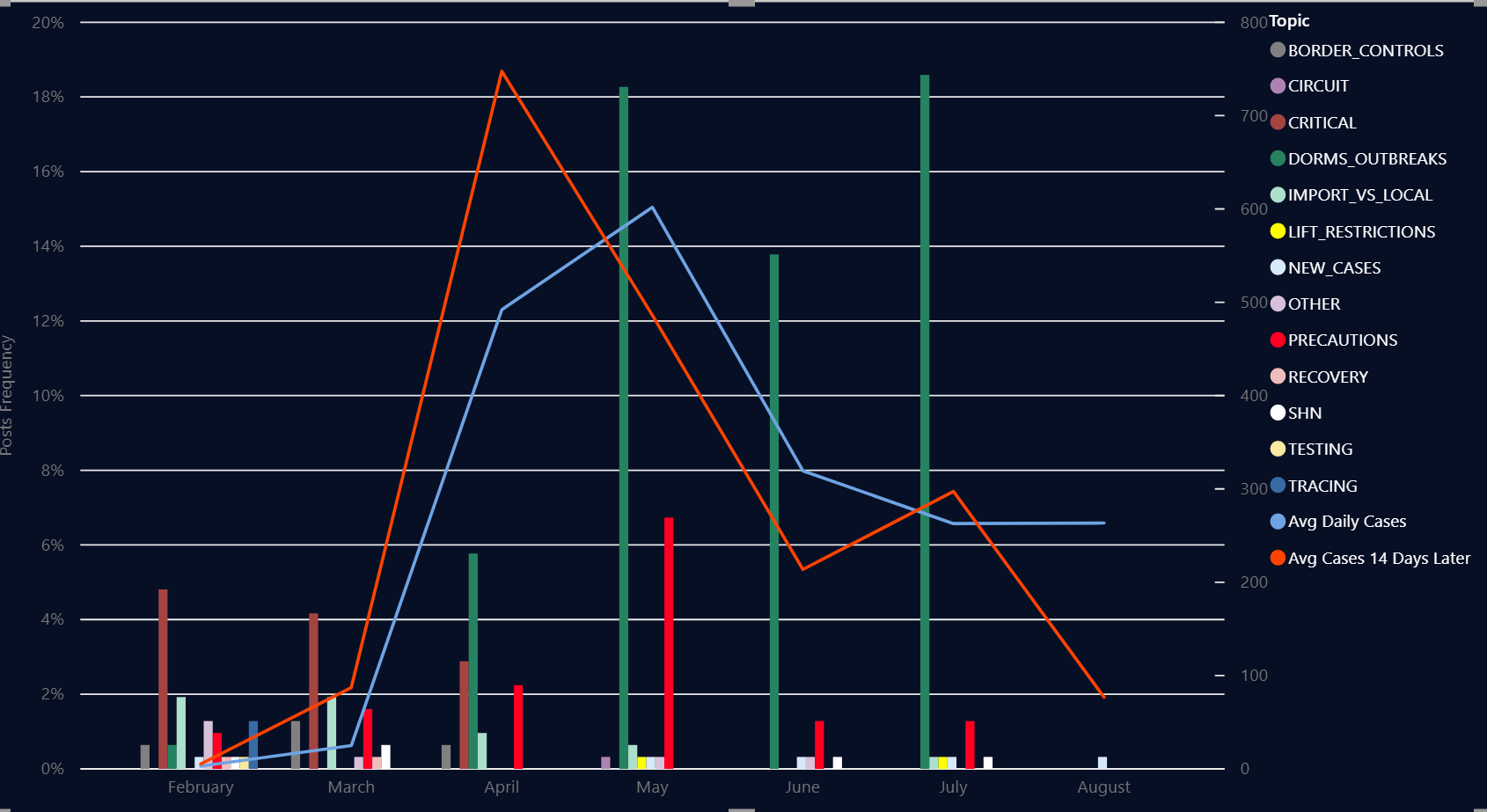


Figure 30 Topic Frequency vs Infection Trends – Facebook Posts

### Behavior Frequency vs Infections Trends

Under the *Precautions* topic, there are multiple behaviors that are either encouraged or discouraged in the social media posts. These behaviors were identified by keyword match on all posts.

We then calculate the percentage of posts containing references to each behavior. We use SPSS to calculate the Pearson correlation between each behavior and the Daily New Cases Average and the Average of New Cases 14 days after each post. These are the same measures described for the Topic Correlation Analysis.



Figure 31: Behaviors identified in the posts

#### A word on behavior changes

Based on the datasets used in this project, it is not possible to evaluate if individuals have actually altered their behavior in response to educational messages issued by the MOH.

For instance, we don’t have any data on how many people stopped leaving their homes or started wearing masks outside. Our only indication of behavior changes is the number of new infections.

However, it’s worth noticing that changes in the new infection trends are due to many outside factors, ranging from increases in testing to closing of international borders.

All those factors may be documented on social media, but the social media content should not be viewed as a cause for changes originated by government mandates.

#### Correlation Analysis - Avg Daily Cases vs Behavior Messaging Frequency

We will Facebook data for frequency and correlation analysis. There is a much larger number of Twitter posts, because of the 250-character message size, and the higher number of posts dilutes the rate calculations.

Even though the MOH post to social media every day, emphasis on ***Hygiene*** and ***Masks*** usage is relatively low, ranging from 2.5% to 9.52%.

The MOH places greater emphasis on ***Tele-commuting*** and ***Avoiding Crowds***. Both messages peak in March (Telecommute 18.75%, Avoid Crowds 21.88%). ***Travel*** related messages also peak in March, and decrease later, as Travel Advisories are already in place.

Despite the emphasis on educational messages in March, cases still climb in April, following a worldwide trend.



Figure 32: Input variables for Behavior Frequency vs Cases Trend Analysis



Figure 33: Behaviors vs Cases Correlation - SPSS Output

There is one strong correlation (>= 0.70) between ***Average Daily Cases*** and ***Masks*** messages (R = 0.707).

This correlation most likely indicates that the government officials reinforce the usage of masks as cases are increasing, but there is no strong correlation between it and the number of cases 14 Days later (R = 0.458)

There is one other correlation value that is almost at the 0.7 threshold: ***Visitor Caps*** and ***Daily New Cases*** (R = 0.693). However, these messages only exist for 2 months (May and June).

**These correlations most likely indicate that the government eases restrictions on certain activities as the cases are decreasing, but continues to maintain social distance rules.**

Figure 34: Average Daily Cases vs Masks Messaging Frequency

The next graph depicts the frequency of behavior-related messaging month by month, compared to the infection trends.

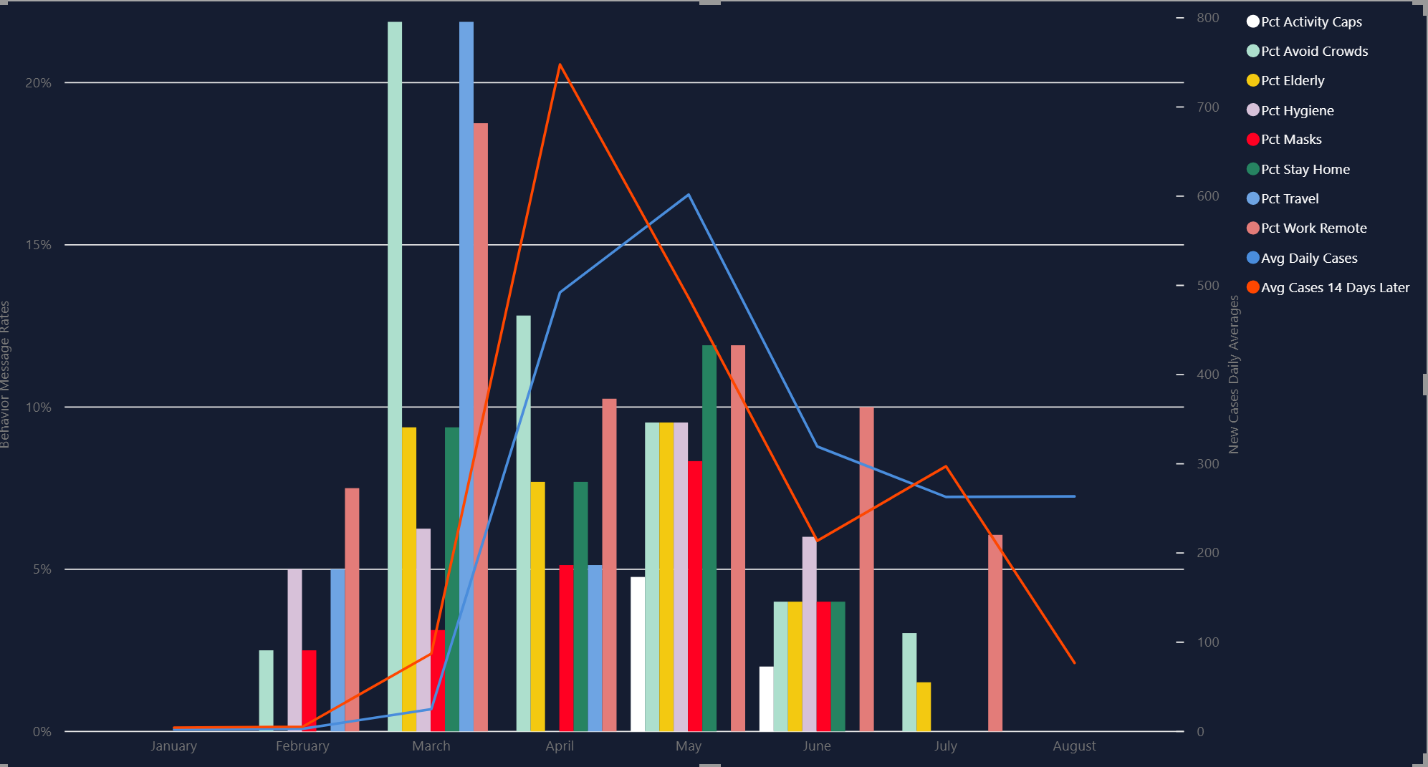


Figure 35 Behavior Message Frequency by Month vs Infection Trends - Facebook

## Are there daily downward trends in positivity rates when mitigation mandates are issued?

Correlation analysis indicate that content of social media messages is created likely as a reaction to the infection trends.

Here we analyze the timeline of announcements of Government-mandated restrictions and how it relates to the New Infections trends.

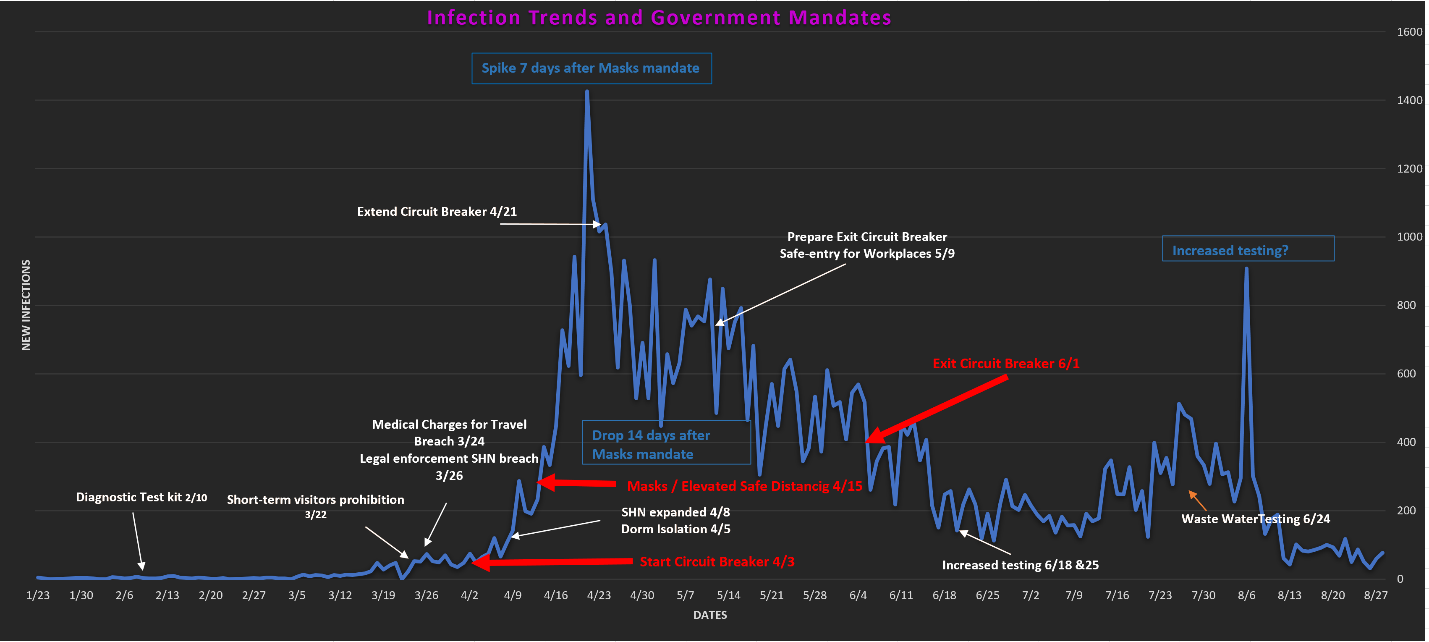


Figure 36: Infections Trends vs Government Mandates

### Stop importation of new cases - Border Control measures

During the month of March, case numbers are still low, but multiple Border Control measures are put in place, including the prohibition of entry or transit to short term visitors

from China, ASEAN countries, Italy and UK.

Retuning residents must self-isolate in quarantine centers or at home (Stay-Home Notice). Authorities are given Legal Enforcement powers in 03/26.

On March 24th the Government enacts a mandate that anyone who gets sick following unauthorized travel will be responsible for the medical bills.

In a similar fashion, on 03/26 a new mandate is put in place to allow legal enforcement of quarantine (Stay-Home Notice) breaches.

**The Travel Advisories and visitor entry restrictions were extremely effective. During the months of Apr and May the number of passenger arrivals and departures drops to below 1% of the annual average.**

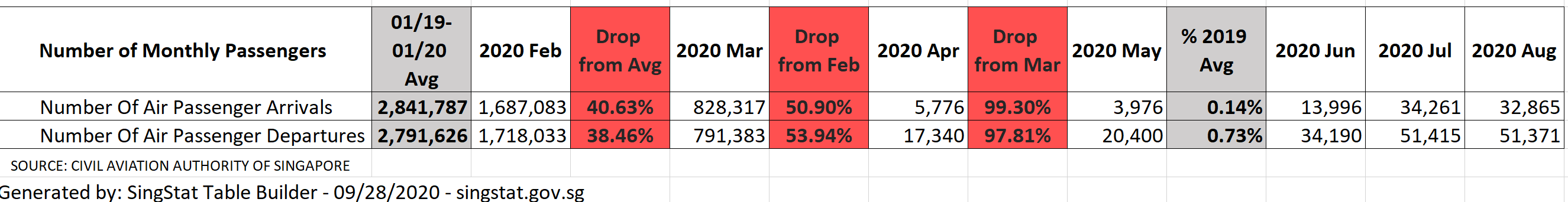


Figure 37: Drop on Arrivals and Departures from Singapore International Airport during Circuit Breaker

### Limit spread among elderly population

Senior-centric activities are cancelled early in March (03/10). Restrictions are partially lifted on 06/17, for individual senior activities.

Messages regarding extra caution for the elderly are as frequent as all other social distancing messages.

### Limit spread in the local population

After all Border Control measures are enacted in March, the infection numbers climb to a peak in mid-April. In addition to preventing importation of new cases, Singapore enacts the Circuit Breaker restrictions on 04/03. Elevated Safe Distancing measures, including a mandate of mask usage in public, is enacted on 04/15. These measures aim at limiting the spread of local infections.

For the duration of the Circuit Breaker, the MOH continues to emphasize the behaviors of ***Staying Home***, ***Working Remote*** and ***Avoiding Crowds***.

**We can notice a drop in new infections around 05/03, or approximately 2 weeks after the enactment of additional Safe Distancing measures.**

### Timeline of Behavior Messaging in Social Media

Travel Restrictions are emphasized in Feb and March, when cases are still low. Despite all announced border control and travel advisories, new infections spike in the following month (April).

Wearing masks and practicing good hygiene are mostly emphasized during the Circuit Breaker (April and May). However, the MOH focuses more messages on avoiding social contact.

***Avoid Crowds***, ***Work Remote*** and ***Stay Home*** are consistently the most frequent messages from march thru May. Cases still climb from March to April, but steadily decline from Apr to June.

Along with crowd avoidance, messages emphasize caution for the elderly, at a similar rate and distribution.

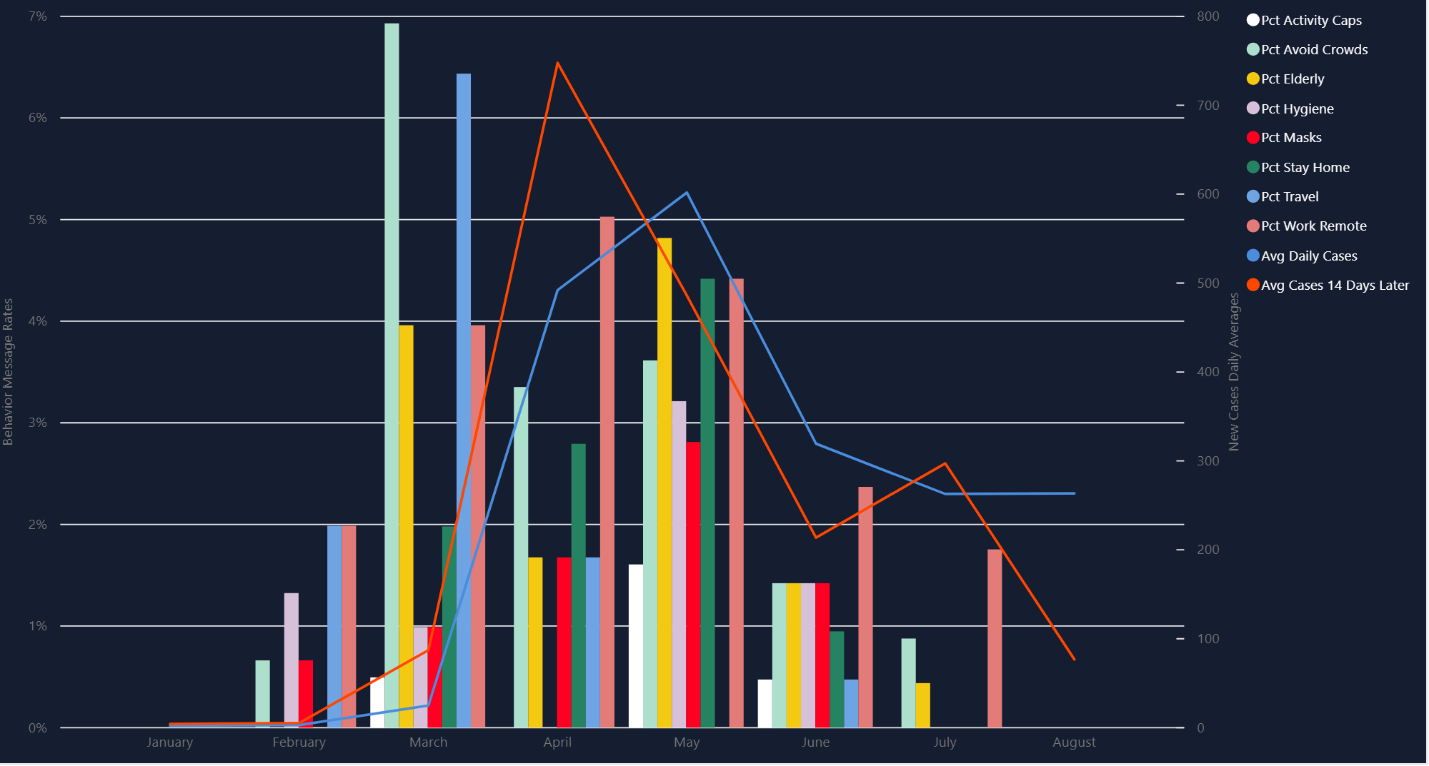


Figure 38: Behavior Message Frequency vs New Case Trends Plot – All Social Media Channels

# Conclusion

This project was created to find statistical evidence that Government Health Officials choice of topics in Social Media can influence the public to adopt heathy behaviors that in turn could have a positive effect in the efforts to contain the spread of the COVID-19 virus in the population.

While our analysis did not find any conclusive statistical proof that the frequency certain topics are emphasized in social media has a direct effect in the number of new cases within 14 days, we have determined that the timing of the enactment of certain restrictions seems to explain the abrupt drop in new infections following the steep climb in the month of April.

Quick action to limit the importation of new cases and the enactment of restrictions in the workplace combine to contain the spread from external sources and within the community. However, we don’t have any data to indicate if behavior changes originated mostly from adherence to educational messages (in social media or elsewhere) or simply due to the existence of government-mandated and legally enforced restrictions.

## Project Summary

This project was successful in its goal to provide an opportunity for students to learn how to analyze a real-life scenario based on measurable evidence, rather than instinct and individual or collective bias.

From a technology perspective, the Text Analytics component of the analysis was an invaluable opportunity to learn and exercise concepts of Big Data Analytics, Deep Learning and Artificial Intelligence. We learned that while these technologies have a lot to offer, they are not a miracle solution for every problem, nor do they work without human intervention and professional analysis of the results.

## Limitations and Future Direction

Despite different levels of restrictions imposed by multiple governments worldwide, the virus continues to spread to this day, baffling experts and worrying government officials and citizens alike.

The greatest limitation of all contemporary studies of the social-economic effects of the COVID-19 Pandemic is the novel nature of the phenomenon. While pandemics of this nature have happened before (the Spanish Influenza of 1918, the Hong Kong Flu in 1968, the Swine Flu in 2009, the SARS outbreak in 2003), socio-economic factors make this pandemic unique: influence of traditional and social media on the general public’s perception of risk, the impact of total or partial lockdowns in the economy, the availability of Big Data tools to extract and study data from the internet, regardless of relevance. All these factors combine to create a sense of urgency around the understating of how the disease spreads; however, as of the date this report was written (Nov 2020), the pandemic is still not under control.

A specific limitation for this project was the inability to find any evidence of actual behavior change in response to government requests. We could not find any datasets that could measure the number of employees working remote vs onsite, for example. For future direction, this project’s analysis could be enhanced with economic indicators such as industrial output, school attendance, online shopping increases, etc., that could better measure behavior changes as inputs to a complete Epidemiological Model.

# References

Moreno, Ángeles; Fuentes-Lara, Cristina; Navarro, Cristina (2020). “Covid-19 communication management in Spain: Exploring the effect of information-seeking behavior and message reception in public’s evaluation”. El profesional de la información, v. 29, n. 4, e290402. <https://doi.org/10.3145/epi.2020.jul.02>.

Wei, WE, Li Z, Chiew CJ, Yong SE, Toh MP, Lee VJ. (2020). Presymptomatic Transmission of SARS-CoV-2-Singapore, January 23-March 16, 2020. *MMWR Morb Mortal Wkly Rep*, 69, 411-415. DOI: <http://dx.doi.org/10.15585/mmwr.mm6914el>

# Appendix A – Text Classification and Deep Learning Resources

## Introduction to Text Classification

Text Classification: Applications and Use Cases:

<https://towardsdatascience.com/text-classification-applications-and-use-cases-beab4bfe2e62>

A Comprehensive Guide to Understand and Implement Text Classification in Python:

<https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/>

Best Practices for Text Classification with Deep Learning:

<https://machinelearningmastery.com/best-practices-document-classification-deep-learning/>

Approaching (Almost) Any NLP Problem on Kaggle:

<https://www.kaggle.com/abhishek/approaching-almost-any-nlp-problem-on-kaggle>

## Introduction to Deep Learning and Word Embeddings

Contextual Word Representations: Putting Words into Computers:

<https://cacm.acm.org/magazines/2020/6/245162-contextual-word-representations/fulltext>

Beyond Word Embeddings Part 1. An Overview of Neural NLP Milestones:

<https://towardsdatascience.com/beyond-word-embeddings-part-1-an-overview-of-neural-nlp-milestones-82b97a47977f>

Beyond Word Embeddings Part 2. A primer in the neural NLP model:

<https://towardsdatascience.com/beyond-word-embeddings-part-2-word-vectors-nlp-modeling-from-bow-to-bert-4ebd4711d0ec>

Neural Networks From Scratch in Python & R:

<https://www.analyticsvidhya.com/blog/2020/07/neural-networks-from-scratch-in-python-and-r/>

Activation Functions in Neural Networks:

<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>

Recurrent Neural Network | Fundamentals Of Deep Learning:

<https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/>

Long Short Term Memory | Architecture Of LSTM:

<https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/>

Illustrated Guide to LSTM’s and GRU’s: A step by step explanation:

<https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>

Layers in a Neural Network explained:

<https://deeplizard.com/learn/video/FK77zZxaBoI>

# Appendix B – SPSS and Statistical Modeling Resources

The following resources were instrumental in performing statistical modeling on the data utilized in this project:

Daniel, T. (2017). How to do a Pearson Correlation in SPSS (13-8) [Video]. Retrieved from <https://www.youtube.com/watch?v=rR99bpl0rKM>

Daniel, T. (2017). How to do Simple Linear Regression by Hand (14-4) [Video]. Retrieved from <https://www.youtube.com/watch?v=YxvoByLcFfw>

Daniel, T. (2017). How to do Simple Linear Regression in SPSS (14-5) [Video]. Retrieved from <https://www.youtube.com/watch?v=6xcQYmPDqXs>

Daniel, T. (2019). 4:31 / 8:22 Four Ways to Compare Groups in SPSS and Build Your Data Handling Skills [Video]. Retrieved from <https://www.youtube.com/watch?v=Ju-M3gHHI2Y>

Grande, T. (2015). Creating and Interpreting Boxplots in SPSS [Video]. Retrieved from <https://www.youtube.com/watch?v=M58Ww51xhso>

# Appendix C – Additional Resources for Data Context Interpretation

The following resources were utilized to add context to the data sources used in this project:

**The Global State of Digital in 2019 Report**

Global and country-specific reports on utilization and reach of social media platforms. Used to frame the significance of the Likes and Shares statistics on the MOH social media sites.

<https://hootsuite.com/pages/digital-in-2019#accordion-148291>

**Statistics Singapore**

Official Singapore website providing all sorts of demographics and economics statistics. Used to assess the actual impact of Border Control Restrictions and Travel Advisory as measures by the variations in the number of arriving and department air passengers during February and July 2020.

<https://www.singstat.gov.sg/find-data>

# Acknowledgements

Our group would like to give special recognition to our project sponsor, Dr. Shirley Tian, for her guidance through this process. Additionally, we would like to recognize Dr. Meng Han, our advisor, for supporting our project throughout this semester.

# 