Singapore University of Technology and Design
Information Systems Technology and Design

50.038 Computational Data Science

Analysis of Covid-19 Sentiments on Social Media

Jisa Mariam Zachariah Fion Yao Yuechi Varsha Venkatesh
Xie Han Keong

December 14, 2020

Abstract

This report aims to find a correlation between Covid-19 sentiments on social media and its infection rate. We trained four different machine learning models on a publicly-available dataset of tweets with labelled sentiments, and found that the model which achieved the highest accuracy was the BERT with fine-tuning model. We used it to predict the sentiments of social media posts related to Covid-19 which we scraped from Twitter and Reddit. We applied various data visualisation tools to visualize our findings. We found that there was no significant correlation between Covid-19 sentiments and the infection rate. We provide several explanations why. Mainly, we think that it is because the nature of the sentiment of social media content is not linear as people use social media to both spread positive vibes and express their frustrations, whereas the Covid-19 infection rate has so far only seen an exponential growth in trend.

1 Introduction

When Covid-19 started in end-2019, no one thought that it would become the pandemic it is today. Over the months, as the number of cases increased exponentially, many people have taken to various social media platforms to express their thoughts and feelings about Covid-19. Some hot topics have been, to list a few, the number of cases in their home country, who contracted/recovered from/succumbed to the virus, and government policies and measures to curb the spread of Covid-19. Some of these people, unable to adjust to the new norms, have also refused to adhere to recommended practices such as social distancing and wearing face masks which reduce the chance of infection. In doing so, they are indirectly contributing to the Covid-19 infection rate as the chance of them being infected and/or passing on an infection is higher.

2 Problem

In our paper, we aim to investigate the correlation between the public's general sentiments about Covid-19 and the infection rate. Our hypothesis is that there is a negative correlation between the public's general sentiments toward Covid-19 and the infection rate. Several assumptions of our hypothesis are:

1. The comments and posts on social media are representative of the general public
2. The sentiment of the comments and posts generally reflects people's actions and their intentions

3 Dataset And Collection

We collected three datasets in total: two from Twitter and one from Reddit. One of the Twitter datasets was labelled, whereas the other Twitter dataset and the Reddit dataset were unlabelled. The unlabelled Twitter dataset and the Reddit dataset were created by us from scratch.

3.1 Twitter dataset (unlabelled)

We decided to use Twitter to test our hypothesis because it is a very popular social media platform where users post quick and instant thoughts about their lives. Using scrapehero.com, we managed to amass 8,600 tweets filtered by Covid-19 keywords (See Table 3.1), time, location, and likes. The tweets were dated from 1 January to 26 October 2020 and had a minimum of 200 likes in order to represent the popular opinion of the public. Figure 3.1 shows the top 5 keywords and Figure 3.2 shows the distribution of the tweet lengths. The median length of tweets is about 30 words, which is expected since Twitter imposes a limit of 280 characters on the text content of a tweet.

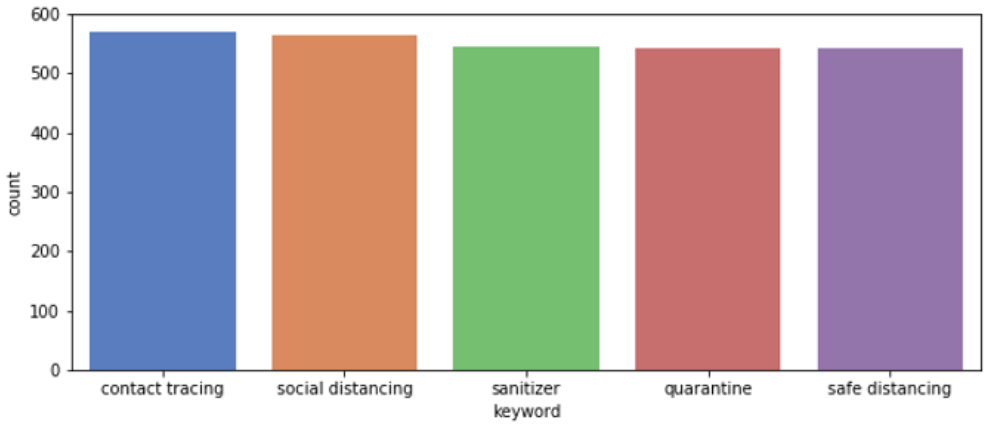

Figure 3.1: Number of tweets for top 5 keywords

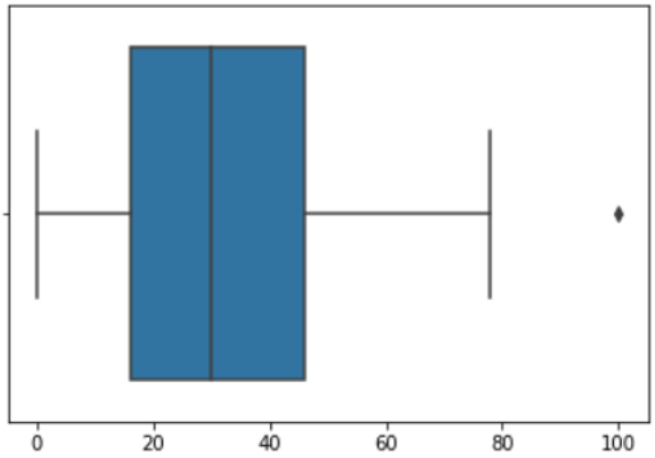

Figure 3.2: Distribution of tweet lengths

3.2 Reddit dataset

We also decided to use Reddit to test our hypothesis because it is another popular social media platform where many people discuss their opinions and thoughts about many different topics. We used pushshift.io to fetch data for all submissions in r/singapore from 1 January to 1 December 2020, amounting to over 51,000 submissions. We filtered 7,000 out using a list of Covid-19 keywords (See Table 3.2) and extracted all their comments which added to 150,000. Figure 3.3 shows the distribution of comment and submission lengths. The median length is about 20 words for comments and 0 for submissions. This is expected because Reddit submissions can also be in image, video, news article, or other formats which do not require any accompanying text. Figure 3.4 shows that the top 85% of the comments belong to 30% of the submissions. This indicates that the majority of comments came from a minority of submissions in the top page.

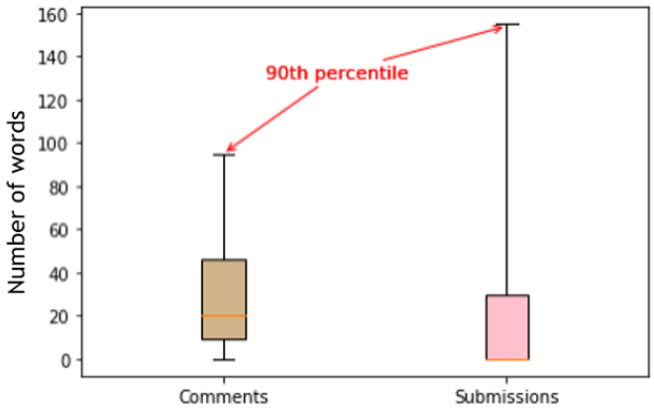

Figure 3.3: Distribution of comment and submission lengths

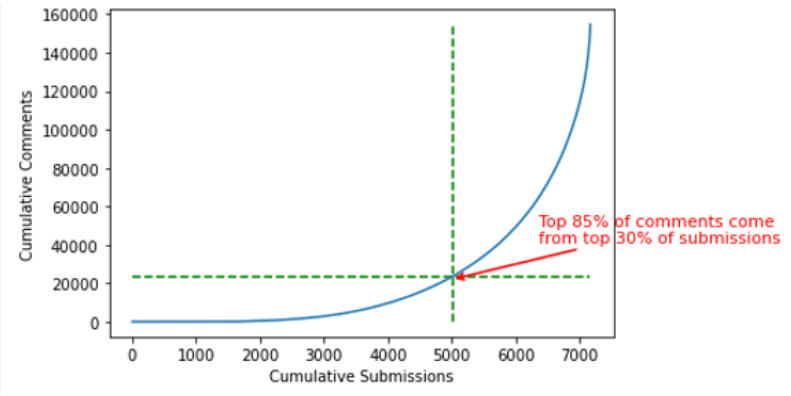

Figure 3.4: Distribution of comments across submissions

Table 3.1: Covid-19 keywords used to filter tweets

border closure	cb	check in
check out	circuit breaker	newnormal
opening up	pandemic	community cases
contact tracing	covid19	coronavirus
dorscon orange	mask	online classes
e-learning	pandemic	zoom
vaccine	quarantine	flights
hand sanitizer	imported cases	lockdown
safe distancing	safeentry	sanitizer
sgunited	social distancing	tracetogether
wfh	work from home	covid19
covidhoax	flights	freedom
stayhome	staymasked	temperature taking

Table 3.2: Covid-19 keywords used to filter submissions

cb	circuit breaker	corona
coronavirus	covid	covid19
dorscon	epidemic	ncov
outbreak	pandemic	pneumonia
safe distance	safe entry	safeentry
shn	social distance	stay home notice
trace together	tracetogether	virus
wfh	work home	wuhan

3.3 Twitter dataset (labelled)

The last dataset we obtained was a dataset from data.world containing 40,000 tweets with 13 sentiment labels. The purpose of acquiring a labelled dataset was to train our models. We selected this particular dataset because it contained five basic sentiments that we were looking for to represent Covid-19 sentiment: anger, happiness, neutral, sadness and surprise. We selected a sample of 1,000 tweets for each of the five sentiments to ensure that the classes were balanced (See Figure 3.5). Figure 3.6 shows that the median length of tweets is about 15 words, which is half that of our unlabelled Twitter dataset.

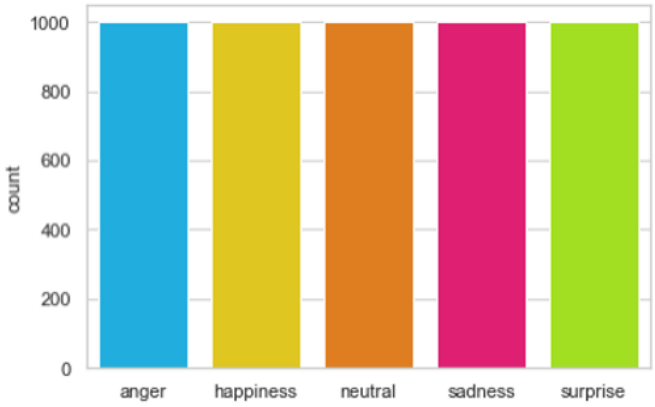

Figure 3.5: Class populations for the 5 sentiments

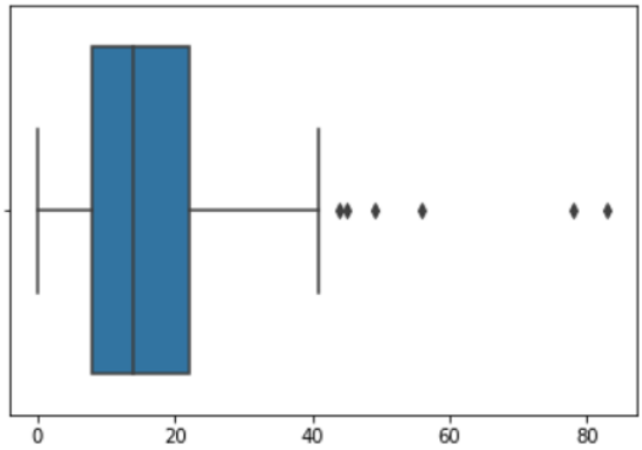

Figure 3.6: Distribution of tweet lengths (labelled)

4 Data Pre-processing

After all our datasets had been collected, we used regex to clean all the data in the following ways:

1. User mentions were removed (e.g. "@user123")
2. Hashtags were removed (e.g. "#covid" became just "covid")
3. Emoticons were removed (e.g. ":-)")
4. URLs were removed (e.g. "http://example.com")
5. Alphabets were converted to lowercase

An example of the data cleaning process is shown in Figures 4.1a and 4.1b. After the data was cleaned, we tokenized the words and removed stopwords. At this stage, the text was ready to be fed into our models for training or testing.

	tweet_id	sentiment	author	content
0	1956967341	anger	xoshayzers	@tiffanylue i know i was listenin to bad habi...
1	1956967666	anger	wannamama	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	anger	coolfunky	Funeral ceremony...gloomy friday...
3	1956967789	anger	czareaquino	wants to hang out with friends SOON!
4	1956968416	anger	xkilljoyx	@dannycastillo We want to trade with someone w...
5	1956968477	anger	xxxPEACHESxxx	Re-pinging @ghostridah14: why didn't you go to...
6	1956968487	anger	ShansBee	I should be sleep, but im not! thinking about ...
7	1956968636	anger	mcsleazy	Hmmm. http://www.djhero.com/ is down
8	1956969035	anger	nic0lepaula	@charviray Charlene my love. I miss you
9	1956969172	anger	Ingenue_Em	@kelcouch I'm sorry at least it's Friday?
10	1956969456	anger	feinyheiny	cant fall asleep
11	1956969531	anger	dudeitsmanda	Choked on her retainers
12	1956970047	anger	Danied32	Ugh! I have to beat this stupid song to get to...
13	1956970424	anger	Samm_xo	@BrodyJenner if u watch the hills in london u ...
14	1956970860	anger	okiepeanut93	Got the news
15	1956971077	anger	Sim_34	The storm is here and the electricity is gone
16	1956971170	anger	poppygallico	@annarosekerr agreed
17	1956971206	anger	brokenangel1982	So sleepy again and it's not even that late. I...
18	1956971473	anger	LCJ82	@PerezHilton lady gaga tweeted about not being...
19	1956971586	anger	cleepow	How are YOU convinced that I have always wante...

(a) Before data pre-processing

```
0          i know  i was listenin to bad habit earlier a...
1        Layin n bed with a headache  ughhhh...waitin o...
2                     Funeral ceremony...gloomy friday...
3                    wants to hang out with friends SOON!
4         We want to trade with someone who has Houston...
5        Re-pinging 14: why didn't you go to prom? BC m...
6        I should be sleep, but im not! thinking about ...
7                                                    Hmmm.
8                          Charlene my love. I miss you
9                     I'm sorry  at least it's Friday?
10                                      cant fall asleep
11                                Choked on her retainers
12       Ugh! I have to beat this stupid song to get to...
13        if u watch the hills in london u will realise...
14                                          Got the news
15           The storm is here and the electricity is gone
16                                                agreed
17       So sleepy again and it's not even that late. I...
18        lady gaga tweeted about not being impressed b...
19       How are YOU convinced that I have always wante...
20        oh too bad! I hope it gets better. I've been ...
21       Wondering why I'm awake at 7am,writing a new s...
22       No Topic Maps talks at the Balisage Markup Con...
23       I ate Something I don't know what it is... Why...
24       so tired and i think i'm definitely going to g...
25       On my way home n having 2 deal w underage girl...
26         i'm sorry people are so rude to you, isaac, ...
27       Damm servers still down  i need to hit 80 befo...
28       Fudge.... Just BS'd that whole paper.... So ti...
29           I HATE CANCER. I HATE IT I HATE IT I HATE IT.
                                  ...
4969      Same here!  I just wanted it to keep going an...
4970      if you could get down to easton, you could jo...
```

(b) After data pre-processing

Figure 4.1: Data pre-processing

5 Algorithms and Models Used

We selected four models suitable for multi-class classification to train with our labelled Twitter dataset: Logistic Regression, Naive Bayes, Support Vector Machine, and BERT with Fine-Tuning.

5.1 Logistic Regression

We chose to use logistic regression because it is one of the basic machine learning models that is suitable for multi-class data. In a logistic regression model, the training data is fit to a linear model using a set of weights and biases, and the logistic function is applied to project the output space between 0 and 1. This can be used to represent the confidence score of the classification, i.e. a value close to 1 signifies a strong prediction whereas a value close to 0 signifies a weak one. In a multi-class problem, the softmax is taken over all classes to find the prediction with the highest confidence score.

5.1.1 Implementation

We used the `LogisticRegression` class from the `sklearn` library with the default settings and the following parameters:

1. `multi_class = "multinomial"` for multi-class classification

5.2 Naive Bayes

We chose to use a naive Bayes classifier because it is a popular choice used for document classification based on word frequencies. A naive Bayes classifier is a probabilistic model based on Bayes' Theorem which aims to maximize the conditional probability of the output class variable given a set of input features. It relies on the assumptions that the input predictors are independent and have equal weight towards the outcome. Multinomial Naive Bayes is a type of Naive Bayes classifier that employs a multinomial distribution for each input feature, which is suitable for problems such as text classification based on word counts and our problem of predicting tweet sentiments.

5.2.1 Implementation

We used the `MultinomialNB` class from the `sklearn` library with the default settings and the following parameters:

1. `fit_prior = False` since the class ratios were already balanced

5.3 Support Vector Machine

We chose to use a support vector machine because it is a popular model which performs well on many machine learning problems. In a support vector machine, the objective is to find the optimal hyperplane (also known as the decision boundary) with the greatest margin that separates the input space into two classes. To tackle the problem of non-linearly separable data, the kernel trick can be employed to project the input space into a much higher dimensional and linearly separable feature space with minimum computation involved. From there, a suitable decision boundary can be found using an analytical solver. One of the most widely-used kernels is the radial basis function (RBF) kernel, which projects the input space into an infinite-dimensional space and always results in a linearly separable solution. In order to apply support vector machines to a multi-class problem, two strategies can be used: "ovo" which stands for one-versus-one and "ovr" which stands for one-versus-rest.

5.3.1 Implementation

We used the `SVC` class from the `sklearn` library with the default settings and the following parameters:

1. `decision_function_shape = "ovo"` for more precise comparison between sentiment labels
2. `kernel = "rbf"` after optimization
3. `C = 45` after optimization
4. `gamma = 1` after optimization

5.3.2 Optimization

Kernel We tried to experiment with four different kernels: linear, RBF, polynomial and sigmoid, and found that RBF performed the best (See Table 5.1). Thus, we decided to use the RBF kernel.

Table 5.1: Accuracy of different SVM kernels

Kernel	Accuracy
Linear	0.255
Polynomial	0.353
RBF	**0.376**
Sigmoid	0.119

GridSearch In order to further improve the accuracy of our SVM, we decided to use the `GridSearchCV` class from `sklearn` to find the optimal values of `C` and `gamma`

for the RBF kernel. In doing so, we found that `C = 45` and `gamma = 1` gave the best results, and the accuracy improved from 0.376 to 0.410. Thus, we decided to use these parameters in our final SVM test.

5.4 BERT with Fine-Tuning

We chose to use BERT with Fine-Tuning because we wanted to experiment and see the effectiveness of what the state-of-the-art technology is for many Natural Language Processing (NLP) tasks today. We try to provide a brief overview of the BERT model here.

5.4.1 Overview

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a sequence-to-sequence model based on the Transformer architecture. Given an input sequence, it is passed through a series of encoding layers that converts each word in the input sequence into a context vector all at once. These context vectors are used by another series of decoding layers to identify patterns in the input sequence and generate the corresponding output sequence.

BERT was pre-trained for two tasks:

1. The Cloze task, which involves masking 15% of the input sequence and predicting the missing words
2. Next Sentence Prediction (NSP), which is to determine, given two sentences, if the second comes after the first

However, we can train BERT to perform a variety of NLP tasks using a process called fine-tuning. This involves adapting the inputs into a format which we can pass into BERT, feeding the outputs of BERT into an output layer, and training the entire model end-to-end.

5.4.2 Implementation

We would like to use BERT for multi-class sentiment analysis, so we came up with the following approach:

1. Convert the tweets into a format which we can pass into BERT
2. Design an output layer which can take in BERT's outputs and return a corresponding label for the sentiment
3. Train and optimize the entire model end-to-end

We used the `BertTokenizer` class from the `transformers` module to first tokenize the tweet content into the required format for BERT. Then, we used the `TFBertModel` class with the following parameters:

1. `pretrained_model_name = "bert-base-uncased"` which consists of 110M parameters
2. `max_sequence_length = 50` as we found that all tweets with a few exceptions had no more than 50 tokens

We decided to use a simple fully-connected neural network as the output layer with two hidden layers of 24 neurons each and five neurons in the output layer representing each of the sentiment classes. A dropout of 0.2 was applied after each hidden layer. See Figure 5.1 for a visualization of our model architecture.

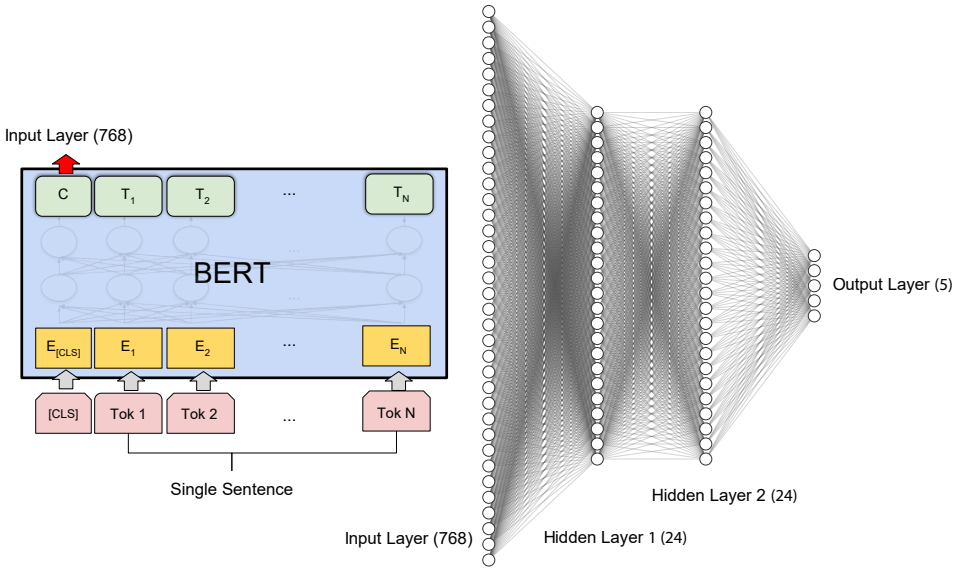

Figure 5.1: Model architecture of our fine-tuned BERT model

5.4.3 Optimization

Batch Size From initial experiments, we found that our machine could not handle a batch size any larger than 25 as it ran out of memory. Thus, we settled with a batch size of 25.

Optimizer We used `Adam` because it is a very popular and effective optimizer which reaps the benefits of RMSProp and AdaGrad. Since our batch size was small, we decided to set a relatively low `learning_rate = 3e-5`.

Epoch Since our learning rate was relatively low, we needed more epochs to train our model. We tried various epochs of 5, 20, and 100, and found that an epoch of 40 was where the validation accuracy started to plateau off. So, we used an epoch of 40 in our final test.

Dropout We tried to use a dropout of 0.5, 0.25, and 0.2, and found that 0.2 consistently delivered the best results. We realised that dropout 0.5 was too high as it caused all the tweets to be predicted as neutral. Thus, the final dropout that we used was 0.2.

6 Evaluation Methodology

Train-Test Split We employed a 70-30 train-test split to train our Logistic Regression, Naive Bayes, and SVM models. However, for our BERT with Fine-Tuning model, we decided to use an 80-10-10 train-validation-test split because we wanted to optimize the number of epochs and batch size used to train our model, and we felt 80-10-10 was a reasonable split since our dataset had 5000 examples.

Loss For the BERT with Fine-Tuning model, we used `SparseCategoricalCrossentropy` with `from_logits = True` as the loss function, since we were dealing with a multi-class classification problem. We used the sparse version with logits because the sentiment labels were encoded as integers and the final output layer was activated with softmax.

Metric Since we were not concerned with false positives or false negatives and were more interested in the percentage of correct predictions, we decided to use accuracy as the metric to compare our models.

7 Results and Discussion

7.1 Results

We report the results of our models in Table 7.1.

Table 7.1: Accuracy of different models

Model	Accuracy (%)
Logistic Regression	21
Naive Bayes	35
Support Vector Machine	41
BERT with Fine-Tuning	**82**

Since the BERT model with Fine-Tuning had the highest accuracy, we chose to use the BERT model with Fine-Tuning to predict the sentiment labels for our unlabelled Twitter and Reddit datasets. We discuss about the data that was generated in the next section.

7.2 Visualisations

We created three interactive visualisations in order to better understand and summarize the data that was generated. They are able to display more detailed and numerical information about the selected region upon hovering over with the mouse cursor. We further describe the visualizations in the following sections.

Keyword vs. Sentiment In our interactive stacked bar graph (See Figure 7.1), we can explore the distributions of sentiments for each keyword. This allows us to see which keyword had the most number of tweets for a certain sentiment. For example, "sgunited" had the largest number of sad tweets.

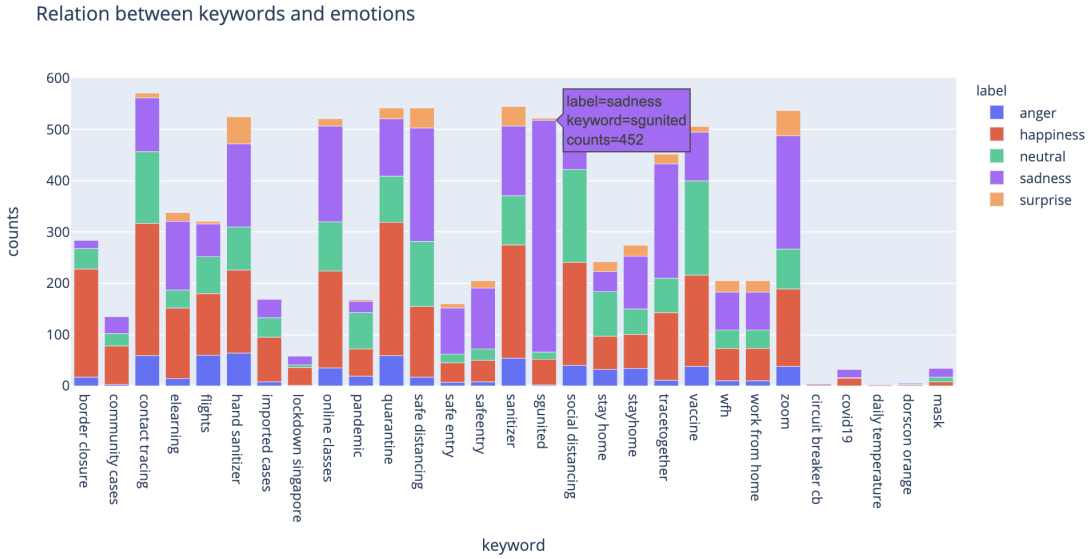

Figure 7.1: Relationship between keywords and sentiments

Sentiment vs. Topic We categorized the keywords into five broad topics in order to generalize the results. The topics are:

1. Singapore Measures
2. Online Work / School
3. Safety Measures
4. Travel and Social Restrictions
5. General

In our Sankey diagram (See Figure 7.2), we can see how the five topics and five sentiments are distributed both ways. For example, in Figure 7.3, we can see how the sentiments for the topic 'Online Work / School' is distributed. Also, in Figure 7.4, we can see how the topics for the sentiment 'sadness' is distributed. With our Sankey diagram, we can explore the relationships between sentiments and topics in closer detail.

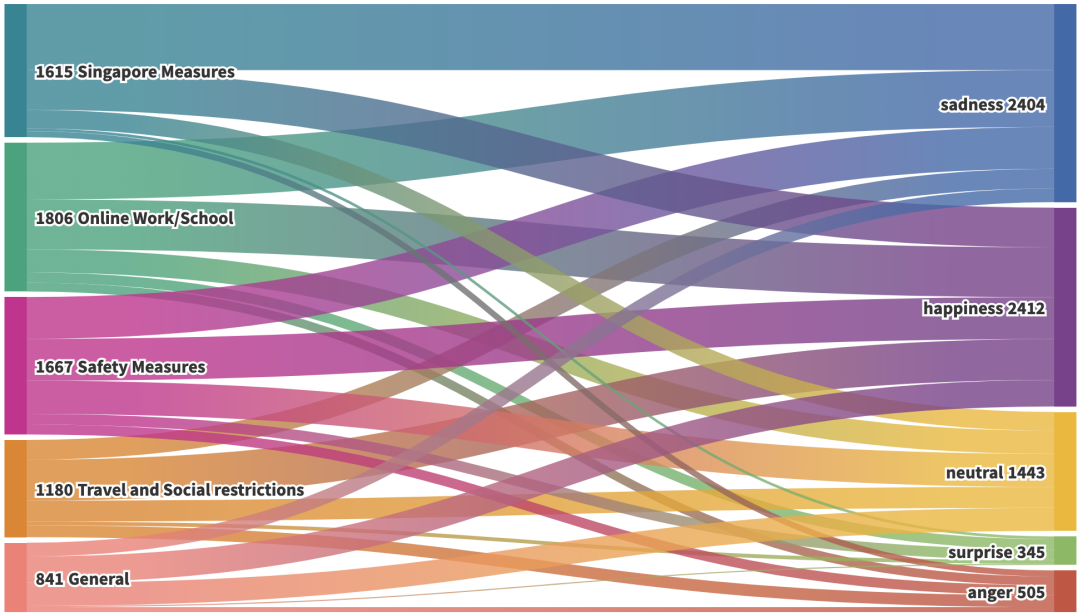

Figure 7.2: Relationship between sentiments and topics

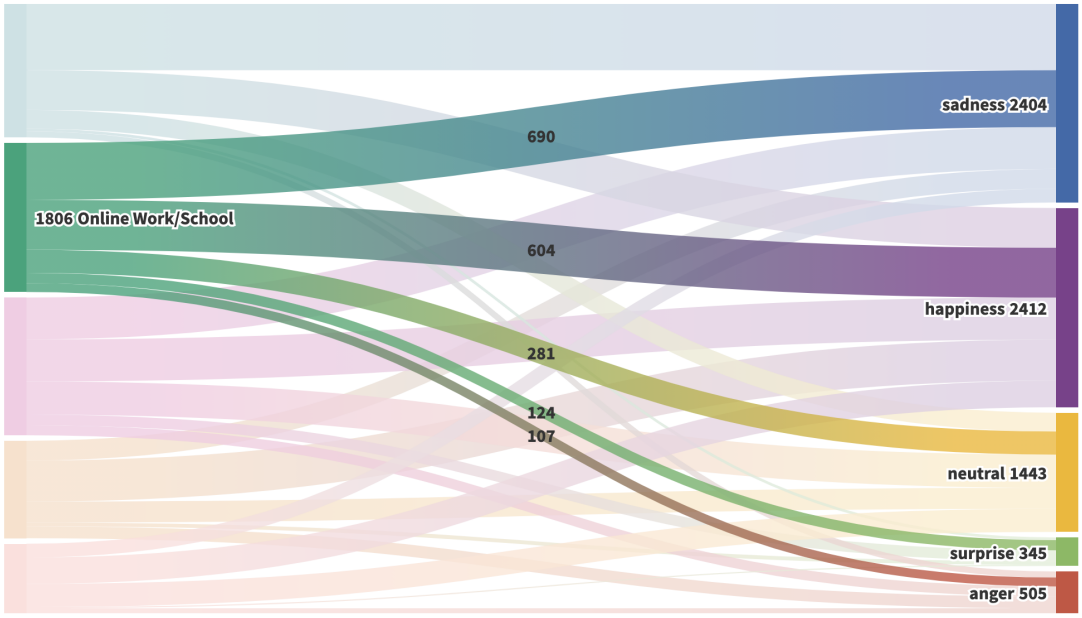

Figure 7.3: Sentiment distribution for online work / school

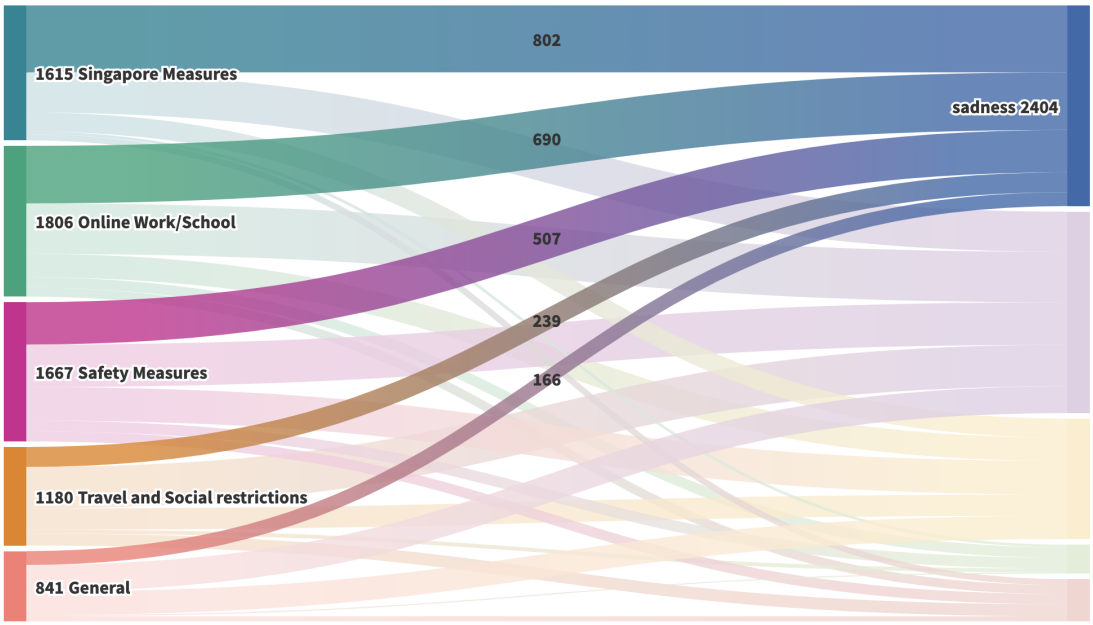

Figure 7.4: Topic distribution for sadness

Sentiment vs. Infection Rate The purpose of this final visualization is to find out if our hypothesis was correct or not. In our hybrid graph (See Figure 7.5), we can examine the monthly distribution of sentiments and the Covid-19 infection rate over time. Now, we can study this graph to test our hypothesis.

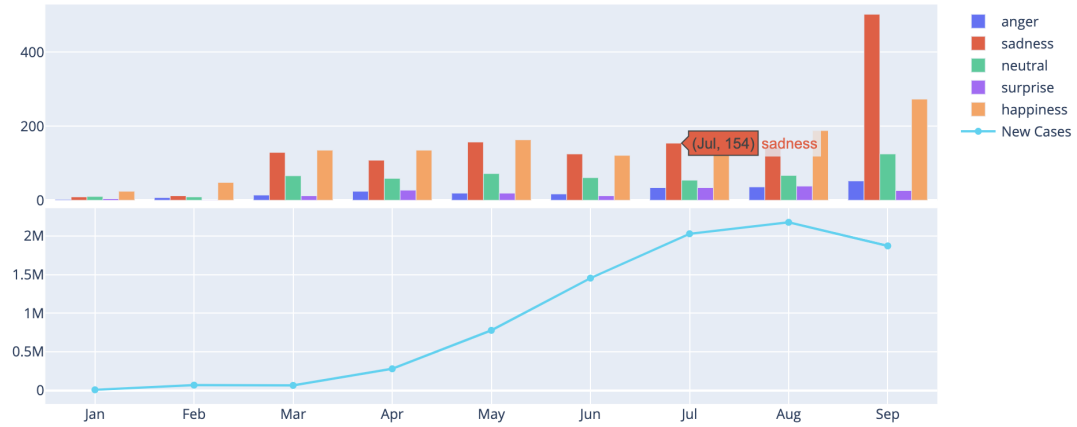

Figure 7.5: Monthly distribution of sentiments and Covid-19 infection rate over time

7.3 Observations

From our visualizations, we were able to come up with the following observations:

1. From Figure 7.1, we see that happiness and sadness are always the prevailing sentiments across all keywords.
2. Figure 7.2 reveals some interesting insights:
   a) People were the saddest about Singapore measures
   b) People were the happiest about online work / school
   c) People were the most neutral about safety measures and precautions
   d) All the topics had a roughly even split between happiness and sadness
3. From Figure 7.5, we can see that the number of sad tweets went up especially in September, which roughly corresponded to the exponential rise in new cases.
4. However, we can also see from Figure 7.5 that the number of happy tweets also increased, although to a smaller extent.

7.4 Analysis

We think that our results do not show that there is a strong correlation between the public's sentiments and the infection rate. Even though the number of sad tweets increased as the number of new cases increased, the number of happy tweets also increased albeit not as much. We think that this is because people in general use social media platforms to both vent their frustrations and spread positivity to friends and family, especially in this period due to social distancing rules. Moreover, sentiments toward Covid-19 are affected by many individual, social, and political factors which vary from person to person and place to place. Because of this, the nature of social media sentiments will not always be linear, and it will not be easy to find obvious trends using our approach. This is in contrast to the Covid-19 infection rate which has seen an exponential growth since it hit the largest countries in the world such as India and the United States.

7.5 Limitations

Choice of Training Data One problem with the training data that we used is that the tweets inside are not related to Covid-19. This limits the transferability of the model that we train with it to the datasets that we created which are exclusively related to Covid-19.

Choice of Data Collected One big problem with the data we collected ourselves is that it was not selected with the training data in mind. This led to the Reddit dataset having no relevance to the model that we trained, which resulted in very poor results (which we decided not to show). This is because the nature of Reddit submissions and comments is very different from Twitter tweets. Reddit submissions and comments come from a forum format where people share their opinions in response to others, and their responses do not have a character limit unlike Twitter. As a result, the way people write Reddit comments and submissions is very different from the way people write Twitter tweets, and thus it is not suitable to use a model trained on Twitter tweets to predict Reddit submissions and comments.

Type of Data Collected There were serious problems with the Twitter data that we collected as well. Due to the way the scraping tool we used worked, we got a much larger proportion of tweets in the later months of 2020 like August, September, and October as compared to the earlier few months. This caused the distribution of tweets to be very skewed towards that period which happened to be when the Covid-19 infection rate started to increase exponentially. Thus, the results we obtained do not accurately depict the sentiments in the earlier half of 2020, which makes it a very weak source to test our hypothesis.

Choice of Train-Validation-Test Split One caveat that we observed during the training of our fine-tuned BERT model was that the validation and test set might have been too small. This is because the validation accuracy was higher than the training accuracy over all epochs during training, which is contrary to what we would expect (See Figure 7.6). This indicates that there could be too little samples in the validation and test set as compared to the training set. Perhaps it would be better to decrease the proportion of the train-validation-test split to 60:20:20 in order to ensure that the model is able to truly perform as well as it did.

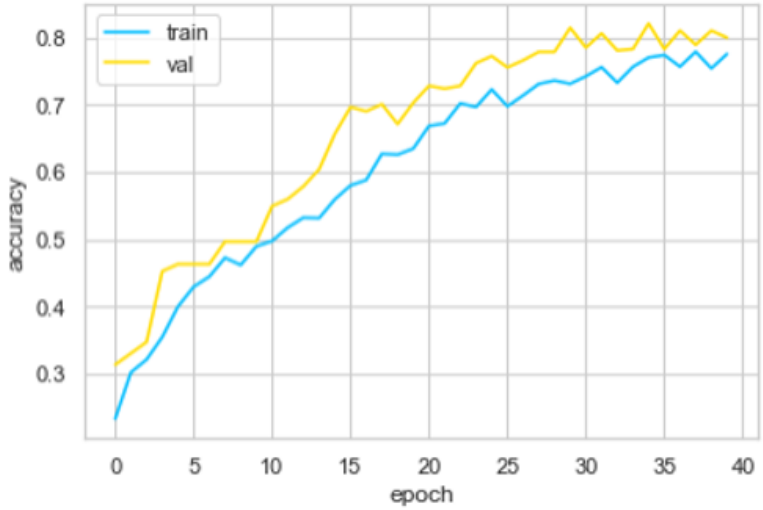

Figure 7.6: Validation and training accuracy versus epochs

7.6 Conclusion

In general, we feel that public sentiment, though interesting, is not a good predictor of Covid-19 infection rate, because not all people who express negative sentiments toward Covid-19 on social media platforms will refuse to follow recommended practices such as social distancing and wearing a mask. Perhaps other more relevant indicators such as government-preparedness and public-orderliness may be more suitable. In the grand scheme of things, our work was only experimental and exploratory in nature, which allowed us to explore the data creatively and come up with various theories, explanations, and reasoning from both a data science as well as a sociological point of view for the various observations that we found.