

Twitter Sentiment analysis on COVID-19 vaccination



July 2021

Name: Yanwen Wang

Student number: 428942

Supervisor: Prof. dr. Rommert Dekker

Co-reader: dr. Clement Bellet

Study program: MSc Data Science and Marketing Analytics

Abstract

COVID-19 has caused substantial public attention. Since 2021, the whole society's attention on COVID-19 vaccination increases. Social media such as Twitter is one of the sources to analyze people's attitudes on the vaccine. This research uses Pfizer and AstraZen different brand vaccine-related datasets and general COVID-19 vaccine dataset to analyze public sentiments. VADER technique, Naïve Bayes, and Support Vector machine method are used for the sentiment classification. The SVM model results have the highest accuracy (approximately 95% accuracy based on the COVID-19 vaccine twitters dataset). Five related hypotheses are tested by two sample t-test, ANOVA, Person's Chi-squared test, multiple linear regression and ordinary logistic regression methodologies. The Covid-19 general statistics and Covid-19 government policies datasets are used to analyze the hypothesis. The result of the hypothesis shows people hold different sentiments on different brands of vaccines. People from developing countries are more optimistic about COVID-19 vaccinations compared to developed countries. The next finding is people's sentiments are more positive on COVID-19 vaccine when COVID-19 confirmed cases of death increase. With the increased number of vaccinations, such as new vaccinations or total vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive. The last finding is different COVID-19 related government policies affect people's attitudes on vaccination differently.

Acknowledgements

I am extremely thankful for Prof. dr. Rommert Dekker great support, guidance and patience.

Table of Contents

Chapter 1: Introduction	5
1.1 Background information	5
1.2 Research questions and hypothesis	6
1.3 Thesis structure.....	7
Chapter 2: Literature Review	8
2.1 Overall sentiment analysis literature	8
2.2 COVID-19 vaccination literature and twitter literature	12
2.3 Hypothesis description.....	14
Chapter 3: Methodology	17
3.1 Sentiment analysis.....	17
3.2 Methodology to address the hypothesis	23
Chapter 4: Data	32
4.1 Dataset 1 : Pfizer Vaccine Tweets	32
4.2 Dataset 2 : Covid -19 Vaccine.....	33
4.3 Dataset 3 : Astrazen Vaccine	33
4.4 Dataset 4 COVID- 19 General Statistics dataset.	34
4.5 Dataset 5 COVID -19 government policy.....	35
Chapter 5: Results	38
5.1 Pre-process.....	38
5.2 Pfizer Vaccine dataset.....	40
5.3 Astrazen vaccine dataset	47
5.4 Vaccine dataset.....	51
5.5 Sentiment analysis.....	57
5.6 Hypothesis 1.....	61
5.7 Hypothesis 2.....	63
5.8 Hypothesis 3.....	65
5.9 Hypothesis 4.....	71
5.10 Hypothesis 5.....	77
Chapter 6: Conclusion and limitation	85
Sentiment analysis.....	85
Hypothesis 1	85

Hypothesis 2.....86
Hypothesis 3.....87
Hypothesis 4.....88
Hypothesis 5.....88
Limitation and future research.....89
References.....92
Appendix.....97

Chapter 1: Introduction

1.1 Background information

COVID-19 has rapidly spread across the world in 2020 and 2021. It is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and has been declared a pandemic by the World Health Organization in 2020. As of February 2021, there are more than 100 million confirmed cases and more than 2.4 million deaths. The virus is mainly spread through the air or direct contact. Therefore, most countries adopted social distance and national lockdown policy to prevent the spread of pandemics.

COVID-19 has serious impacts on global health, economy and social life. Large social and economic disruption has been observed since 2020. The pandemic's direct impact is the reduction of the GDP by a few percentage points (McKibbin & Fernando, 2020). The closing of schools causes a sharp reduction in labor supply when the workers are forced to take time off to take care of their kids. This further led to a reduction of national GDP. It caused both supply shock, such as manufacturing, mining and services and demand shock in specific sectors, such as the transport industry. Some industries suffer severely due to both supply shock and demand shocks, such as entertainment, restaurants and tourism (del Rio-Chanona, etc, 2020).

The development of COVID-19 vaccination is critical to prevent the further spread of COVID-19. On 8th December 2020, a 90 year old woman in the UK was the first to receive the COVID - 19 vaccine. Currently, more countries and companies are developing COVID-19 vaccines, and more countries have started the vaccinations. This can gradually increase the supply and allow more people to get access to the vaccine. On the one hand, accessibility is one issue and on the other hand. On the other hand, people's opinion and attitudes on the vaccine is another concern. If the public is highly against the vaccination and doesn't accept vaccines even when available to access the spread of the pandemic can still not be restricted and stopped.

Many methods can be used to understand the public's attitudes and opinions on vaccination. One way is by conducting surveys such as the online survey conducted by Ward et al. in April 2020. Another possible way to study people's opinions is by performing sentiment analysis on social media platforms. During this pandemic period, people are actively sharing their insights,

thoughts and feelings on social media. Therefore, there is much information that reflects their attitudes on vaccination.

Social-networking platforms, such as Twitter and Facebook have been rapidly growing these years. They are internet-based platforms that enable users to create and share content without the need to be approved by an editor. Social media have profoundly changed our life and the way we interact with people and the world. People use social media to share their opinions and obtain information. The recent events in their daily lives are discussed on social media, representing users' personal opinions. Another social network feature is the user can be anonymous, which allows them to have fewer concerns when sharing their thoughts. People can also "follow" each other or "like" the posts. One-third of the world population is currently using social media, and two-thirds of world internet users are also social media users (Ruiz-Frau, 2020). Therefore, the social media contents are treated as the resource to show a large percentage of people's attitude to a selected topic.

COVID-19 vaccine contents are widely created and shared across different social media platforms. There are many popular social media platforms such as Instagram, Youtube, Facebook, Twitter, Tiktok, etc. This paper uses Twitter data to conduct further analysis. Twitter is a social networking service that allows people to post and interact with posts and messages. Till the first quarter of 2019, Twitter has an average of 330 million monthly active users. In 2020, due to the spread of COVID-19 , there is an increased usage of this platform. The posts on Twitter about COVID-19 vaccination can be powerful tools to understand people's emotions and opinions, which will be further used and analyzed in this paper.

1.2 Research questions and hypothesis

Based on the previous discussed backgrounds and reasons, this thesis is aiming to solve the following central question:

What are the sentiments of people from different countries regarding COVID-19 vaccine and Pfizer Vaccine based on Twitter data and what are the effects of COVID-19 related statistics on the population's sentiments?

1.3 Thesis structure

The research is organized as follows. In section 2, the relevant literature is described, and the hypotheses are developed. Section 3 explained the methods used to answer the central questions. Section 4 describes the data used in this research. Section 5 discusses the results and the findings based on the dataset. The last section concludes the paper, discusses the limitations, and provides some suggestions for future research.

Chapter 2: Literature Review

Many researchers have used the text mining method to study textual data. Text mining is a method used to obtain meaningful information from unstructured text. Based on Tan's study in 1999, around eighty percent of the company's information is textual data. Text mining is used to find the trends or patterns from the unstructured textual data (He, Zha & Li, 2013). In this research, the sentiment analysis is applied to analyze the textual data. The literature reviews are structured as follows. First, the overall sentiment analysis literature is discussed. The lexicon-based approach, machine learning approach are generally reviewed. Then the COVID-19 twitter literature and vaccination-related literature are further discussed. Afterwards, the literatures about the hypothesis are discussed. Finally, the literature summary will be conducted.

2.1 Overall sentiment analysis literature

The sentiment analysis, also called opinion mining, is used to detect whether a textual dataset holds a positive or negative opinion, emotions, or attitudes towards topics or issues (Liu, 2012). Nasukawa and Yi first introduced the sentiment analysis in 2003. Sentiment analysis can process Natural Language processing tasks at document, sentence and phrase level. (Agarwal,ect., 2011) Feldman's research in 2013 further summarized these specific problems.

The document-level sentiment analysis is the simplest form of sentiment analysis by assuming there is one opinion expressed by the author of the document (Feldman,2013). The main approaches for document-level sentiment analysis are supervised learning and unsupervised learning. Supervised learning assumes the number of the classified classes is finite. Besides, it also assumes the training data available for each class is available. The texts have been labeled manually into different classes, for example, positive and negative classes. The training data is used to train the classification model by applying different classification methods, such as SVM, Naive Bayes and KNN. Then we will use a trained machine learning model to classify the sentiment of the new documents (Feldman,2013). Unsupervised learning does not use pre-labeled libraries to classify opinions. It is based on determining some phrases' semantic orientation and comparing the semantic orientation with the pre-defined positive and negative threshold (Feldman,2013). Turney's research in 2002 has classified the recommended and not recommended review by predicting the average semantic orientation of the phrases. This paper

uses pre-tagged part-of-speech to identify phrases. Then to estimate the semantic orientation of a phrase by Pointwise Mutual Information and Information Retrieval method to measure the similarity of pairs of words. In other words, the similarity of the given phrase with the positive reference word and with negative reference word is calculated to obtain the semantic orientation of a given phrase. In this paper, “excellent” are the positive reference word and “poor” is the negative reference word (Turney, 2002).

One document might have different opinions, which requires the use of sentence - level sentiment analysis. This method assumes each sentence or phrase contains one opinion and assumes the sentence contains the entities' identity (Feldman,2013). The sentiment of the subjective sentences is further analyzed. The supervised learning method and unsupervised learning methods can also be applied to analyze the sentence's sentiment.

One sentence can refer to more than one entity and different opinions can be expressed in one sentence. This type of textual data requires aspect – based sentiment analysis. The aspect can refer to the category, feature or discussed topic. This approach categorizes the aspect within the whole document and identifies each aspect's sentiment (Feldman,2013). When the sentence contains indirect opinion such as the comparison sentence, the comparative sentiment analysis can be to extract the preferred entity. Jindal & Liu studied the comparative sentence identification problem in 2006. After categorized the comparative sentences into different types, they implemented the machine learning approach to identifying comparative sentences from the whole documents. After identifying the comparative sentence, the opinion mining approach can be further employed to identify the sentiment (Ding, Liu & Zhang, 2009).

The growth of social media usage allows individuals or organizations to use textual content to make decisions. Many researchers have performed sentiment analysis on social media textual datasets, such as Twitter. Go, etc. have studied sentiment analysis on Twitter data in 2009. They introduce machine learning algorithms approach, such as Naïve Bayes, Maximum Entropy, and support vector machine (SVM), to classify Twitter messages' sentiment. They use positive and negative emoticons to classify the emotions. Based on their results, SVM outperforms other classifiers.

Kharde and Sonawane applied sentiment analysis on Twitter data streams in 2016. In 2018, Jianqiang, etc. introduced the word embeddings method obtained by unsupervised learning, which has higher accuracy than only analyzing lexical and syntactic features. Go et al. (2009) first carried out the study for Twitter sentiment analysis. In this research, they classify the tweets' sentiment into a binary category, which is positive and negative. To reduce the time used to tag the sentiments tweets manually, they introduced the distant supervision method to automatically classify the tweets with emoticons. Their results have above 80% accuracy based on the Naïve Bayes (NB), Maximum Entropy and Support Vector Machine (SVM) method. Giachanou and Crestani (2016) conducted a survey about the algorithms used in Twitter sentiment analysis (TSA) based on 50 articles. Based on these articles, the Support Vector Machine (SVM) and Naive Bayes (NB) classifiers are the most popular methods used on TSA. Besides, they also conclude that the TSA is still an "open field" for the research.

2.1.1 Supervised sentiment classification: Machine-learning based approaches

There are two methodologies in machine learning, which are supervised learning and unsupervised learning. Supervised learning has both the independent variables and the dependent variable associated. Unsupervised learning contains only the independent variables for each observation (Kwartler, 2017). Sentiment classification is the text classification problem, the existing supervised learning method can be directly applied (Liu, 2020). In 2002, Pang et al. applied naïve Bayes, maximum entropy classification and support vector machines to classify the movie reviews. The results show that using bag – of – unigram features with all three classifiers can perform well for particular categories.

The principle of sentiment classification is applied features' effectiveness. Liu has made a complete summary for this (2000). First feature is the unigram and n-grams with their frequency counts. TFIDF (term frequency-inverse document frequency) weighting shows high effective rate in the traditional text classification. TFIDF is calculated by multiplying the term frequency of a word in a document and the inverse document frequency of the word across a set of documents (Liu, 2000). Part of speech (POS) tagging is the second feature to consider. It categorizes the words in a corpus with a specific tag. Another feature is sentiment words and phrase. Seminar words are the words or phrases that express positive or negative sentiments,

such as good or bad. There are many other features for sentiment classification. This paper will not go into details with them.

There are several machine learning methods, such as Maximum Entropy (MaxEnt), Random Forest, Naïve Bayes, Support vector machine (SVM), etc. Many researchers use a machine learning approach to perform sentiment analysis. Naïve Bayes is implemented to calculate the probability of data to be positive or negative. Based on Singh and Husain's study in 2014, Naïve Bayes is originally given by Thomas Bayes. Naïve Bayes is one of the most efficient algorithms to compute. It is based on Baye's Theorem, assuming that the attributes are independent among each other, which makes the algorithm not valid all the time in the real world (Singh & Husain, 2014). MaxEnt uses the estimated probability distribution to perform sentiment classification. (Nigam, Lafferty, & McCallum, 1999). Unlike Naïve Bayes, it does not make independence assumptions for its attributes. Therefore we can add attributes like bigrams and phrases without overlapping problem. Go, etc.'s research in 2009 shows that MaxEnt has high accuracy for sentiment classification. Random forest is the algorithm proposed by Breiman in 2001. It is the ensemble method based on decision tree algorithm. The forecast is an ensemble of decision trees and usually are trained by the "bagging" method. The principle of "bagging" is to increase the results by combination of learning models. Based on Gupte's study in 2014, Random forest classifier has high accuracy and great performance advantages. However, it might be over-fits when the number of trees and vague links are too large. SVM was original introduced by Cortes & Vapnik in 1995. It is first time used in text classification by Joachims in 1998. It maps the optimal boundaries to separate positive and negative training samples. In 2004, Gamon conducted a sentiment classification task on short and very noisy customer feedback data by using SVM. The results show deep linguistic analysis features improve the performance of the classifiers. In general, SVM method had outstanding performance Amrani, etc.'s research in 2018.

2.1.2 Unsupervised sentiment classification: Lexicon-based approach

Sentiment lexicon means a set of words associated with the positive and negative sentiment orientation. Sentiment words contain both individual words and phrases. Sentiment words have two types, base type and comparative type. The base type contains words, such as beautiful, bad,

etc. The comparative type contains words that express comparative opinions, such as better, worse, etc (Liu, 2020).

There are three main approaches to obtain the sentiment lexicon, including manual approach, dictionary - based approaches and corpus – based approaches. Manual approach means people manually code the lexicon by hand. This approach is labor-intensive and time-consuming because each domain requires its own lexicon. Therefore, this method is not feasible and will not further be discussed in this paper (Feldman,2013). The dictionary - based approach using a dictionary containing synonyms and antonyms to expand the sets of words. Specifically, this method first uses a small set of seed sentiment words for the chosen domain. The seed words are collected manually. Then the set of words expanding by applying dictionaries (e.g. Word Net) synonyms and antonyms. However, the obtained sentiment lexicon based on this method is domain-independent (Feldman, 2013). To acquire domain – specific sentiment lexicon, the corpus – based approach is required. This approach uses syntactic patterns to find sentiment words in a large corpus with the set of seed sentiment words. The statistical approach can determine the polarity of the word by calculating the frequency of co-occurrence with another word. More specifically, if the word frequently appears among the positive texts, then the polarity of the word is positive. If it often appears among the negative texts, then the polarity of the word is negative. If it appears equal frequencies among negative texts and positive texts, it is categorized as a neutral word (Rajput & Solanki 2016).

2.2 COVID-19 vaccination literature and twitter literature

Some researches focus on analyzing people's attitudes towards COVID-19 vaccines in 2020 and 2021. Some studies applied survey and interview methods to collect people's opinions. An online survey of more than 18000 adults from 15 countries is conducted in October 2020. The survey results show that 73% of people agree to get a vaccine once it is available and Respondents in Asia tend to have a higher agree rate than people in Europe or North American. The main reason for not accepting vaccines is that the responders are worried about the side effects and clinical trials. However, people's attitudes towards the COVID-19 vaccine may vary among different periods and focus groups. For example, one research uses a survey from April to December to study people's likelihood of getting the COVID-19 vaccine once a vaccine was available. The

survey results show people's self-reported likelihood to get vaccine declined from 74% in April to 56% in early December 2020 (Szilagyi, etc., 2021). Based on another recent interview on 1117 US adults results, only half of American adults plan to get COVID-19 vaccine (Neergaard & Fingerhut, 2020). Only 3 out of 10 respondents are confident in the safety and effectiveness of the COVID-19 vaccine. Graffigna etc. performed a survey in 2020 to get insights on 1004 Italian adult citizens' attitudes towards vaccines during the early days of Italian reopening after the lockdown. Their research shows 15% of respondents would refuse the vaccine and 27% of respondents would be hesitant about it.

Other researches focus on analyzing social network text datasets to understand peoples' sentiment on COVID-19 vaccine. Based on the existing research, in Asia, less than half of the population tends to hold positive opinions towards the covid-19 vaccine. This finding is based on Ritonga research in 2021 and Sv, etc's research in 2021. Ritonga, etc. (2021) studied people's opinions in January on the COVID-19 vaccine in Indonesian by performing sentiment analysis on Twitter data. They applied the Naïve Bayes method on 6000 tweets and found 56% of people are negative towards vaccine, 39% are positive, and 1% holding neutral attitudes during this period. However, this study only contains one-month data that cannot represent the general attitudes in Indonesian. Sv, etc. analyzed India's attitude towards the COVID-19 vaccine using twitter data from September to December 2020. Their results show 35% of social media posts about COVID-19 vaccines are positive. Moreover, their study also shows the major concerns for Indian citizens about COVID-19 vaccines are health and allergic reactions. This theory is further supported by Dhingra, etc.'s research in 2021. They analyzed COVID -19 vaccine sentiment by analyzing 24000 tweets. The author specifically focused on China and India and further compared these two countries' sentiment analysis results. This result shows nearly 55% of tweets are positive towards the vaccination, 30% of tweets holding neutral attitudes and almost 15% of tweets are negative towards vaccination. China covid-19 vaccination has around 40% positive statements compared to 35% positive tweets in India. China also has 13.6% people holding negative attitudes on vaccination, which is 1.6% higher than India. However, the author did not clarify the period range of the data in this paper. The results of the finding might be varied in different periods.

2.3 Hypothesis description

To address the central question, the following sub-questions are formulated.

Sub-question 1: *Do people hold the same sentiments on different brands of COVID-19 vaccine (such as Pfizer, or Astrazen) ?*

Hypothesis 1 : *People's sentiments on different brands of vaccine are different.*

To answer this question, this paper analyzes two different brands vaccine datasets, viz. Pfizer vaccine tweets and Astrazen vaccine tweets. By comparing these two dataset results, people can understand whether people hold different attitudes on different vaccine brands. If people are more positive towards a particular vaccine brand, the social acceptance of the COVID-19 on this specific brand is expected to be higher also. The results of this hypothesis can be helpful for the government to design the COVID-19 vaccination strategy. Currently, there is no publicly available research for this hypothesis.

Sub-question 2: *What is the difference in people's sentiment on COVID -19 vaccination between developed countries and developing countries?*

Hypothesis 2: *People in the developed countries are more positive towards COVID -19 vaccination than people in the developing countries.*

Based on the above session literature, people's attitudes towards COVID-19 vaccination vary among the geographical areas. There is no available research about the variation of people's attitudes towards COVID-19 vaccination in developing and developed countries. Therefore, this paper will use other vaccination findings to support further this hypothesis. Van Essen's research in 2003 shows levels of Influenza vaccine use are higher in rapidly developing countries than in developed countries. Their research also indicates that vaccination usage varies widely between countries and none of the countries managed to implement its national vaccination recommendations fully. This result helps us shape the second hypothesis that people in rapidly developing countries such as India and China are more positive towards COVID-19 vaccination than people in developing countries.

Sub-question 3: *Will people's sentiment be more positive on COVID-19 vaccine if the confirmed new cases, death, total confirmed cases or total death increase?*

Hypothesis 3: *people's sentiments will be more positive on COVID-19 vaccine when COVID-19 confirmed new cases, death, total confirmed cases or total death increase.*

This paper would like to study the effects of confirmed COVID-19 new cases and death on people's sentiment in this sub-question. Similar to the previous sub-question, limited research has studied this topic. Sv, etc. 's analysis in 2021 shows a positive correlation between COVID-19 confirmed cases and the positive sentiment towards COVID-19 vaccines in India. Based on their research result, the hypothesis for this sub-question is a positively correlated between COVID-19 confirmed new cases and positive sentiment towards the vaccine. Besides, people also tend to be more positive towards the COVID-19 vaccine when the death of COVID-19 increases.

Sub-question 4: *Will the sentiment of the twitters be more positive if the total number of vaccination or new vaccination increases?*

Hypothesis 4: *With the increased number of vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive.*

Since December 2020, multiple countries started the COVID-19 vaccination. With the increasing number of vaccination, it is interesting to know the effects of vaccination on people's sentiments based on the related twitters.. There is no available research that studied this question. Based on Shiller's herd behavior concept, people who interact with each other tend to think similarly (Shiller, 1995). Based on these concepts, this paper assumes that the increasing number of vaccinations can increase people's positive attitudes towards the COVID-19 vaccine.

Sub-question 5: *What is the effect of the government policy on people's sentiments on COVID-19 vaccine?*

Hypothesis 5: *People are more optimistic about COVID-19 vaccination when the government policies are stricter, such as closedown schools etc.*

This sub-question analyzes the effect of government policy on people's sentiments on the COVID-19 vaccine. There is no available published research on this question. People showed negative emotion on COVID-19 at the beginning of the lockdown. However, as the reopening starts, people's negative attitudes decreased (Ahmed, Rabin & Chowdhury, 2020). This finding indicates that the strict lockdown policies tend to affect people's emotions on COVID-19 negatively. When people lose their job or need to take care of the children who study at home, it is assumed they are more likely to accept the COVID-19 vaccine to speed up the reopening process. Therefore the hypothesis for this sub-question is people tend to hold more positive attitudes about the COVID-19 vaccine when there are strict government policies. In this hypothesis, different government policies such as school closing, working place closing, cancel public events. Restriction on gathering, closing public transportation, stay at home requirements, restrictions on internal movement, international travel control and vaccine policy. Because for different countries and in different periods, government policies changes a lot due to the spread cycle, government policies for each country on daily basis are used in this hypothesis.

Chapter 3: Methodology

This session will discuss the techniques used in the sentiment analysis and the methods used to exam the hypothesis.





3.1 Sentiment analysis

3.1.1 Text preprocessing

Pre-processing the data is the process of cleaning and preparing the text for sentiment classification. The original twitter data contains lots of noise and not useful parts such as punctuations, HTML tags, etc. Many words such as “the”, “a”, etc. doesn't contain informative messages. Because each word in the twitters text is treated as a dimension, keeping these words in the text will make the classification harder due to the dimensionality problems. Text data usually requires a long time to get the results. Keeping this noise will also reduce the computation speed. This requires us to reduce the effects of the noise in the twitter text and further improve the model's performance.

The Twitter text needs to be preprocessed before performing the sentiment analysis. The pre-processing in this paper includes the following steps:

- Lowercasing all the letters
- Removing stop words
- remove the numbers
- removing the punctuations
- removing the empty spaces
- remove URLs,
- Remove hashtags,
- Remove mentions
- Remove white spaces
- Remove general time
- Remove AM, PM

- Remove newline
- Remove some particular terms    
- Stemming words

In this paper, both the Twitter text and location text contains a lot of noise and need to be cleaned. The location data will be used to categorize the country into developing and developed countries. This will be further explained in the results session. The variables' dates need to transform into the same format. The date and location data will also be used to merge different datasets and examine the hypothesizes.

3.1.2 Unsupervised sentiment classification: lexicon based approach

Several ways can be used to address the sentiment classification problems. Manually creating the sentiment lexicon is the most labor-intensive method, and there could be an error during the labeling process. Therefore, this paper will use dictionary-based methods. The first dictionary is NRC sentiment dictionary. It can be used to calculate the eight different emotions of the text data and obtain the polarity of the text. This dictionary is used to investigate the emotions of Twitter. The results will be visualized in the frequency plots. This dictionary can be accessed through “syuzhet” R package. Eight emotions can be calculated including: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust". The positive and negative polarity are also generated. The second dictionary is “bing” sentiment dictionary. This dictionary is a general-purpose English sentiment lexicon dictionary. The last dictionary is valence aware dictionary and sEntiment Resoner (here after “VADER”). This dictionary is one of the best-unsupervised methods for social media text, especially in Twitter This method is introduced by Hutto, and Gilbert in 2014. It is sensitive to both the sentiments expressed in the Twitter text and is generally applicable to both the polarity and intensity of the sentiments expressed in the Twitter text. This method used the existing well-established sentiment word banks and introduce many common sentiment expression in microblogs, including “Western_style emoticons”(e.g. :-)), “sentiment-related acronyms and initialisms” (e.g.LOL), and “commonly used slang which contains sentiment value” (e.g. nah) (Hutto & Gilbert, 2014). This paper uses Twitter data to analyze people’s sentiment on COVID-19 vaccination, therefor VADER method is used, and the output is described in the results session.

3.1.3 Supervised sentiment classification – Machine learning based approach: Naive Bayes

Bayesian network classifiers are the supervised classification method, which is widely used in sentiment analysis. The intuition of the naïve Bayes classifier is to represent the text document by a bag of words. Thus, the position of the words is ignored. It uses word frequencies as the feature to judge the document's categories.

Bayes theorem is used to calculate conditional probability. The conditional probability means “ the probability of one event occurring, given the other event has already occurred” (Gut, 2013) . The following equation represents the theorem:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$P(x|y)$: Conditional probability of event x occurring, given the event y

$P(x)$: Probability of event x occurring

$P(y)$: Probability of event y occurring

$P(y|x)$: Conditional probability of event y occurring, given the event x

Naïve Bayes is a probabilistic classifier method based on Bayes Theorem. Each predictor is independent of each other in this model (Jurafsky & Martin, 2020).

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|d) = \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

\hat{c} represents the estimated correct class out of all classes c. It has the maximum posterior probability given the document d. $P(d)$ can be dropped because it is the same for each class.

$P(c|d)$ is the probability of class c given document d. The above equation can be further simplified as follows:

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|d) = \operatorname{argmax}_{c \in C} P(d|c)P(c)$$

$P(c|d)$ represents the likelihood of document;

$P(c)$ represents the prior probability of the class.

The document d can be represented by a set of features f_1, f_2, \dots, f_n , the above formula can be write as follows:

$$\hat{c} = \operatorname{argmax}_{c \in C} P(f_1, f_2, \dots, f_n | c) P(c)$$

Naïve Bayes classifiers have two assumptions, which are bag of words assumption and naïve Bayes assumption. Bag of words assumption assumes the position doesn't have the effect on the classification. Naïve Bayes assumption is the conditional independence assumption. which states that features are independent of each other given the class c , which means $P(f_1, f_2, \dots, f_n | c) = P(f_1 | c) * P(f_2 | c) * \dots * P(f_n | c)$. Based on this assumption. The equation can be rewritten as below:

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) * P(f_1 | c) * P(f_2 | c) * \dots * P(f_n | c) = \operatorname{argmax}_{c \in C} P(c) \prod_{f \in F} P(f | c)$$

In the text document, the features are the words at different position in the text.

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) \prod_{i=1}^m P(w_i | c)$$

w_i represents the word at position i ; m represents the position.

$$P(c) = \frac{N_c}{N_{doc}}$$

N_c is the number of documents in the training data with class c . N_{doc} is the total number of documents.

$$P(w_i | c) = \frac{\text{count}(w_i, c)}{\sum_{w \in V} \text{count}(w, c)}$$

$P(w_i | c)$ is the “fraction of times the word w_i appears among all words in all documents with topic c .”. V is the vocabulary which contains the union of all the words in all classes.

However, if there are no training documents that contains the word w_i and are classed as one class c , $P(w_i | c) = 0$ will cause the probability of the class equal to 0. To solve this problem the equation can add one smoothing, as follows:

$$P(w_i | c) = \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} (\text{count}(w, c) + 1)} = \frac{\text{count}(w_i, c) + 1}{(\sum_{w \in V} \text{count}(w, c)) + |V|}$$

For the unknown words in the test set that haven't appear in the training set, we can remove them from the test set. The stop words will be removed by a predefined stop word list.

3.1.4 Supervised sentiment classification – Machine learning based approach

Support Vector machine (SVM) is a supervised machine learning algorithm used for classification problems. It predicts the group or label by finding the hyperplane or boundary line which separates between classes.

SVM draws that hyperplane by transforming the data by the “Kernels” functions. Kernels include linear, sigmoid, RBF, non-linear, polynomial. Since the problem in this paper is just positive and negative linear problems, we will go for “linear SVM”

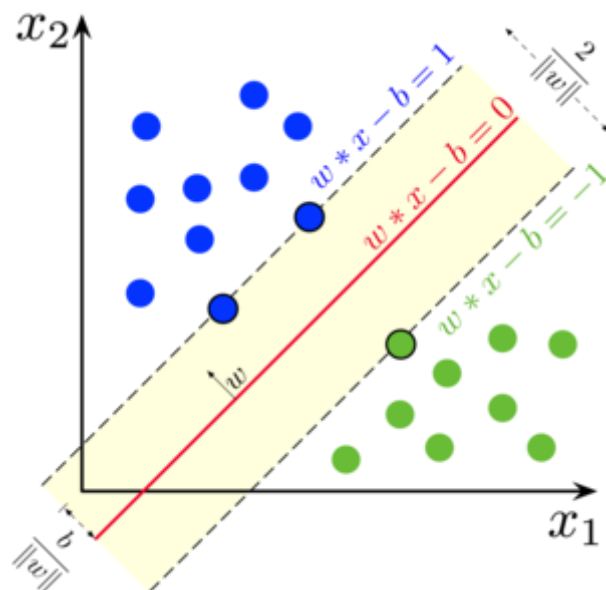


Figure 1: The optimal separating hyperplane between two classes (*from https://en.wikipedia.org/wiki/Support-vector_machine*)

We need first to split the Twitter into training and testing sets and then vectorize the test data to build the model. Then we will create the linear SVM model and evaluate the results.

3.1.5 Evaluation of the sentiment analysis

To evaluate the performance of the sentiment analysis, we need to build the confusion matrix. It is a table visualizing how well an algorithm performs compared to the actual labels.

Table 1 : confusion matrix (two-class classification)

		Actual classes		
		Actual positive	Actual negative	
Predicted classes	Predicted positive	True positive (TP)	False positive (FP)	<i>Precision</i>
	Predicted negative	False negative (FN)	True negative (TN)	
		<i>Recall</i>		<i>Accuracy</i>

Precision measures the correctness of positive predictions. Accuracy measures the correctness of the total number of predictions.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- TP means the total number of correctly classified positive words
- FP means the total number of wrongly classified positive words
- FN means the total number of wrongly classified negative words
- TN means the total number of correctly classified negative words
- TP+ FP means the total number of predicted positive words
- TP+FN means the total number of actual positive words
- TP+TN means the total number of correctly classified words

In this paper we have more than two classes of sentiments, including positive, negative and neutral. Therefore, we need to update the confusion matrix to the three-class classification (Gut, 2013).

Table 2 : confusion matrix (three-class classification)

Actual classes

		positive	neutral	negative	
		Predicted classes	positive	x_1	x_2
neutral	x_4		x_5	x_6	$Precision_{neu}$
negative	x_7		x_8	x_9	$Precision_{neg}$
		$Recall_{pos}$	$Recall_{neu}$	$Recall_{neg}$	

$$Recall_{pos} = \frac{x_1}{x_1 + x_4 + x_7}$$

$$Recall_{neu} = \frac{x_5}{x_2 + x_5 + x_8}$$

$$Recall_{neg} = \frac{x_9}{x_3 + x_6 + x_9}$$

$$Precision_{pos} = \frac{x_1}{x_1 + x_2 + x_3}$$

$$Precision_{neu} = \frac{x_5}{x_4 + x_5 + x_6}$$

$$Precision_{neg} = \frac{x_9}{x_7 + x_8 + x_9}$$

3.2 Methodology to address the hypothesis

3.2.1 Hypothesis 1 :

People's sentiments on different brand of vaccine are different.

This hypothesis is used to verify statistically significant differences in people's sentiment on the AstraZeneca vaccine and Pfizer vaccine. The following methods are used in this hypothesis.

Method 1: two sample t-test:

Since we cannot possibly capture all the tweets sent in this period, therefore the Twitter posts used are the samples taken from the broader population. Since the Twitter posts are scraped from Twitter randomly, we assume the dataset reflects the broader population.

The first method applied is the t-test. It is the method used to test whether the two populations' mean is equal. The sample size for the AstraZeneca vaccine dataset and Pfizer vaccine dataset are not the same. The definition of a two-sample t-test for unpaired data is Null hypothesis (Ho): the two means are equal. The alternative hypothesis (Ha): the two means are not equal.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 is the mean value of the first group

\bar{x}_2 is the mean value of the second group

n_1 is the size of the first group

n_2 is the size of the second group

s_1 is the standard deviation of the first group

s_2 is the standard deviation of the second group

Reject the null hypothesis when t is larger than the critical value of the t distribution with v degree of freedom.

The degree of freedom for unpaired (unequal variance) is

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_1)^2}{n_2 - 1}}$$

The degree of freedom for paired (equal variance) is

$$v = n_1 + n_2 - 2$$

Method 2: Anova

ANOVA is used to test whether there are significant differences among the single brand vaccine group and the overall covid-19 vaccination. ANOVA is a method that decides whether the mean value of two or more groups is different. (Scheffe,1999) The null hypothesis (Ho) is the test

statistic there is no difference in means. The alternative hypothesis is that the means are not equal. The probability value(P-value) tells the probability of getting the obtained result by chance if the null hypothesis were true. It determines whether we should reject the Ho. The significance level is the probability of rejecting the null hypothesis when Ho is true. If the p-value is small (i.e. smaller than 5% significance level), this implies strong evidence against the null hypothesis.

Method 3: Person's Chi-squared test

Instead of sentiment score, we can also use sentiment label to compared the sentiment difference between Pfizer and Astrazen vaccine twitters. Sentiment label is the categorical variable with three levels, we can use Person's Chi-squared test to compare the category variables. The Null hypothesis is there is no relationships between categorical variables. The Alternative hypothesis is there is relationships between categorical variables.

Chi-square is based on difference between the actual observed and the expected relationship between variables. The formula of the Chi-square is as follows:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

f_0 is the observed frequency

f_e is the expected frequency is there is no relationship between the variables.

3.2.2 Hypothesis 2:

People in the developed countries are more positive towards COVID -19 vaccination than people in the developing countries.

Method 1: visualization

First, we will check the sentiment frequency for both the developing country and developed county. The frequency will be visualize in the bar plot. To better understand the sentiment

changes overtimes. The daily sentiment plot or monthly sentiment plot for developed country and developing country are generated and presented in the beginning part of the result session.

Method 2: Two sample t-test

Two sample t-test is used to check whether the developed country average sentiment score is difference to developing country average sentiment score.

Method 3: ANOVA

ANOVA is used to test whether there are significant differences among developing country and developed country. One-way ANOVA is performed to see the effects of country category on Covid-19 vaccine sentiment score. As mentioned before, the null hypothesis of ANOVA is there is no difference in means, and the alternate hypothesis is there is difference in means.

Method 4: ordinary logistic regression

When the dependent variable is a categorical variable, the ordinary least regression model cannot be used. Sentiment label has three groups with the natural order. Ordinary logistic regression is performed to measure the relationship between predictors and ordinal response variables.

Several assumptions of ordinal logistic regression need to be tested. First, the dependent variable should be measured on an ordinal level. The sentiment label is categorized with ordinal level, which is positive, neutral and negative. Second, the variable should be continuous, categorical and ordinal. Third, no multi-collinearity should be the case. This means there are two or more independent variables that are highly correlated with each other. Fourth, each of the observations should be independent and doesn't depend on any of the others. The last assumption is there is proportional odds, which means the relationship between each pair of outcome groups is the same. This ensures the odds ratios across all categories are the same. Because the relationship between all pairs of groups is the same, only one set of coefficients is obtained. Otherwise, we need different sets of coefficients to describe the relationship between every pair of the outcome groups.

Based on the Bilder and Loughin ‘s study in 2014, we can conduct the following interpretation and formulas for the ordinary logistic regression.

Y is the ordinal outcome with J categories. $P(Y \leq j)$ is the cumulative probability of Y less than or equal to one of the category j . The odds of less than or equal to one of the category j is defined as follows

$$\frac{P(Y \leq j)}{1 - P(Y \leq j)} = \frac{P(Y \leq j)}{P(Y > j)} \quad \text{for } j = 1, \dots, J - 1$$

The log odds (or logit) is equal to

$$\log \frac{P(Y \leq j)}{P(Y > j)} = \text{logit}(P(Y \leq j))$$

The ordinal logistic regression model is represents by the following function:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_{j1}x_1 + \dots + \beta_{jp}x_p$$

β_{j0} is the intercepts, $\beta_{j0}, \dots, \beta_{jp}$ are the coefficients with p independent variable for category $j = 1, \dots, J-1$.

Due to the proportional odds assumption, the slope for each category in the model are the same across response categories. Therefore the ordinal logistic regression model can be simplified as follows:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_1x_1 + \dots + \beta_px_p$$

The ordinal logistic regression equation will be calculated through R package ‘‘polr’’. In this package the definition of this model is as follows.

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - k_1x_1 - \dots - k_px_p$$

where $k_p = -\beta_p$.

In this hypothesis, we have only one independent variable. The equation for this hypothesis is as follows:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - \beta_1 * \text{country category}.$$

Country category contains only two levels, developing country and developed country. So although the sentiment has three levels, positive, neutral, and negative, the coefficient of the country category stays the same for both developing country categories and developed country category. So

$$\begin{aligned}\text{logit}(P(Y \leq j | \text{country} = \text{developing})) &= \beta_{j0} - \beta_1 \\ \text{logit}(P(Y \leq j | \text{country} = \text{developed})) &= \beta_{j0}\end{aligned}$$

Here $-\beta_1$ means one unit change in log odds of being in one group versus other group .

The $\text{logit}(P(Y \leq j | \text{country} = \text{developing})) - \text{logit}(P(Y \leq j | \text{country} = \text{developed})) = -\beta_1$.

The proportional odds assumptions make sure the odd ratios for all J -1 categories are the same. In this paper, this means the odds of being negative sentiment versus neutral or positive sentiment is the same as the odds of being negative and neutral sentiment versus positive sentiment. The odds ratios can be calculated by apply log-odds metric to the coefficients. After exponentiate both sides of the above formula,

$$\frac{P(Y \leq j | \text{country} = \text{developing})}{P(Y > j | \text{country} = \text{developing})} / \frac{P(Y \leq j | \text{country} = \text{developed})}{P(Y > j | \text{country} = \text{developed})} = \exp(-\beta_1)$$

To simplify the above formula, we can rewrite it as follows:

$$\frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{1}{\exp(\beta_1)}$$

3.2.3 Hypothesis 3:

People's sentiments are more positive on COVID-19 vaccine when COVID-19 confirmed new cases and death increase.

Method 1: multiple linear regression

Linear regression models are one of the key part of the supervised learning models. The multiple linear regression is a method used to predict the dependent variable based on several independent variables.

$$\text{Sentiment score} = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n + e$$

X_n is the independent variable, which can be total deaths, new deaths , total confirmed cases or new confirmed cases in this hypothesis. We will also include country categorical variable if the previous hypothesis hold. β_n is the coefficient of the independent variable X_n . α is the constant term. e is the error term.

Method 2: ordinary logistic regression

This hypothesis also uses the ordinary logistic regression. The equation for this hypothesis based on R package “polr” is as follows:

$$\text{logit} (P(Y \leq j)) = \beta_{j0} - \beta_1 * \text{country category} - \beta_n * X_n$$

X_n represents confirmed cases or deaths. The interpretation of the algorithm is same as hypothesis 1, so we will not go through in detail here.

3.2.4 Hypothesis 4:

With the increased number of vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive.

Method 1: multiple linear regression

This hypothesis uses different vaccination statistics to test the effects of vaccination on people's sentiment of COVID-19 vaccine related twitters. The total vaccinations, new vaccinations, total vaccination per hundred, new vaccinations smoothed per million are the statistics about the vaccination and will be used in the regression. The multiple linear regression equation is as follows:

$$\text{Sentiment score} = \alpha + \beta_1 * \text{country category} + \beta_n * X_n + e$$

Method 2: ordinary logistic regression

This hypothesis also use ordinary logistic regression to test the vaccination effects on the category variable sentiment labels. The equation for this hypothesis based on R package "polr" is as follows:

$$\text{logit}(P(Y \leq j)) = \beta_{jo} - \beta_1 * \text{country category} - \beta_n * X_n$$

X_n represents vaccination related statistics. The interpretation of the algorithm is same as hypothesis 1, so we will not go through in detail here.

3.2.5 Hypothesis 5:

People are more optimistic about COVID-19 vaccination when the government policies are stricter, such as closedown schools.

Method 1: multiple regression

This hypothesis aims to test the effects on government policies on people's sentiment. Nine policies are tested and the correlated multiple regression formula is as follows:

$$\begin{aligned} \text{Sentiment score} = & \alpha + \beta_1 * \text{country category} + \beta_2 * \text{number of vaccinations} + \beta_3 * \\ & \text{School closing category} + \beta_4 * \text{working place closing} + \beta_5 * \text{cancel public event} + \beta_6 * \\ & \text{restriction on gathering} + \beta_7 * \text{closing public transport} + \beta_8 * \\ & \text{stay at home requirements} + \beta_9 * \text{restrictions on internal movements} + \beta_{10} * \\ & \text{international travel controls} + \beta_{11} * \text{vaccine policy} + e \end{aligned}$$

Method 2: ordinary logistic regression

The equation of ordinary logistic regression for this hypothesis based on R package “polr” is as follows:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - \beta_1 * \text{country category} - \beta_2 * X_2 - \beta_n * X_n.$$

$X_2 \dots X_n$ represents different vaccination related government policies. The interpretation of the algorithm is same as hypothesis 1, so we will not go through in detail here as well.

Chapter 4: Data

Different datasets will be used in this paper: Pfizer Vaccine tweets dataset, Astrazen vaccine tweets dataset, Covid-19 vaccine twitters dataset, COVID - 19 general Statistics dataset, and COVID -19 government policy dataset. The below sections describe the details of each dataset.

4.1 Dataset 1 : Pfizer Vaccine Tweets

The first dataset uses the tweets about Pfizer and BioNTech Vaccine from 2020-December-12th till 2021-June-23rd. In total there are 8927 observations. No retweeted tweets are included. This dataset is obtained from the Kaggle database.

Table 3: Pfizer Vaccine Tweets data descriptions.

No	Columns	Descriptions
1	<code>user_name</code>	The name of the user, as they've defined it.
2	<code>user_location</code>	The user-defined location for this account's profile.
3	<code>user_description</code>	The user-defined UTF-8 string describing their account.
4	<code>user_created</code>	Time and date, when the account was created.
5	<code>user_followers</code>	The number of followers a account currently has.
6	<code>user_friends</code>	The number of friends a account currently has.
7	<code>user_favourites</code>	The number of favorites a account currently has
8	<code>user_verified</code>	When true, indicates that the user has a verified account
9	<code>date</code>	UTC time and date when the Tweet was created
10	<code>text</code>	The actual UTF-8 text of the Tweet
11	<code>hashtags</code>	All the other hashtags posted in the tweet along with Pfizer and BioNTech Vaccine
12	<code>source</code>	Utility used to post the Tweet, Tweets from the Twitter website have a source value - web
13	<code>is_retweet</code>	Indicates whether this Tweet has been Retweeted by the authenticating user.
14	<code>Favorites</code>	Indicates how many people click the favorites for the Tweet

Data source: <https://www.kaggle.com/gpreda/pfizer-vaccine-tweets>

The text data are pre-processed before applying the sentiment analysis based on the method described in the methodology session.

Similarly, the `user_location` data are pre-processed as well. The location using non-English words are transformed into English name. The observations using the American state name to describe the location are also transformed into country name.

The 2019 GDP data is applied to differentiate the developing and developed country. If the GDP is above \$12000 per capita, the country is categorized as developed country. Otherwise, the country is labeled as developing country.

4.2 Dataset 2 : Covid -19 Vaccine

This dataset uses the tweets about COVID-19 Vaccine from 2020-August-13th till 2021-June-23rd. In Total there are around 227 million observations. No retweeted tweets are included. The dataset is obtained from the Kaggle database. The preprocessing steps is similar to the Pfizer vaccine dataset.

Table 4: Covid- 19 vaccine Tweets data descriptions.

No	Columns	Descriptions
1	<code>user_name</code>	The name of the user, as they've defined it.
2	<code>user_location</code>	The user-defined location for this account's profile.
3	<code>user_description</code>	The user-defined UTF-8 string describing their account.
4	<code>user_created</code>	Time and date, when the account was created.
5	<code>user_followers</code>	The number of followers a account currently has.
6	<code>user_friends</code>	The number of friends a account currently has.
7	<code>user_favourites</code>	The number of favorites a account currently has
8	<code>user_verified</code>	When true, indicates that the user has a verified account
9	<code>date</code>	UTC time and date when the Tweet was created
10	<code>text</code>	The actual UTF-8 text of the Tweet
11	<code>hashtags</code>	All the other hashtags posted in the tweet along with #CovidVaccine
12	<code>source</code>	Utility used to post the Tweet, Tweets from the Twitter website have a source value - web
13	<code>is_retweet</code>	Indicates whether this Tweet has been Retweeted by the authenticating user.

Data source: <https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets>

4.3 Dataset 3 : Astrazen Vaccine

This dataset use Astrazen Vaccine twitter data from 2020 August to 2021 June 23. The dataset contains 5240 observations. No retweeted tweets are included. This data used to compare with the Pfizer vaccine to address the first hypothesis. The dataset is generated by filter the covid-19 vaccine dataset astrazen vaccine related twitters. The preprocessing steps is similar to the Pfizer and Covid-19 Vaccine dataset.

Table 5: Astrazen vaccine Tweets data descriptions.

No	Columns	Descriptions
1	user_name	The name of the user, as they've defined it.
2	user_location	The user-defined location for this account's profile.
3	user_description	The user-defined UTF-8 string describing their account.
4	user_created	Time and date, when the account was created.
5	user_followers	The number of followers a account currently has.
6	user_friends	The number of friends a account currently has.
7	user_favourites	The number of favorites a account currently has
8	user_verified	When true, indicates that the user has a verified account
9	date	UTC time and date when the Tweet was created
10	text	The actual UTF-8 text of the Tweet
11	hashtags	All the other hashtags posted in the tweet along with #CovidVaccine
12	source	Utility used to post the Tweet, Tweets from the Twitter website have a source value - web
13	is_retweet	Indicates whether this Tweet has been Retweeted by the authenticating user.

4.4 Dataset 4 COVID- 19 General Statistics dataset.

This dataset contains the COVID-19 general statistics such as the total confirmed cases, new confirmed cases, death or vaccinations , testing, etc. from 2020 end of February to 2021 beginning of July. The dataset is collected and maintained by Our World data (Roser,etc., 2020) The complete features in this dataset can be found in appendix. The confirmed cases, death and vaccinations are the main features included in this paper. All information about this dataset can be found in appendix. All the statistics used in this paper is listed in table 6.

Table 6: Covid- 19 general statistics dataset.

Variables	source	category	description
iso_code	International Organization for Standardization	Others	ISO 3166-1 alpha-3 – three-letter country codes
location	Our World in Data	Others	Geographical location
date	Our World in Data	Others	
total_cases	COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Confirmed cases	Total confirmed cases of COVID-19
new_cases	COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Confirmed cases	New confirmed cases of COVID-19
total_deaths	COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Confirmed deaths	Total deaths attributed to COVID-19
new_deaths	COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Confirmed deaths	New deaths attributed to COVID-19
total_vaccinations	National government reports	Vaccinations	Total number of COVID-19 vaccination doses administered
new_vaccinations	National government reports	Vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days)
total_vaccinations_per_hundred	National government reports	Vaccinations	Total number of COVID-19 vaccination doses administered per 100 people in the total population
new_vaccinations_smoothed_per_million	National government reports	Vaccinations	New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population

4.5 Dataset 5 COVID -19 government policy

This dataset is collected by Oxford Covid-19 Government response tracker(OxCGRT). The dataset is about the governments’ measures during the pandemic. This dataset is a project of the Blavatnik School of Government. The dataset contains 20 policies which can be categorized as following four indicators: C - containment and closure policies, E - economic policies, H - health system policies and M - miscellaneous policies.

Most policies are recorded on an ordinal scale. They represent how strict the policies are. “Four of the indicators (E3, E4, H4, and H5) are recorded as a US dollar value of fiscal spending.” The ordinal scales of indicators reflect the following information.

- All indicators show overall government response.
- The C and H indicators implicate the containment and health index.
- C and H1 also reflect the stringency index.

- E indicators show economic support index.

The government policies often vary by country or region. More strict government policy will have a higher ordinal value. However, strict policies do not always apply to a large-scale population. If the most stringent policy is only applied to a limited geographic area, the binary flag variable shows the scope limitation. Nine indicators have the flag, including C1-C7, H1 and H6. When the flag value is 0, it shows the policy is limited to a specific geographical region. When the flag value is 1, it shows that the policy is applicable across the country (Roser, etc., 2020).

The following table contains the general description of the government policies used in this dataset. Other information can be found in appendix 1.

Table 7: government policies

ID	Name	Description	Measurement	Coding
C1	C1_School closing	Record closings of schools and universities	Ordinal scale	0 - no measures 1 - recommend closing or all schools open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - require closing (only some levels or categories, eg just high school, or just public schools) 3 - require closing all levels Blank - no data
C2	C2_Workplace closing	Record closings of workplaces	Ordinal scale	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operation 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors) Blank - no data
C3	C3_Cancel public events	Record cancelling public events	Ordinal scale	0 - no measures 1 - recommend cancelling 2 - require cancelling Blank - no data
C4	C4_Restrictions on gatherings	Record limits on gatherings	Ordinal scale	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data
C5	C5_Close public transport	Record closing of public transport	Ordinal scale	0 - no measures 1 - recommend closing (or significantly reduce volume/route/means of transport available) 2 - require closing (or prohibit most citizens from using it) Blank - no data
C6	C6_Stay at home requirements	Record orders to "shelter-in-place" and otherwise confine to the home	Ordinal scale	0 - no measures 1 - recommend not leaving house 2 - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips 3 - require not leaving house with minimal exceptions (eg allowed to leave once a week, or only one person can leave at a time, etc) Blank - no data
C7	C7_Restrictions on internal movement	Record restrictions on internal movement between cities/regions	Ordinal scale	0 - no measures 1 - recommend not to travel between regions/cities 2 - internal movement restrictions in place Blank - no data
C8	C8_International travel controls	Record restrictions on international travel Note: this records policy for foreign travellers, not citizens	Ordinal scale	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions 4 - ban on all regions or total border closure 0 - No availability
H7	H7_Vaccination Policy	Record policies for vaccine delivery for different groups	Ordinal scale	1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability

Chapter 5: Results





This paper analyzed the sentiments of people using tweets regarding the COVID-19 vaccine.

Three Twitter datasets are used to obtain the sentiment results: Pfizer vaccine dataset, AstraZen vaccine dataset, and all brands vaccine dataset.

5.1 Pre-process

Twitter text

Before performing sentiment analysis, it is important to pre_process the text data. Several types of techniques are applied to preprocessing the twitter text.

- lowercasing all the letters
- Removing stop words
- remove the numbers
- removing the punctuations
- removing the empty spaces
- remove URLs,
- Remove hashtags,
- Remove mentions
- Remove white spaces
- Remove general time
- Remove AM, PM
- Remove newline
- Remove some particular terms    

Bots

Twitter has had the problem of bots for many years. The bots need to be removed to reduce the impact of bots on the performance of the analysis.

First, the repeated twitter is removed from Pfizer, AstraZen, and the Covid vaccine dataset.

However, some of the Twitter change only a few stop words instead of using the same Twitter.

These twitters are also treated as bots in our analysis. After removing the twitters, the observations in each dataset is visualized in Table 8.

Table 8: Bots for each dataset

Dataset	Total twitters	Bots (same text)	Bots (similar text)	Twitters useful
Pfizer	8927	7	804	8116
Astrazen	5240	3	412	4825
Covid vaccine	227307	122	28495	198690

Location

All the dataset's locations are with a lot of noise text. Therefore a similar preprocess step is applied to clean the location information for each text. The detailed cleaning process is as follows:

- lowercasing all the letters
- Removing stop words
- remove the numbers
- removing the punctuations
- removing the empty spaces
- remove URLs,
- Remove hashtags,
- Remove mentions
- Remove white spaces
- Remove general time
- Remove AM, PM
- Remove newline

Some of the location names use the non-English term, such as "Türkiye", "België", etc. are changed into English terms. Some of the Twitters from USA used state names are changed into the country name. The country codes are also further generated to better merge with the GDP data.

The word cloud is one of the basic techniques to visualize the text data.

It is generated to display the frequently used words by using “worldcloud” package in R.

Wordcloud visually represents the word frequency of the text data. Figure 2 shows the words with a minimal frequency of 30 from the Pfizer Vaccine Twitter data. The above word cloud shows that the most frequently used words in the tweets are vaccine, covid, dose. The different colors and size of the words represent the frequency of the words.

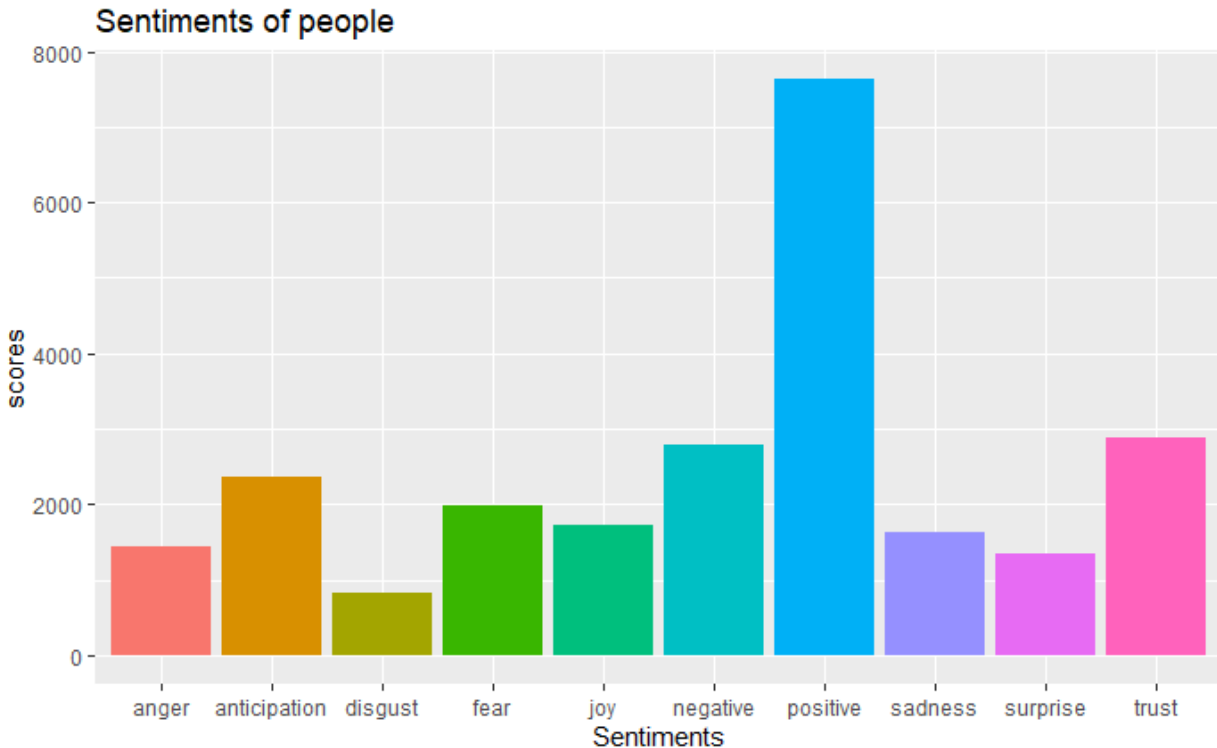


Figure 3: Sentiment of the people’s Twitter on Pfizer vaccine

Figure 3 represents people’s eight emotions and two sentiments positive and negative on Pfizer vaccine. It visualizes people’s different sentiment behind the twitters. Each emotion and sentiment is represented by one color of the bar. The y-axis is “scores”, representing the twitters’ frequency on the certain sentiment. The figure shows positive sentiment bar is the highest, indicating people’s most frequently use optimistic words on their twitters regarding Pfizer vaccine. In order to get the emotion data, the “syuzhet” package is used. 8 emotions are obtained and visualized by the count of words. The trust is the highest bar, followed by anticipation and fear. This can be explained by on the one hand, people tend to trust Pfizer vaccine and indicating

their willingness to trust it. However, on the other hand, people are looking forward to it and fear it might bring other effects.

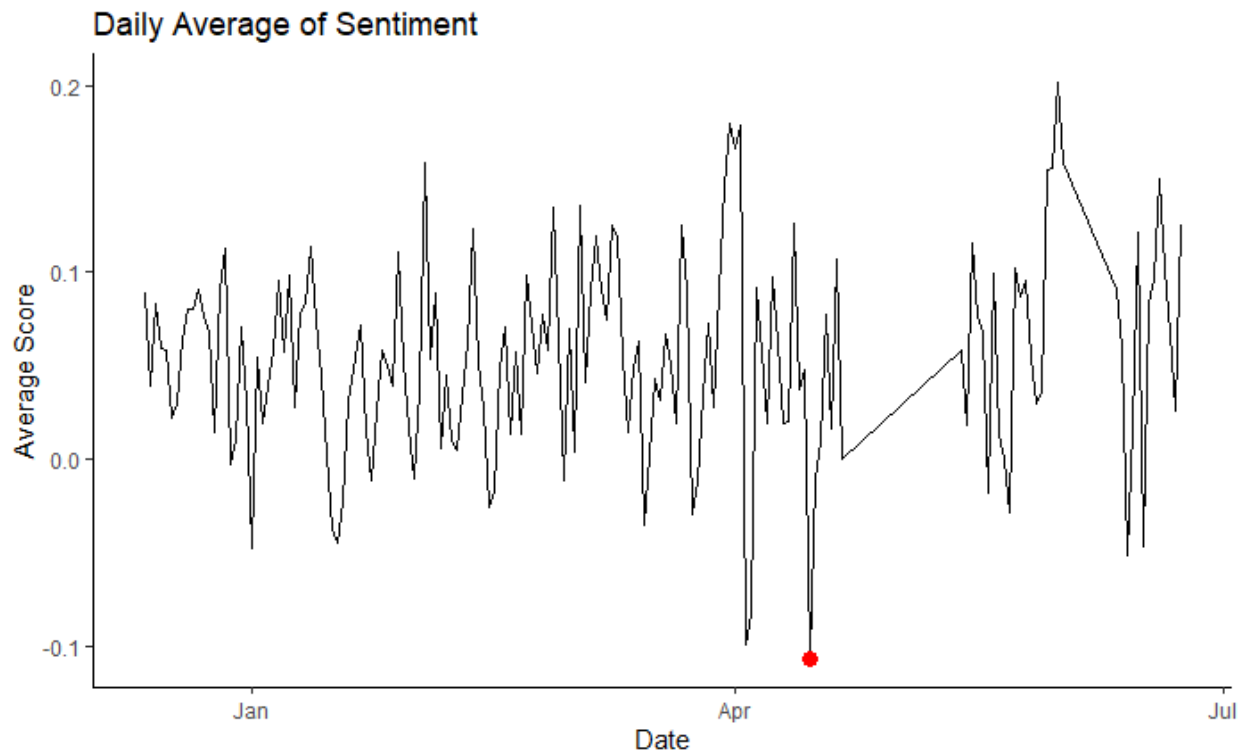


Figure 4: The daily average sentiment score on Pfizer COVID vaccine

Figure 4 shows the daily average sentiment score on Pfizer COVID vaccine from December 12th,2020 till June 23rd,2021. The “Vader” package is used to obtain the polarity scores. The scores have a normalized scale between -1 and 1. It is the lexicon based method and used to categorize the sentiment labels. In this analysis, when the score is 0, the twitter sentiment is labeled as neutral. When the score is above 1, the twitter sentiment is labeled as positive. When the score is below 1, the twitter sentiment is labeled as negative.

In order to understand the change of sentiment over time, the mean sentiment score based on the date is calculated and visualized in Figure 4. Overall, the sentiment average score is positive and fluctuate with 0.15 range from 2020 mid of December to 2021 end of March. One May 31th, the average sentiment score is the highest, with 0.20. On 15th April, the score is the lowest, with -0.106. From April 21th to May 13th, as well as from June 1st to June 11th, data is missing.

Therefore the plot shows a straight line. To better understand the effects, the plot is split into two plots based on the sentiment label.

Table 10: Top daily sentiment score

Group	date	average score
Top low	2021-04-15	-0.106
	2021-04-03	-0.099
	2021-04-04	-0.082
Top high	2021-05-31	0.201
	2021-03-31	0.179
	2021-04-02	0.178

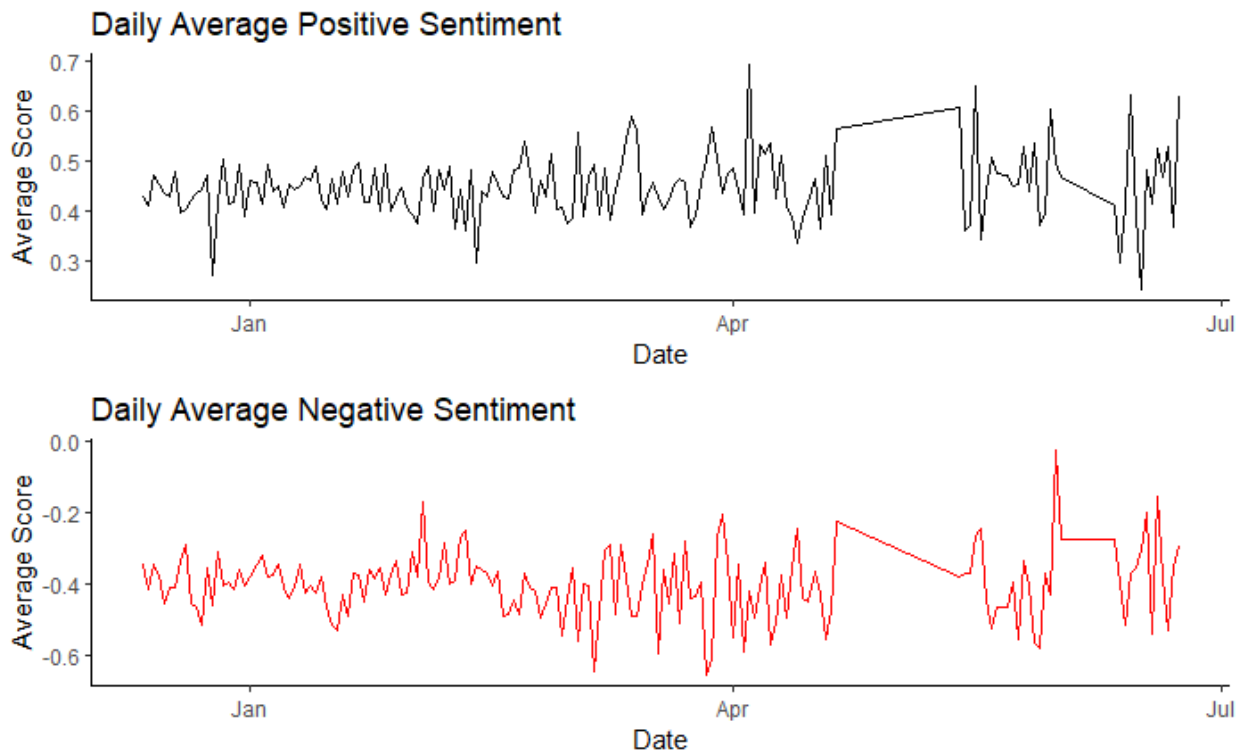


Figure 5: The daily average positive sentiment score and negative sentiment score on Pfizer COVID vaccine

Figure 5 shows the daily average positive and negative sentiment scores. The data is missing from April 21st to May 13th as well as from June 1st to June 11th. From the plot, we can see, on average, people's positive sentiment score stably fluctuates between 0.3 to 0.5 before March 2021. From March onwards, the score is more fluctuated, but higher at the same time. This might be due to people's tend to show more positive attitudes on Pfizer vaccine over time. Based on the daily average negative sentiment plot, the average score is varied between -0.2 to -0.6 before February. From February to March, the average sentiment score has a decreasing trend. This might be because during these periods, different news reports more side effects of COVID-19 vaccine and some severe side effects even cause death. Most countries just starts vaccination process in beginning of the 2021 or later months. People becomes scared to those side effects reports, which leave to the decrease of the average sentiment score during this period. From March onwards, the score is more unstable than before. This might be due to there is more vaccine side effect news from March onwards.

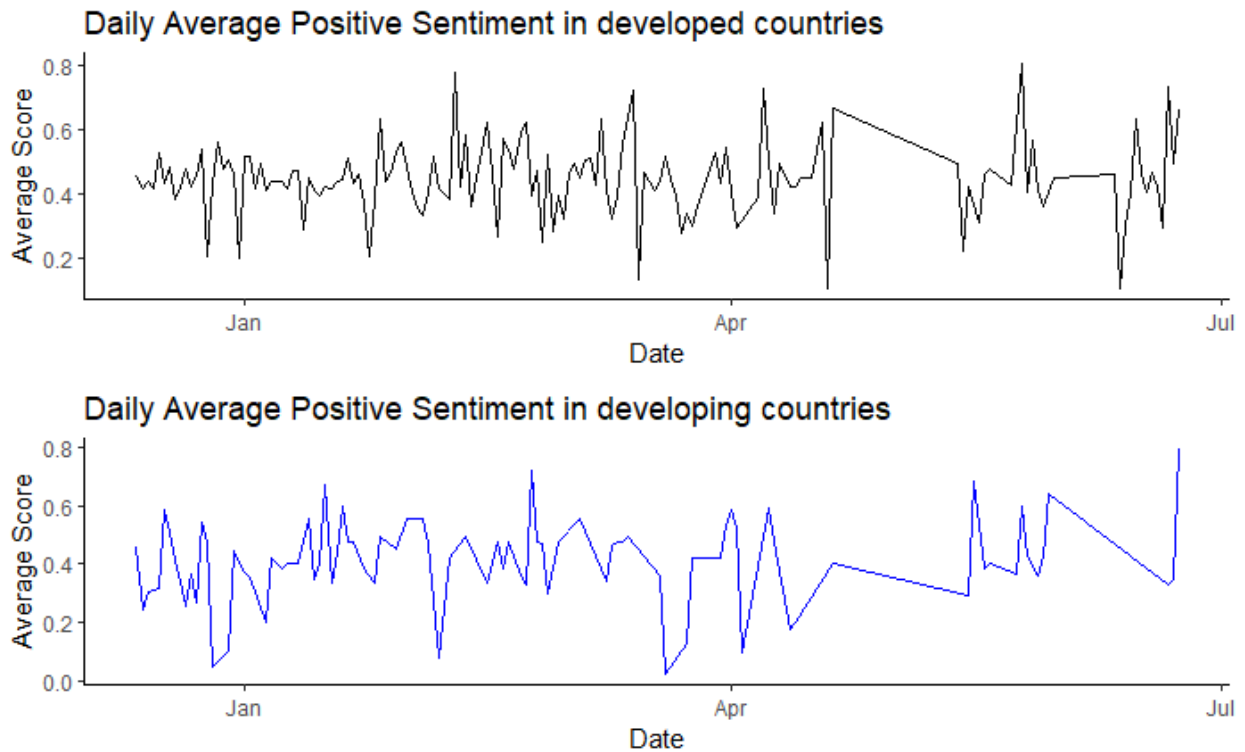


Figure6: The daily average positive sentiment score of developed and developing country on Pfizer COVID vaccine

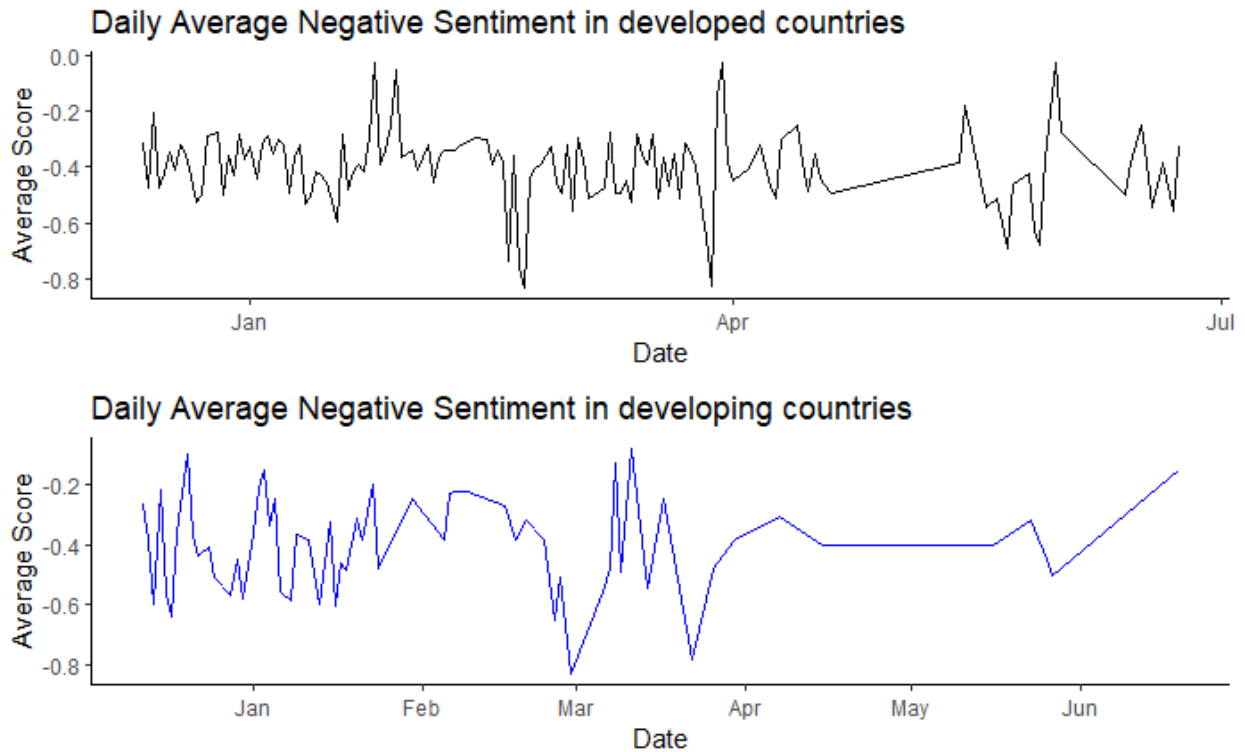


Figure 7: The daily average negative sentiment score of developed and developing country on Pfizer COVID vaccine

Figure 6 and figure 7 visualize the daily average positive and negative sentiment scores of developed and developing countries on Pfizer vaccine. From 12th December 2020 to 23rd June 2021, there are 193 days. However, based on table 11, a lot of dates are missing from the dataset. Therefore, the sentiment trends based on Figure 6 and 7 are not representative. We will not go through these plots in detail in this paper. To understand the difference between different country categories, the mean based on all available date data is calculated and shown in table 11. We can see the developed country has a higher mean score on positive sentiment. Both developed countries and developing countries have a similar mean score on negative sentiment. These might indicate that developed countries tend to be more positive about the Pfizer vaccine than developing countries. However, due to the data limitation, these results might be biased.

Table 11 : Available date for the daily average sentiment plots.

Country category	Available days	Mean score (based on all dates)
------------------	----------------	---------------------------------

Positive : Developed country	142	0.44
Positive: Developing country	87	0.41
Negative : Developed country	132	-0.40
Negative : Developing country	62	-0.40

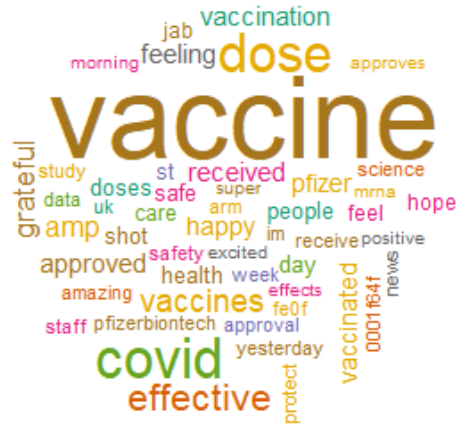


Figure 8: Top 50 common words among positive tweets on Pfizer COVID vaccine

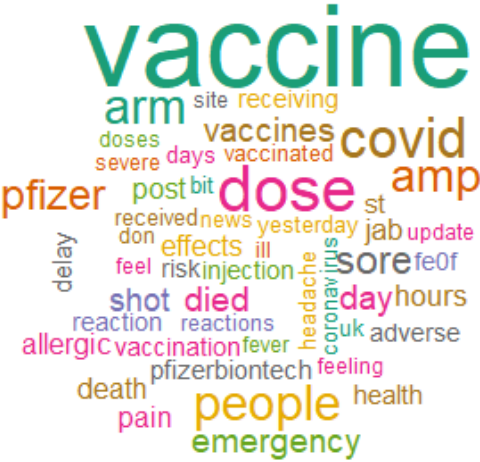


Figure 9: Top 50 common words among negative tweets on Pfizer COVID vaccine

To better understand the difference between the positive tweets and negative twitters on Pfizer vaccine, the world could understand the most common words used in each category. Based on

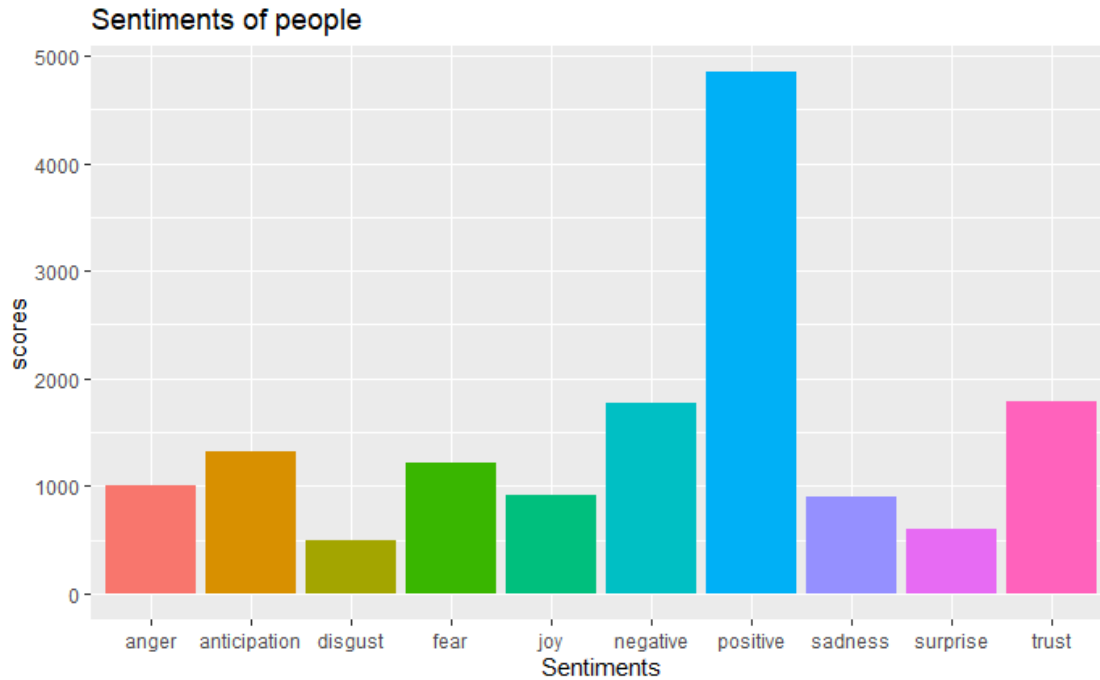


Figure 11: Sentiment of the people's Twitter on Astrazen vaccine

Figure 11 represents people's eight emotions and two sentiments positive and negative on the Astrazen vaccine. Similar to Pfizer dataset, people use more frequently positive words than negative words. More words show trust emotions on Astrazen vaccine than other emotions. Anticipation and fear are the followed emotions.

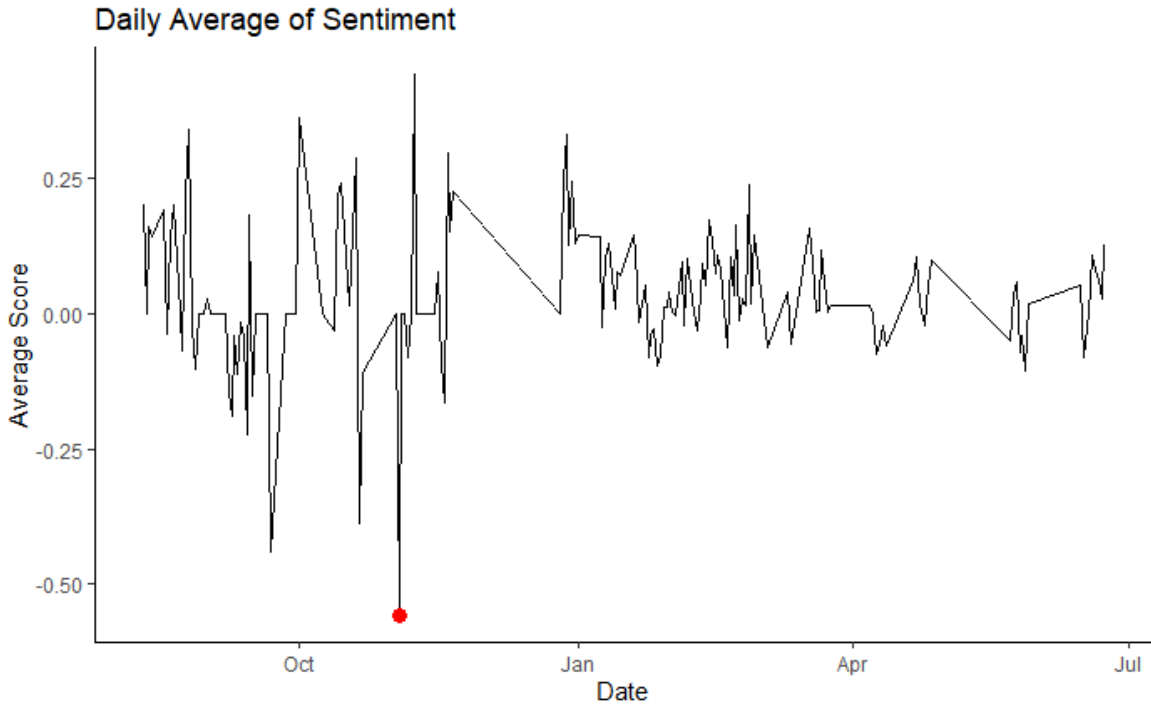


Figure 12: The daily average sentiment score on AstraZen COVID vaccine

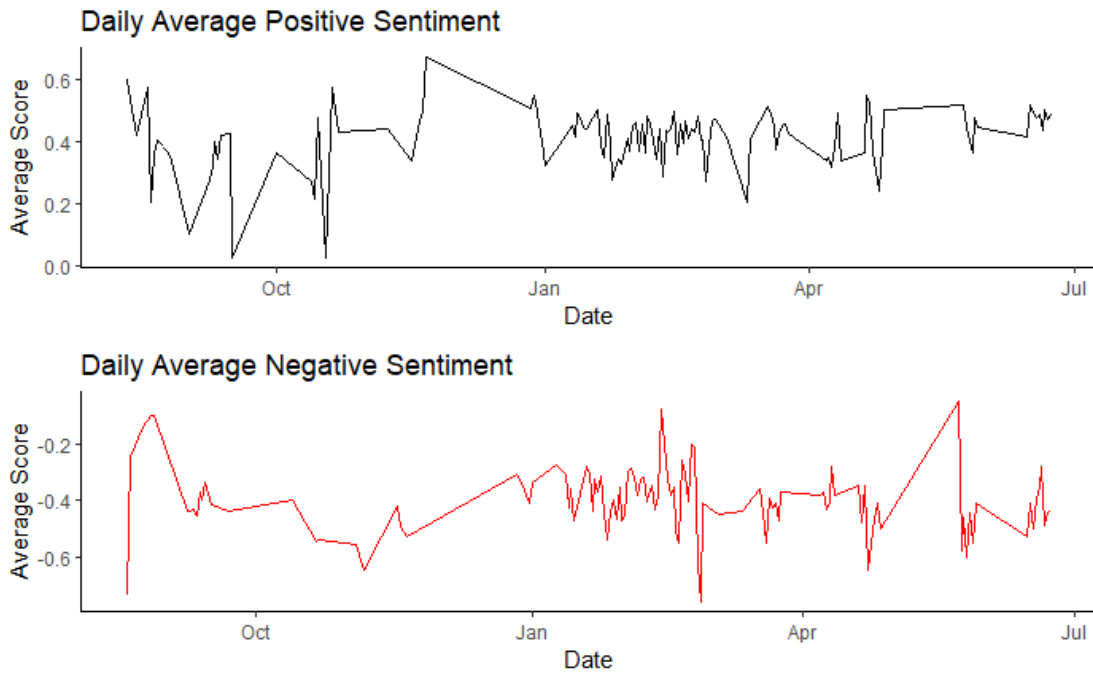


Figure 13: The daily average positive sentiment score and negative sentiment score on AstraZen COVID vaccine

Table 12 : Available date for the daily average sentiment plots on Astrazen vaccine dataset.

Country category	Available days	Mean score (based on all dates)
All	157	/
Positive	122	/
Negative	110	/
Positive : Developed country	78	0.43
Positive: Developing country	66	0.41
Negative : Developed country	77	-0.41
Negative : Developing country	49	-0.38

Figure 12 visualizes the average sentiment scores from 2020 August 11th to 2021 June 23rd of Astrazen vaccine. This period contains a total of 316 days. However, only 157 days include the data. Most of the straight lines represent the missing data on those days. The lowest sentiment average score is on 2020 November 3rd. Figure 13 visualizes the average positive sentiment scores and negative sentiment average sores from 2020 August 11th to 2021 June 23rd of Astrazen vaccine. Similarly, more than half of the dates data is missing, represented by the straight lines in the plots. Due to the large percentage of missing data dates. We do not have enough evidence to conclude the time difference on average sentiment score.

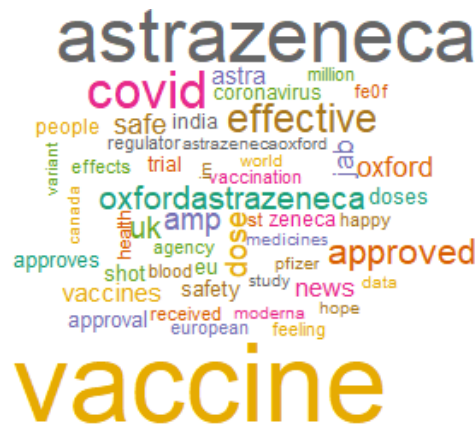


Figure 14: Top 50 common words among positive tweets on Astrazen vaccine

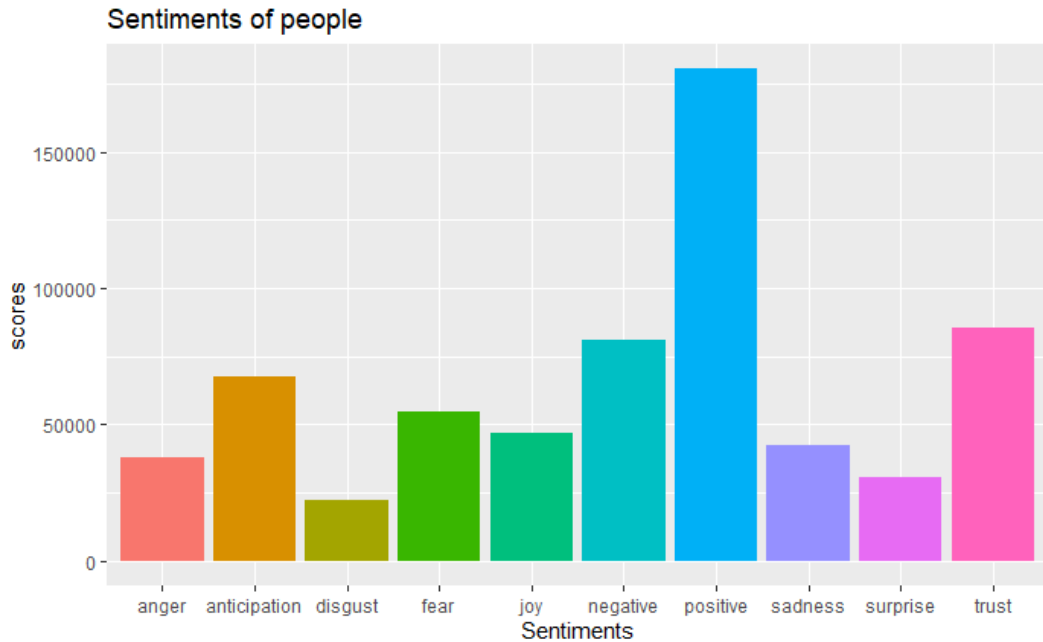


Figure 16: Sentiment of the people’s Twitter on Covid vaccine

Figure 16 represents people’s eight emotions and two sentiments positive and negative on Covid vaccine. Similar to Pfizer and Astrazen dataset, people use more frequently positive words than negative words. More words show trust emotions on Covid vaccine than other emotions.

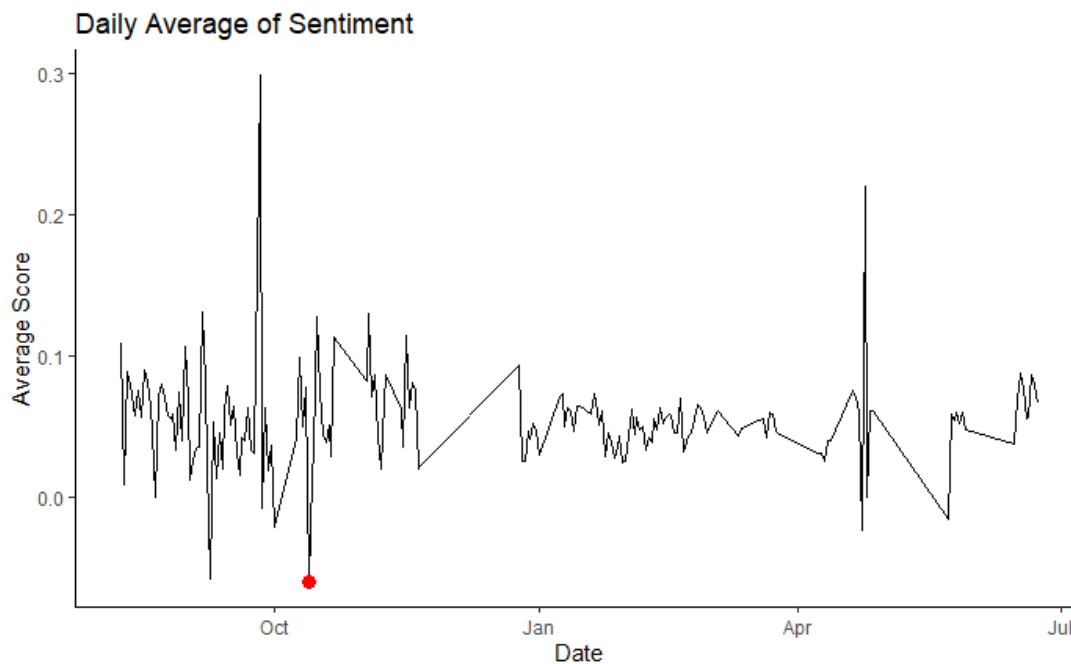


Figure 17: The daily average sentiment score based on date on COVID vaccine

From 9th August 2020 to 23rd June 2021, there are a total of 316 days. 132 days are missing data from this dataset. The dates contain missing data are listed in table 13. Due to the large percentage of the missing date, the average sentiment on the month instead of days is used and presented in Figure 17.

Table 13 : Available date for the daily average sentiment plots on Covid vaccine dataset.

Country category	Available days	Mean score (based on all dates)
All	182days Sep: 25 Oct: 1-8, 23-31 Nov: 1,10-13,22-30 Dec:1-24 Jan: 2-7, 16-18, March: 1-10, 13-17, 26-31 April: 1-6,13-19, 28-30 May: 1-22 23-31 June: 1-14	/
Positive	181	0.44
Negative	178	-0.41
Positive : Developed country	178	0.44
Positive: Developing country	176	0.44
Negative : Developed country	178	-0.42
Negative : Developing country	169	-0.40

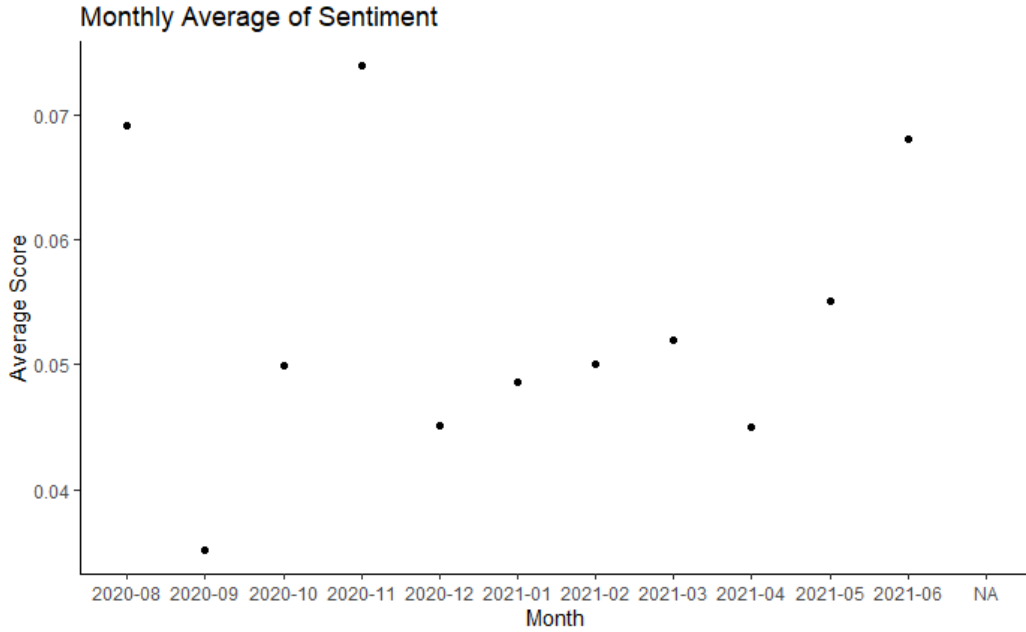


Figure 18: The monthly average sentiment score on COVID vaccine

Figure 18 shows that the monthly average sentiment scores on Covid 19 are all positive. From December 2020 onwards, the sentiment score tend to increase over time.

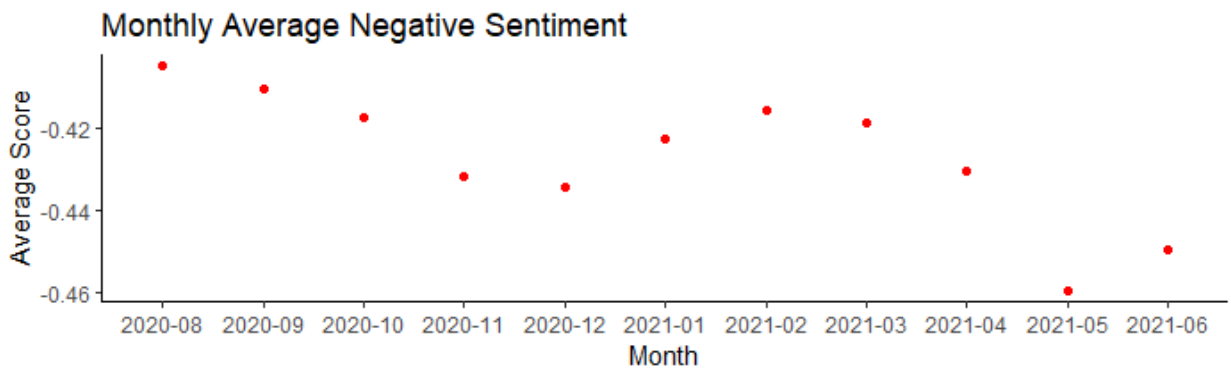
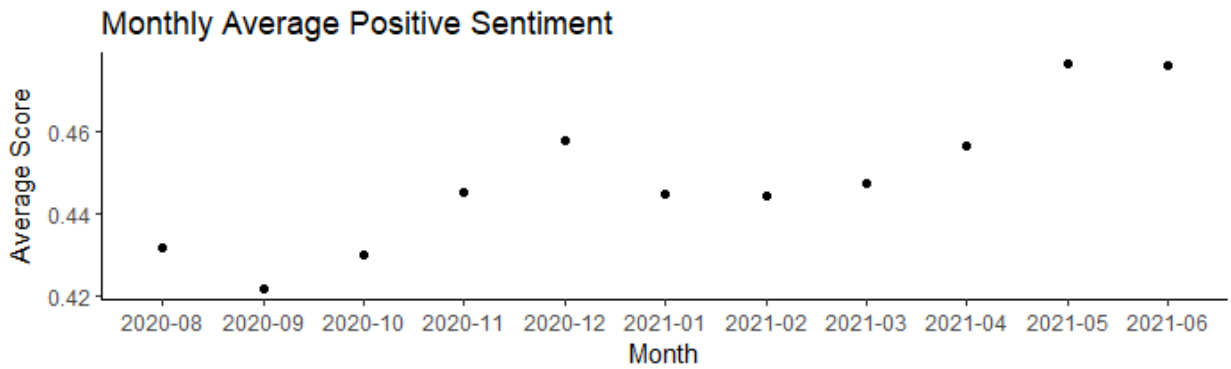


Figure 19: The monthly average positive sentiment score and negative sentiment score based on COVID vaccine

Based on Figure 19, the monthly average positive sentiment score gradually increases from September 2020 to December 2020. However, there is a decrease from December 2020 to February 2021 on average monthly positive sentiment score. From February 2021 onwards, the average positive sentiment score tends to increase over time. The interesting thing is the monthly average negative sentiment score almost shows the opposite trends as positive sentiment. This might be because, over time, people tend to show more positive attitudes towards covid vaccine. However, with more negative side effects news about the vaccine, people are also scared and show negative attitudes on covid vaccine.

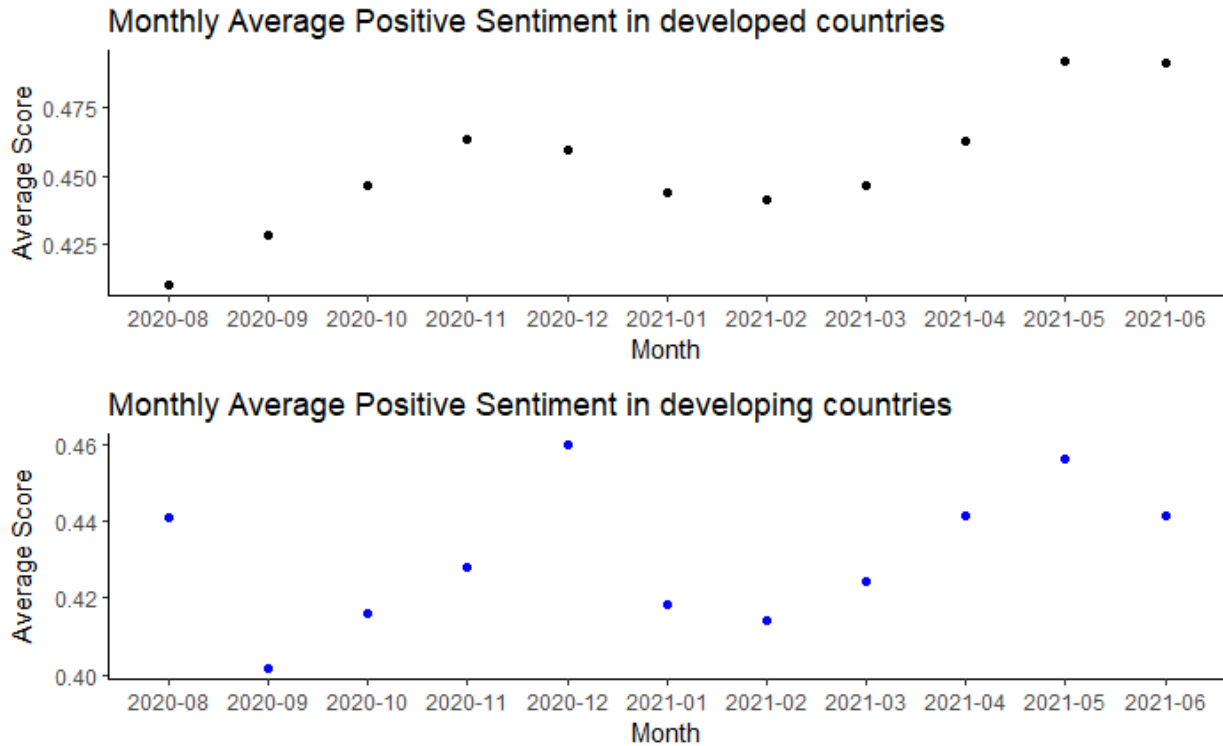


Figure 20: The monthly average positive sentiment score of developed and developing country on COVID vaccine

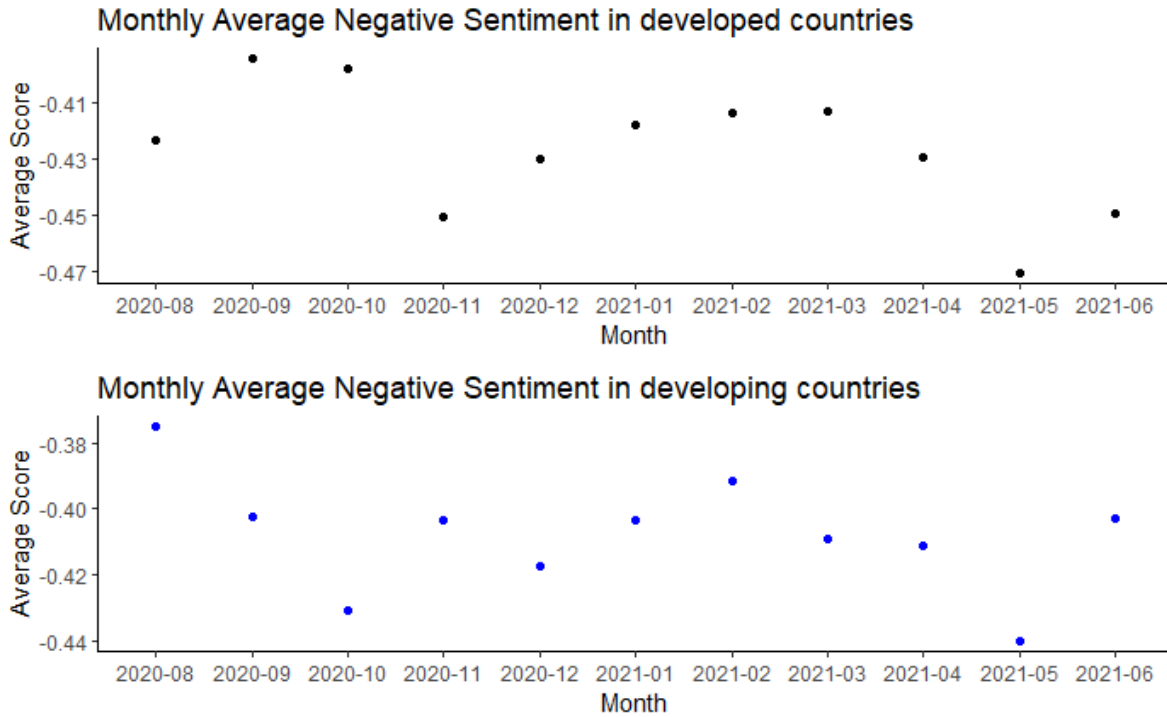


Figure 21: The monthly average positive sentiment score of developed and developing country on COVID vaccine

Figure 20 and figure 21 visualize the daily average positive and negative sentiment scores of developed and developing countries on Covid vaccine. Figure 18 shows that developed countries' monthly average positive sentiment score is stably increased from August 2020 to November 2020 and from February 2021 to May 2021. The positive sentiment score tends to increase over time. Developing country is more fluctuate during this period and has relatively lower monthly average positive sentiment score. Figure 19 shows, both developed and developing countries' monthly average negative sentiment frequently change over time. The average negative sentiment score change range is higher for developing countries compared to developed countries.

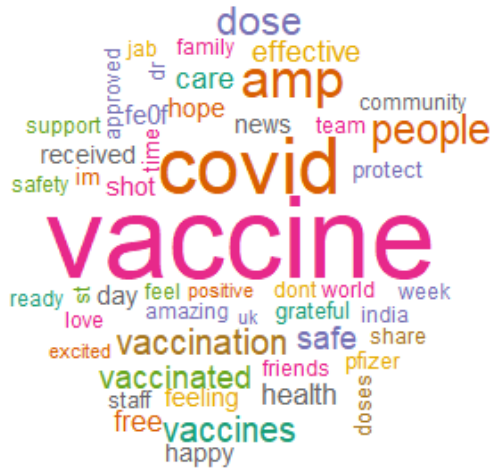


Figure 22: Top 50 common words among positive tweets on covid vaccine

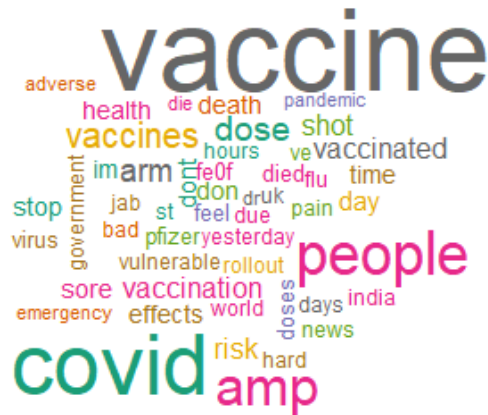


Figure 23: Top 50 common words among positive tweets on covid vaccine

Figure 22 and 23 shows the 50 commons among positive tweets and negative tweets on covid vaccine dataset. The twitters show positive sentiment on the covid vaccine most common use, safe, happy, effective in their twitters. The twitters that show negative sentiment on Pfizer vaccine frequently use words such as death, sore,etc.

5.5 Sentiment analysis

To perform the sentiment analysis, Naïve Bayes and Support Vector machine techniques are used in the paper.

The 70% of the data is used as training dataset. The rest 30% is used as the testing dataset to test the performance. Due to the large dataset and long computation time, the Covid vaccine twitter with missing country data is removed from the dataset to shorten the computation time.

This paper use support vector machine with linear kernel and 10 fold cross validation to predicted the sentiments. SVM is the classification method used to predict the class based on the model obtained from the training dataset.

The confusion matrix measures the performance of the model. The accuracy rate is calculated to compare with different models.

Table 14: confusion matrix based on Naïve Bayes prediction on Pfizer dataset.

predicted	actual			Row total
	neg	neu	pos	
neg	223	113	34	370
neu	128	1123	158	1409
pos	44	170	441	655
column total	395	1406	633	2434

Table 15: contribution matrix based on Naïve Bayes prediction on Astrazen dataset.

predicted	actual			Row total
	neg	neu	pos	
neg	156	65	31	252
neu	95	603	95	793
pos	39	88	274	401
column total	290	756	400	1446

Table 16: contribution matrix based on Naïve Bayes prediction on Covid vaccine dataset part 1.

predicted	actual			Row total
	neg	neu	pos	

neg	675	331	202	1208
neu	322	2486	466	3274
pos	124	368	1122	1614
column total	1121	3185	1790	6096

Table17 : contribution matrix based on Naïve Bayes prediction on Covid vaccine dataset part 2

predicted	actual			Row total
	neg	neu	pos	
neg	655	313	165	1133
neu	298	2665	438	3401
pos	126	365	1119	1610
column total	1079	3343	1722	6144

Table 18: contribution matrix based on Naïve Bayes prediction on Covid vaccine dataset part 3

predicted	actual			Row total
	neg	neu	pos	
neg	697	310	198	1205
neu	298	2567	477	3342
pos	126	391	1156	1673
column total	1121	3268	1831	6220

Table 19: contribution matrix based on Naïve Bayes prediction on Covid vaccine dataset part 4

predicted	actual			Row total
	neg	neu	pos	
neg	635	304	193	1132
neu	341	2499	437	3277
pos	92	380	1183	1655
column total	1068	3183	1813	6064

Table 20: contribution matrix based on SVM prediction on Pfizer dataset

predicted	actual			Row total
	neg	neu	pos	
neg	215	34	26	275
neu	100	1385	110	1595
pos	18	36	510	564
column total	333	1455	646	2434

Table 21: contribution matrix based on SVM prediction on Covid vaccine whole dataset

predicted	actual			Row total
	neg	neu	pos	
neg	10348	106	848	11302
neu	402	29788	456	30646
pos	975	199	16485	17659
column total	11725	30093	17789	59607

Table 14 to Table 21 shows the confusion matrix based on Naïve Bayes and SVM prediction on all three datasets. This is used to calculate the overall accuracy of the prediction.

Table 22: the prediction accuracy based on NB and SVM method for three datasets

	Pfizer	Astrazen	Covid vaccine
Naïve Bayes	73.42%	71.44%	Part 1 (obs.=6096): 70.26% Part 2 (obs.=6144): 72.25% Part 3 (obs.=6220): 71.06% Part 4 (obs.=6064): 71.19%
SVM	86.69%	74.15%	All: 94.99% Part 1: 92.18%

Table 22 shows the accuracy of the Naïve Bayes and SVM for Pfizer, Astrazen, Covid vaccine dataset. All the dataset shows SVM have higher accuracy rate compared to Naïve Bayes. As mentioned before, the Naïve Bayes twitter are split into 4 dataset equally due to the computer

memory limitation. The accuracy for each dataset is around 70% to 72%. By applying SVM to the same part 1 dataset, the prediction accuracy is 92.18%. By using all the twitter, the prediction accuracy increase from 92.18% to 94.99%.

5.6 Hypothesis 1

Hypothesis 1 : *people's sentiments on different brand of vaccine are different.*

This hypothesis is used to verify whether there are statistically significant differences in people's sentiment on AstraZeneca vaccine and Pfizer vaccine.

Method1: Two sample t-test

The two sample t-test can be used to compare means of sentiment scores of Astrazeneca vaccine and Pfizer vaccine. First, the variance of two dataset sentiment scores is compared by F-test .

The Null hypothesis is the variance for the two groups is the same. The alternative hypothesis is the variance is different. F is the ratio between two variances. The more F statistic deviates from 1, the stronger evidence for the unequal variances. If $p > 0.05$, we can assume the two variances are homogenous. The results show $F = 0.89$ and p-value nearly equal to 0.000. The confidence interval for this value is 0.84 to 0.93. Because the p-value is smaller than the significance level 0.05, there is a significant difference between the two variances. Therefore, an unpaired two sample t-test is used in this hypothesis.

The Null hypothesis for the unpaired two-sample t-test is the mean of two groups is equal. The Alternative hypothesis is the mean of the two groups is different. To perform the unpaired two-sample t-test , the `t.test()` function in R is used. The results show t-test statistics value is 2.529. The significance level is p-value = 0.011, which is less than the significance level 0.05. Thus, we can conclude the Pfizer's average sentiment is significantly different from AstraZeneca sentiment score.

Method 2: ANOVA

Next, we can use the one-way analysis of variance (ANOVA) to test whether there are statistically significant differences between the means of sentiment score on two different brands of COVID-19 vaccine twitters.

The null hypothesis of one -way ANOVA is the means for different groups are identical. The alternative hypothesis is the mean of one of the groups is different. A high F-statistics indicates that the null hypothesis holds. To compute the F statistics, we need to divide the between-group variability over the within-group variability. The between-group variability means the differences between the groups within the population. Between-group variability is around 0.642 based on R output. The within-group variability refers to the difference between the groups, which equals approximately 0.097 in this model. R output shows F-statistic is 6.596. The P-value is 0.0102, which is smaller than the threshold of 0.05. Therefore, there is a statistical difference between different brand groups.

Method 3: Chi-test

If we use the sentiment label instead of the sentiment score to compare the sentiment difference between Pfizer and Astrazen vaccine Twitter. The Person’s Chi-squared test can be used to compare the categorical variables. The Null hypothesis is there is no relationship between categorical variables. The Alternative hypothesis is there are relationships between categorical variables. In this hypothesis, we want to test whether there are relationships between the sentiment on the Pfizer vaccine and sentiment on the Astrazen vaccine.

Table 23: Frequency matrix of the sentiment labels for Pfizer and Astrazen dataset

Sentiment label	Pfizer	Astrazen
Negative	1315	961
Neutral	4591	2473
Positive	2210	1391

The frequency matrix in table 23 is created to perform Chi-test. The frequency means the counts of the selected sentiment for the chosen brand of Vaccine. The Chi-test results show the x-squared equals 42.17, and the p-value is much smaller than 0.05 significance level. So we do not

reject the null hypothesis. This means there is no relationship between Pfizer and Astrazen vaccine sentiments.

Both two- sample t-test results and ANOVA test show there is a statistical difference between different brand vaccine's sentiments on Twitter. Chi-test results show there is no relationship between Pfizer and Astrazen vaccine sentiments. Therefore we can conclude hypothesis 1 hold, and people's sentiment on different brands of vaccine are different.

5.7 Hypothesis 2

Hypothesis 2: *people in the developed countries are more positive towards COVID -19 vaccination than people in the developing countries.*

Method 1: visualization :

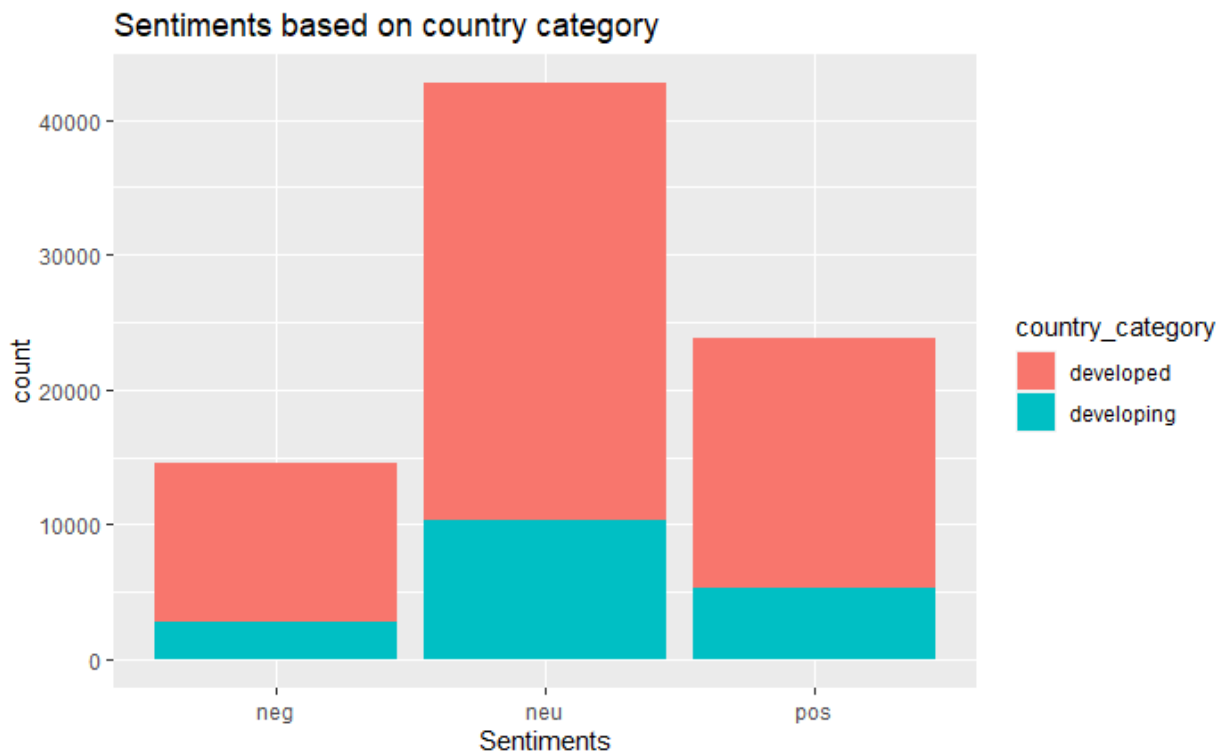


Figure 24 : Sentiment frequency based on country category

Based on the figure 24, most twitters' sentiment on covid vaccine is neutral in both developed country and developing country, followed by positive twitters. The developed country twitters are more than developing country twitters.

The simple visualization doesn't give us a clear view on whether there is difference on twitter's sentiment between developed country and developing country. Therefore, the regression method is applied to further investigate this hypothesis.

Method 2: two sample t-test

The two sample t-test results shows $F= 1.19$ and p-value much smaller than significant level 0.05. Thus, we can conclude the developed country average sentiment is significantly different from developing sentiment score.

Method 3: One-way ANOVA

Furthermore, this part use one-way ANOVA method to check whether country category have different effects on people's sentiment of COVID-19 vaccine. The covid-19 vaccine dataset is used in this hypothesis.

The one-way ANOVA results show F value is 8.919 and P-value is 0.0028. P-value is much smaller than 0.05 threshold. Thus, there is a statistical difference between developing country and developed country's sentiments on Covid-19 vaccine.

Method 4: ordinary logistic regression

Table 24: Ordinary logistic model

variables	coef.		Odds Ratio
country : developing	0.070	**	1.07339
Intercepts:			
negative neutral	-1.505	***	
neutral positive	0.896	***	
Regression statistics			
Residual Deviance	162800.690		

The ordinary output shows for people's twitters from developing country, the log odds of being negative sentiment (versus neutral or positive) is actually 0.070 points lower than twitters in developed country. The formula for the first and second category is as follows:

$$\text{logit}(P(Y \leq 1)) = -1.505 - 0.070 * \text{country}_{\text{developing}}$$

$$\text{logit}(P(Y \leq 2)) = 0.896 - 0.070 * \text{country}_{\text{developing}}$$

$\exp(-0.070) = 0.932$, this means developing country have 93.2% lower odds of being less positive sentiment twitter about Covid-19 vaccine compared to developed country. The R output odd ratio $\exp(\beta_1) = 1.073 = \frac{p_0/(1-p_0)}{p_1/(1-p_1)} = 1/\exp(-0.070)$. This means twitters in developing country have 1.073 times the odds of being positive (vs. neutral or negative) compared to twitters from developed countries.

Overall, sentiments of vaccination twitters in developing country category and in developed country category are different. People in the developing countries hold more positive sentiment towards COVID -19 vaccination is approximately 7.3% higher than people in the developed countries. Therefore, hypothesis 2 does not hold.

5.8 Hypothesis 3

Hypothesis 3: *people's sentiments are more positive on COVID-19 vaccine when COVID-19 confirmed cases or death increase.*

Method 1: multiple regression

This hypothesis tries to investigate effects of the confirmed new cases and death on people's sentiments on Twitter.

Table 25: correlation coefficient of independent variables

	new_cases	total_cases	new_deaths	total_deaths
new_cases	1.00	0.48	0.85	0.43
total_cases	0.48	1.00	0.60	0.98
new_deaths	0.85	0.60	1.00	0.56
total_deaths	0.43	0.98	0.56	1.00

Before performing the multiple regression, the correlation coefficients of independent variables are calculated and presented in table 25 matrix. The total confirmed cases and total death are highly correlated. The new deaths and new cases are also highly correlated with 0.85 correlation coefficient. The new cases and total death correlation coefficient is 0.43, which allows us to include both of them in the multiple regression.

Based on hypothesis 2 output, the country category has effects on the twitter sentiment. Therefore in this part, the country category is a categorical variable in the following regression models. The multiple regressions are performed based on new confirmed cases, total confirmed cases, new deaths, and total deaths separately. Table 26 shows the regression output. Model 1 (M1) tests the effects of new confirmed cases on sentiment score. Model 2 (M2) tests the effects of total confirmed COVID-19 cases on twitter sentiment score. Model 3 (M3) uses new deaths and Model 4 (M4) uses total death. Model 5 (M5) uses both new cases and total deaths as independent variables.

Table 26: regression output

	M1	M2	M3	M4
variables	coef.	coef.	coef.	coef.
constant	0.051 ***	0.049 ***	0.053 ***	0.051 ***
country category				
developing	0.009 **	0.011 ***	0.008 **	0.009 **
new confirmed cases	2.417E-08			
total confirmed cases		2.579E-10 *		
new deaths			2.90E-07	
total deaths				1.941E-07 **
Regression statistics				
R-squared	0.0001	0.0002	0.0001	0.0001

Adjusted R-squared	0.0001	0.0002	0.0000	0.0001
--------------------	--------	--------	--------	--------

*** = significant at 0.1%; ** = significant at 1%; * = significant at 5%; . =significant at 10%

The result shows the total deaths has significantly positive effects on people’s sentiment score on covid vaccine at 5% significant level. One unit increase in total death leads to 1.941E-07 increase in sentiment score based on M4. The total confirmed cases’ coefficient is significant at 10% significant level. One unit increase in total confirmed cases leads to 1.941E-07 increase in sentiment score based on M2.

Table 27: regression output

variables	M5 coef.	
constant	0.046	***
country category		
developing	0.018	**
new confirmed cases	7.666E-07	
total confirmed cases		
new deaths		
total deaths	1.815E-08	**
Regression statistics		
R-squared	0.0002	
Adjusted R-squared	0.0001	

*** = significant at 0.1%; ** = significant at 1%; * = significant at 5%; . =significant at 10%

Table 27 shows Model 5's regression output. The result shows one unit increase of total deaths will significantly increase people’s sentiment score on covid vaccine by 1.815E-08 at 5% significant level. The new confirmed cases’ coefficient is not significant. This implies when the total deaths due to Covid-19 increases, people tend to hold more positive sentiment on vaccine.

Method 2: ordinary logistic regression

Before performing the ordinary logistic regression, we should first check the assumptions. The sentiment label has three ordinary levels. There is no multicollinearity. Each observation is independent. The proportional odds assumption still need to be tested. There are available packages in R that can be used to test this assumption. However, Harrell has criticized in 2001 that those packages might get the wrong results, which will reject the null hypothesis that the sets

of coefficients are the same. In order to test the parallel slopes assumption, Harrell suggested to use a graphical method to visually test this assumption. The linear predicted values from logit model is displayed in the plot. In order to create this graph, the Hmisc package is used. We will plot a graph showing the predicted logits from each logistic regressions with single predictor.

The first step is to create a function to estimate the logit value to be plotted. The dependent variable has three levels. So we created the function containing three levels sentiment label variable with log odds greater than or equal to 1, log odds greater than or equal to 2, log odds greater than or equal to 3. The probability is transformed into logit by “qlogis” function. Then we can check the parallel slopes assumption by several binary logistic regressions with different levels of independents and check equality of the different levels coefficients. The first set of coefficients are treated as reference point and normalized to be zero for better visualization. The results are presented in Figure 25. The plot shows the distances between the coefficients for the sets for all the independent variables are similar. This suggests that the proportional odds assumption holds for all the independent variables (Harrell, 2001).

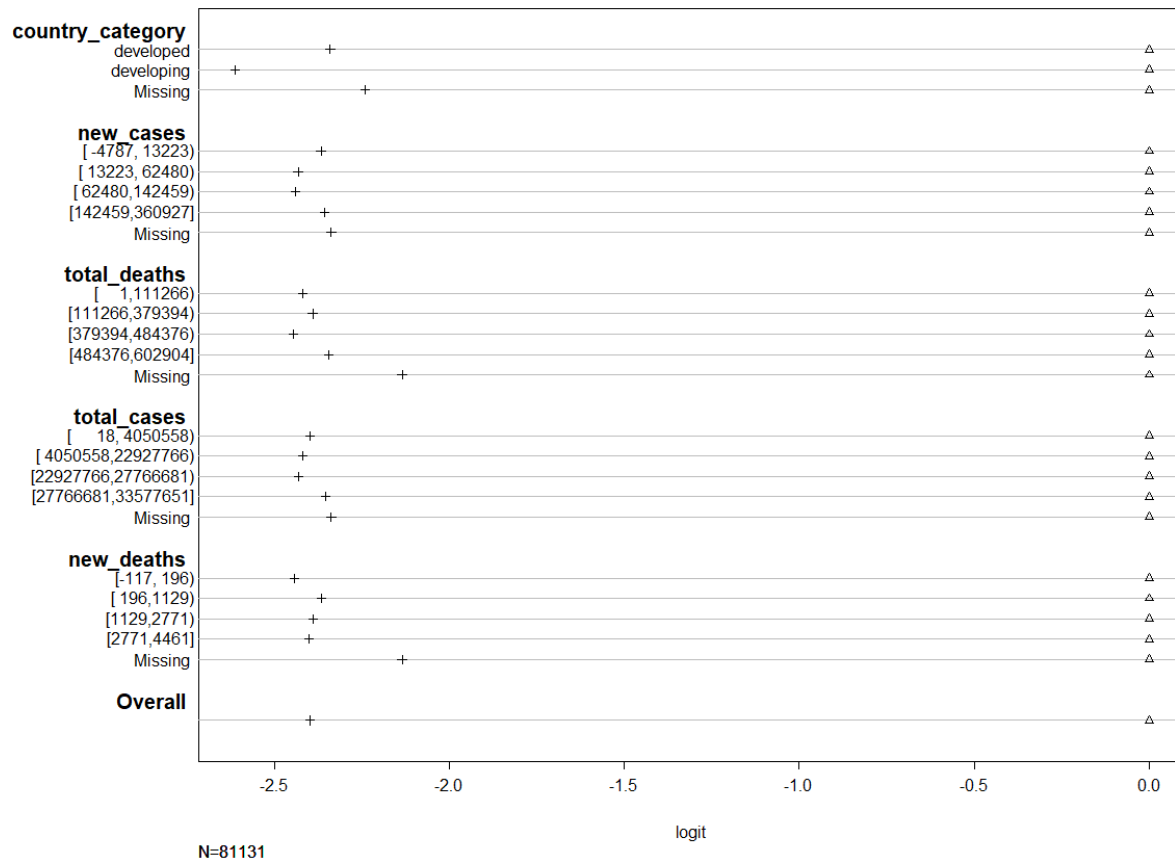


Figure 25: logit model plot for proportional odds assumption

All the assumptions of the ordinary logistic regression are met. So the next step is to estimate the model. The “polr” function in “MASS” package is used in this paper.

Table 28: Ordinary logistic model

	M1		M2		M3		M4	
variables	coef.		coef.		coef.		coef.	
country : developing	0.079	***	0.088	***	0.077	***	0.098	***
scale(new_cases)	0.017	**						
scale(total_cases)			0.020	***				
scale(new_deaths)					0.009			
scale(total_deaths)							0.026	***
Intercepts:								
negative neutral	-1.503	***	-1.501	***	-1.503	***	-1.499	***
neutral positive	0.898	***	0.900	***	0.898	***	0.903	***

Regression statistics

Residual Deviance	162692.74	162690.91	162653.85	162644.09
AIC	162700.74	162698.91	162661.85	162652.09

*** = significant at 1%; ** = significant at 5%; * = significant at 10%;

New case, total cases, new deaths, and total deaths are scaled. Table 28 shows the ordinary logistic regression results. The p-value is calculated by t-value against the standard normal distribution, generated based on the R output. The residual Deviance and AIC can be used to compare the models.

The results in table 28 can be written as the following estimated models.

Model 1 (M1):

$$\text{logit}(P(Y \leq 1)) = -1.503 - 0.017 * \text{new cases (scaled)} - 0.079 * \text{developing country}$$

$$\text{logit}(P(Y \leq 2)) = 0.898 - 0.017 * \text{new cases (scaled)} - 0.079 * \text{developing country}$$

Model 2 (M2):

$$\text{logit}(P(Y \leq 1)) = -1.501 - 0.020 * \text{total cases (scaled)} - 0.088 * \text{developing country}$$

$$\text{logit}(P(Y \leq 2)) = 0.900 - 0.020 * \text{total cases (scaled)} - 0.088 * \text{developing country}$$

Model 3 (M3):

$$\text{logit}(P(Y \leq 1)) = -1.503 - 0.009 * \text{new deaths (scaled)} - 0.077 * \text{developing country}$$

$$\text{logit}(P(Y \leq 2)) = 0.898 - 0.009 * \text{new deaths (scaled)} - 0.077 * \text{developing country}$$

Model 4 (M4):

$$\text{logit}(P(Y \leq 1))$$

$$= -1.499 - 0.026 * \text{total deaths (scaled)} - 0.098 * \text{developing country}$$

$$\text{logit}(P(Y \leq 2)) = 0.903 - 0.026 * \text{total deaths (scaled)} - 0.098 * \text{developing country}$$

The confidence intervals (CI) and odd ratios are calculated for all the models and displayed in table 29. For new cases, we can say that for one unit increase in scaled new cases, given all other variables held constant, we expect a 0.017 increase in expected value of sentiment label on the

log odds scale. To better interpret the output, we convert the coefficients into odd ratios. For one unit increase in scaled new cases, the odds of having more positive sentiment twitters (positive or neutral sentiment versus negative sentiment) is multiplied 1.017 times (increase 17%), given all other variables held constant. For one unit decrease in scaled new cases, the odds of having less positive sentiment twitters (negative sentiment versus neutral or positive sentiment) is multiplied 1.017 times, given all other variables held constant. Similar interpretation can be applied to the other 3 models. Therefore, the hypothesis three holds. People's sentiments are more positive on COVID-19 vaccine when COVID-19 confirmed cases or death increase.

Table 29: odd ratios and CI of the ordinary logistic model

		odd ratios	2.50%	97.50%
M1	country:developing	1.082	0.047	0.111
	scale(new_cases)	1.017	0.003	0.030
M2	country:developing	1.092	1.056	1.129
	scale(total_cases)	1.020	1.006	1.035
M3	country:developing	1.080	1.045	1.117
	scale(new_deaths)	1.009	0.995	1.023
M4	country:developing	1.103	1.065	1.143
	scale(total_deaths)	1.026	1.011	1.041

Both multiple linear regression and ordinary logistic regression suggests people's sentiments are more positive on COVID-19 when confirmed cases (new confirmed or total confirmed) or increase of COVID-19 death (new death or total death) increase. Therefore, this hypothesis holds true.

5.9 Hypothesis 4

Hypothesis 4: *With the increased number of vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive.*

Method 1: Multiple regression

Table 30 is the correlation coefficients matrix used to test the multicollinearity between independent variables. If the correlation coefficients is too large (larger than 0.7), there is multicollinearity and the two variables are highly correlated. Therefore, 7 regression models are generated and the results are displayed in table 30 and table 31.

Table 30: correlation coefficients

	total_vaccinations	new_vaccinations	total_vaccinations_per_hundred	new_vaccinations_smoothed_per_million
total_vaccinations	1.000	0.606	0.762	0.418
new_vaccinations	0.606	1.000	0.267	0.408
total_vaccinations_per_hundred	0.762	0.267	1.000	0.687
new_vaccinations_smoothed_per_million	0.418	0.408	0.687	1.000

Table 31: multiple regression

	M6	M7	M8	M9
	Estimate	Estimate	Estimate	Estimate
(Intercept)	0.0496 **	0.0505 **	0.0460 **	0.0461 **
country_category				
developing	0.0035	0.0049	0.0093 *	0.0085 *
total_vaccinations	4.76E-11 **			
new_vaccinations		1.19E-09		
total_vaccinations_per_hundred			0.0003 **	
new_vaccinations_smoothed_per_million				1.58E-06 **
Regression statistics				
R-squared	0.0001625	0.00005483	0.000428	0.00015
Adjusted R-squared	0.0001298	0.00001945	0.000395	0.00012

*** = significant at 0.1%; ** = significant at 1%; * = significant at 5%; . = significant at 10%

M6 and M8 shows there is significantly positive effects of total vaccination numbers on the sentiment scores. M9 shows there is significant positive effects of new vaccinations smoothed per million on the sentiment scores too. However, because the coefficients value is quite small, the effects are relatively small.

Table 32: multiple regression

	M10		M11		M12	
	Estimate		Estimate		Estimate	
(Intercept)	0.0450	***	0.0458	***	0.0498	***
country_category						
developing	0.0125	**	0.0076		0.0053	
total_vaccinations			3.06E-11		6.75E-11	**
new_vaccinations	-1.07E-09				-1.71E-09	
total_vaccinations_per_hundred	0.0003	***				
new_vaccinations_smoothed_per_million			1.06E-06			
Regression statistics						
R-squared	0.0005862		0.0002034		0.00024	
Adjusted R-squared	0.0005331		0.0001538		0.00019	

*** = significant at 0.1%; ** = significant at 1%; * = significant at 5%; . = significant at 10%

M10 to M12 shows the multiple regression output by adding two vaccine related independent variables. The results for the total vaccine (or total vaccinations per hundred) are significant indicating twitter's sentiment score increase when the total numbers of COVID-19 vaccinations increase.

Method 2: ordinary logistic regression

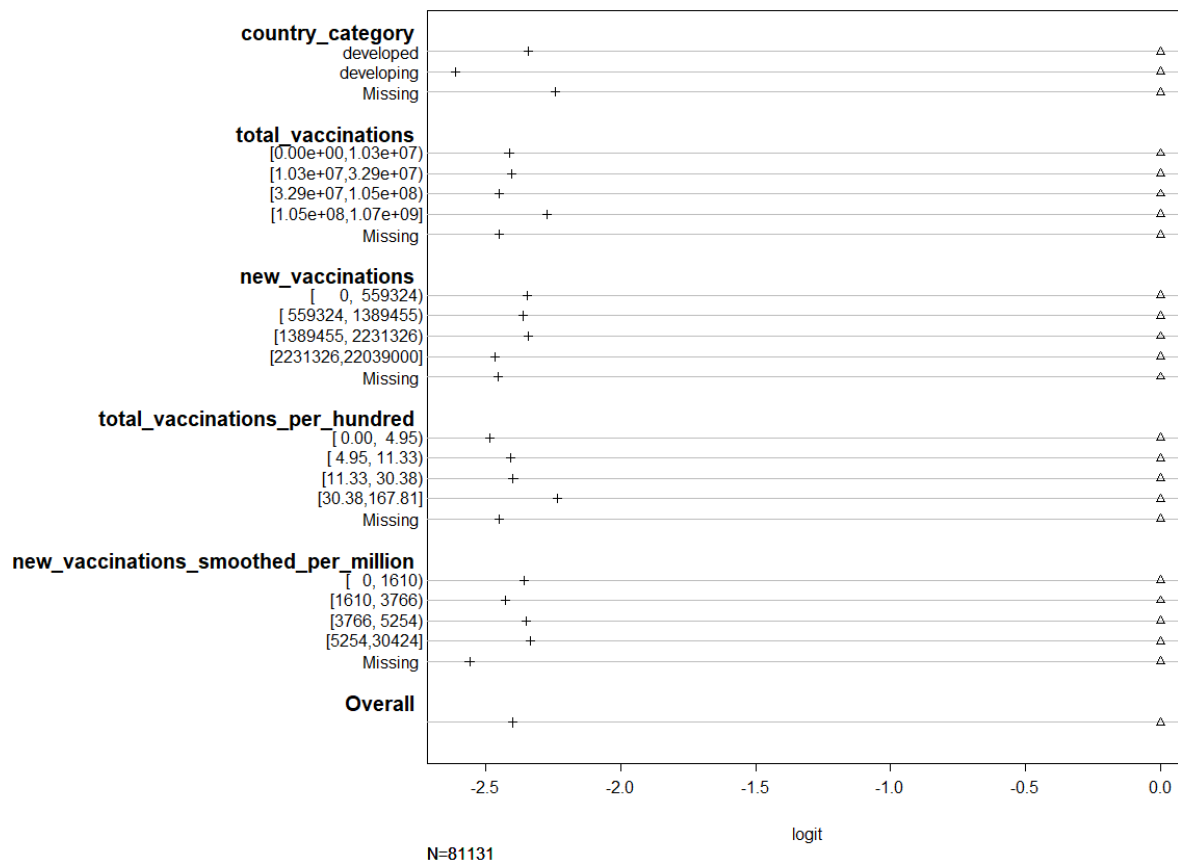


Figure 26: proportional odds assumption

Similar to the previous hypothesis, to perform the ordinary logistic regression, we need to first check the proportional odds assumption. Figure 26 shows for all vaccination variables, the distance between two sets of coefficients is similar. This suggests the proportional odds assumption holds for all variables.

Table 33: Ordinary logistic model

	M1		M2		M3		M4	
variables	coef.	p value	coef.	p value	coef.	p value	coef.	p value
country : developing	0.042	**	0.049	**	0.068	***	0.060	***
scale(total vaccinations)	0.018	**						
scale(new vaccinations)			0.008					
scale(total vaccinations per hundred new vaccinations smoothed per million)					0.031	***		
							0.000	*
Intercepts:								

negative neutral	-1.489 ***	-1.483 ***	-1.486 ***	1.451 ***
neutral positive	0.895 ***	0.895 ***	0.899 ***	0.918 ***
Regression statistics				
Residual Deviance	123004.39	114095.37	122995.04	134985.05
AIC	123012.39	114103.37	123003.04	134993.05

*** = significant at 1%; ** = significant at 5%; * = significant at 10%;

After assessing the assumptions, the ordinary logistic regression is performed and the regression output is presented in table 33 .

We can rewrite the output into the following estimated models.

Model 1 (M1):

$\text{logit}(P(Y \leq 1))$

$$= -1.489 - 0.018 * \text{total vaccinations (scaled)} - 0.042 * \text{developing country}$$

$\text{logit}(P(Y \leq 2))$

$$= 0.895 - 0.018 * \text{total vaccinations (scaled)} - 0.042 * \text{developing country}$$

Model 2 (M2):

$\text{logit}(P(Y \leq 1))$

$$= -1.483 - 0.008 * \text{new vaccinations (scaled)} - 0.049 * \text{developing country}$$

$\text{logit}(P(Y \leq 2))$

$$= 0.895 - 0.008 * \text{new vaccinations (scaled)} - 0.049 * \text{developing country}$$

Model 3 (M3):

$$\begin{aligned} \text{logit}(P(Y \leq 1)) &= -1.486 - 0.031 * \text{total vaccinations per hundred (scaled)} - 0.068 \\ &\quad * \text{developing country} \end{aligned}$$

$$\begin{aligned} \text{logit}(P(Y \leq 2)) &= 0.899 - 0.031 * \text{total vaccinations per hundred (scaled)} - 0.068 \\ &\quad * \text{developing country} \end{aligned}$$

Model 4 (M4):

$$\begin{aligned} \text{logit}(P(Y \leq 1)) &= -1.451 - 0.000 * \text{new vaccinations smoothed per million} - 0.060 \\ &\quad * \text{developing country} \end{aligned}$$

$$\begin{aligned} \text{logit}(P(Y \leq 2)) &= 0.918 - 0.000 * \text{new vaccinations smoothed per million} - 0.060 \\ &\quad * \text{developing country} \end{aligned}$$

Table 34: odd ratios and CI of the ordinary logistic model

model	Variables	odd ratios	2.50%	97.50%
M1	country:developing	1.043	1.000	1.087
	scale(total vaccinations)	1.018	1.002	1.034
M2	country:developing	1.051	1.005	1.098
	scale(new vaccinations)	1.008	0.992	1.024
M3	country:developing	1.070	1.025	1.117
	scale(total vaccinations per hundred)	1.032	1.015	1.048
M4	country:developing	1.062	1.016	1.110
	new vaccinations smoothed per million	1.000	1.000	1.000

Table 34 shows the confidence intervals (CI) and odd ratios are calculated for all the models. For one unit increase in scaled total vaccination, the odds of having more positive sentiment twitters (positive or neutral sentiment versus negative sentiment) is multiplied 1.043 times (increase 4.3%), given all other variables held constant. Similar interpretation can be applied to M2 and M3. For M4, the odd ratios is approximately equal to 1.000. This means for one unit increase in

scaled total vaccination, the odds of having more positive sentiment twitters is approximately not change (or slightly change), given all other variables held constant.

Overall, both multiple regression results and ordinary logistic regression results indicates with the increased number of vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive. Therefore, hypothesis 4 also holds true.

5.10 Hypothesis 5

Hypothesis 5: People are more optimistic about COVID-19 vaccination when the government policies are stricter, such as closedown schools.

Method 1: Multiple regression

Sentiment score = $\alpha + \beta_1 * \text{country category} + \beta_2 * \text{total vaccinations} + \beta_3 * \text{School closing category} + \beta_4 * \text{working place closing} + \beta_5 * \text{cancel public event} + \beta_6 * \text{restriction on gathering} + \beta_7 * \text{closing public transport} + \beta_8 * \text{stay at home requirements} + \beta_9 * \text{restrictions on internal movements} + \beta_{10} * \text{international travel controls} + \beta_{11} * \text{vaccine policy} + e$

Table 35: multiple regression

Variables	Estimate	Sig
(Intercept)	0.06	
country_categorydeveloping	0.02	.
scale(total_vaccinations)	0.00	
C1_School_closing1	0.02	
C1_School_closing2	0.01	
C1_School_closing3	0.02	
C2_Workplace_closing1	-0.04	
C2_Workplace_closing2	-0.03	
C2_Workplace_closing3	-0.03	
C3_Cancel_public_events1	-0.11	*
C3_Cancel_public_events2	-0.11	*

C4_Restrictions_on_gatherings1	0.27	*
C4_Restrictions_on_gatherings2	0.08	.
C4_Restrictions_on_gatherings3	0.06	
C4_Restrictions_on_gatherings4	0.09	*
C5_Close_public_transport1	0.04	***
C5_Close_public_transport2	0.00	
C6_Stay_at_home_requirements1	0.01	
C6_Stay_at_home_requirements2	-0.01	
C6_Stay_at_home_requirements3	0.00	
C7_Restrictions_on_internal_movement1	-0.01	
C7_Restrictions_on_internal_movement2	-0.01	
C8_International_travel_controls2	0.04	.
C8_International_travel_controls3	0.00	
C8_International_travel_controls4	-0.02	
H7_Vaccination_policy1	0.01	
H7_Vaccination_policy2	0.01	
H7_Vaccination_policy3	0.00	
H7_Vaccination_policy4	0.00	
H7_Vaccination_policy5	0.02	

Regression statistics

R-squared	0.002
Adjusted R-squared	0.001

*** = significant at 0.1%; ** = significant at 1%; * = significant at 5%; . = significant at 10%

Table 35 shows the multiple regression output. Compared to no measures, all levels of work place closing measurement, cancel public events, internal movement restrictions will have negative effects on the sentiment score. Compared to no measures, all levels of school closing, restriction on gathering, close public transport policies, vaccination policy are positively affects peoples' sentiment. The effects of stay at home requirement policy and internal movement restrictions, internal travel controls on sentiment score are mixed. When some policies becoming stricter, their effects on sentiment scores varies across different type of policies. Let's use

vaccination policy output as an example. Vaccination policy records the policies for vaccine delivery for different groups. The baseline in this model is 0, indication no availability for all groups. For 1st level it is available to one of the groups (key workers, clinically vulnerable groups or elderly groups). For 2nd level it is available to two of the groups (key workers, clinically vulnerable groups or elderly groups). Higher number indicating larger groups available to be vaccinated. The regression shows compared to not available for all groups, one group and two group available policies have positive effects on sentiment score. 3rd level and 4th level have no effects on sentiment score compared to not available policy. When the vaccination is universal available (at 5th level), the effects on sentiment score is 0.02 units higher than not available. All the estimates are not statistically significant. Similar interpretation can applied to other policy variables. Overall, the multiple regression results shows, stricter government policies will have mixed effects on people’s sentiments on covid-19 vaccine.

Method 2: Ordinary logistic regression

Table 36: proportional odds assumption

variables	neu_plus	pos	diff
(Intercept)	9.320	0.050	-9.270
country_categorydeveloping	0.258	0.002	-0.256
C1_School_closing1	0.014	0.080	0.066
C1_School_closing2	0.174	-0.001	-0.175
C1_School_closing3	0.166	0.074	-0.092
C2_Workplace_closing1	-0.458	-0.221	0.237
C2_Workplace_closing2	-0.381	-0.148	0.233
C2_Workplace_closing3	-0.335	-0.074	0.261
C3_Cancel_public_events1	0.132	-0.365	-0.497
C3_Cancel_public_events2	0.175	-0.477	-0.652
C4_Restrictions_on_gatherings1	1.107	0.962	-0.145
C4_Restrictions_on_gatherings2	0.712	0.675	-0.038
C4_Restrictions_on_gatherings3	0.478	0.540	0.062
C4_Restrictions_on_gatherings4	0.511	0.505	-0.005

C5_Close_public_transport1	0.111	0.056	-0.055
C5_Close_public_transport2	0.221	-0.040	-0.261
C6_Stay_at_home_requirements1	-0.087	0.091	0.178
C6_Stay_at_home_requirements2	-0.171	0.030	0.201
C6_Stay_at_home_requirements3	-0.128	0.015	0.142
C7_Restrictions_on_internal_movement1	-0.031	-0.056	-0.025
C7_Restrictions_on_internal_movement2	-0.123	-0.022	0.100
C8_International_travel_controls1	-8.152	-1.069	7.083
C8_International_travel_controls2	-8.016	-0.899	7.116
C8_International_travel_controls3	-8.088	-0.922	7.166
C8_International_travel_controls4	-8.215	-1.044	7.171
H5_Investment_in_vaccines	0.012	-0.003	-0.015
H7_Vaccination_policy1	-0.029	-0.014	0.015
H7_Vaccination_policy2	0.011	-0.030	-0.041
H7_Vaccination_policy3	-0.156	-0.098	0.058
H7_Vaccination_policy4	-0.379	0.089	0.468
H7_Vaccination_policy5	-0.300	0.196	0.496

To check the proportional odds assumption, we performed stratified binomial models on the data to check the difference of the coefficients of the independent variables. Two binomial logistic regression model for both neutral and positive sentiment and for positive sentiment are created and the coefficients of both models and their difference are presented in table 36. If the differences are large, the proportional odds assumption is likely violated. In our model, the international travel controls difference scores are too large, which violates the proportional odds assumption and we will remove it from the ordinary logistic model. Without considering the intercept, the differences for the rest coefficients are relatively small. Thus, the proportional odds assumption is hold for them and we include these variables in the ordinary logistic regression.

Table 37: Ordinary logistic model

variables	coef.
country : developing	0.070 **

C1_School_closing1	0.111	
C1_School_closing2	0.122	*
C1_School_closing3	0.159	**
C2_Workplace_closing1	-0.335	**
C2_Workplace_closing2	-0.259	*
C2_Workplace_closing3	-0.234	
C3_Cancel_public_events1	-0.197	
C3_Cancel_public_events2	-0.258	
C4_Restrictions_on_gatherings1	0.938	***
C4_Restrictions_on_gatherings2	0.681	***
C4_Restrictions_on_gatherings3	0.536	***
C4_Restrictions_on_gatherings4	0.517	***
C5_Close_public_transport1	0.074	**
C5_Close_public_transport2	0.030	
C6_Stay_at_home_requirements1	0.087	*
C6_Stay_at_home_requirements2	0.016	
C6_Stay_at_home_requirements3	0.099	
H7_Vaccination_policy1	-0.022	
H7_Vaccination_policy2	-0.019	
H7_Vaccination_policy3	-0.125	***
H7_Vaccination_policy4	-0.102	***
H7_Vaccination_policy5	0.003	
Intercepts:		
negative neutral	-1.297	***
neutral positive	1.115	***
Regression statistics		
Residual Deviance	160482.65	
AIC	160532.65	

*** = significant at 1%; ** = significant at 5%; * = significant at 10%;

Table 37 shows ordinary logistic model output. Based on the output, restriction on gathering coefficients is statistically significant at 1% significant level for all categories. So, for School closing policy, we expect a 0.111 increase in the expected value of apply on the log odds scale for school closing at 1st level compared to no measures, given all the other variables in the model are held constant. For workplace closing policy, the effects are negative. In other words, given all other variables in the model held constant, compared to not closing, 1st level of closing is expected to have 0.335 decrease in the expected value on the log odds scale. Same as previous hypothesis, the estimates have two cutpoints (or intercepts). Neg vs neu has negative intercepts and neu vs pos have positive intercept.

The $X_{m,n}$ represents different independent variables at different category levels. $\beta_{m,n}$ is the coefficients of $X_{m,n}$. Therefore, we can rewrite the estimated models as follows:

$$\text{logit}(P(Y \leq 1)) = -1.297 - \beta_{m,n} * X_{m,n}$$

$$\text{logit}(P(Y \leq 2)) = 1.115 - \beta_{m,n} * X_{m,n}$$

Table 38: odd ratio and CI of the ordinal logistic regression

variables	OR	2.50%	97.50%
country_categorydeveloping	1.073	1.005	1.145
C1_School_closing1	1.117	0.968	1.290
C1_School_closing2	1.130	0.981	1.301
C1_School_closing3	1.172	1.020	1.347
C2_Workplace_closing1	0.715	0.536	0.955
C2_Workplace_closing2	0.772	0.583	1.021
C2_Workplace_closing3	0.792	0.596	1.052
C3_Cancel_public_events1	0.821	0.598	1.126
C3_Cancel_public_events2	0.772	0.567	1.051
C4_Restrictions_on_gatherings1	2.554	1.339	4.874
C4_Restrictions_on_gatherings2	1.975	1.473	2.649
C4_Restrictions_on_gatherings3	1.710	1.283	2.280
C4_Restrictions_on_gatherings4	1.678	1.260	2.233

C5_Close_public_transport1	1.076	1.009	1.148
C5_Close_public_transport2	1.031	0.951	1.117
C6_Stay_at_home_requirements1	1.090	0.990	1.201
C6_Stay_at_home_requirements2	1.017	0.913	1.131
C6_Stay_at_home_requirements3	1.104	0.968	1.259
H7_Vaccination_policy1	0.978	0.919	1.041
H7_Vaccination_policy2	0.981	0.931	1.034
H7_Vaccination_policy3	0.882	0.832	0.935
H7_Vaccination_policy4	0.903	0.846	0.964
H7_Vaccination_policy5	1.003	0.906	1.109

To better interpret the output, we convert the coefficients into odd ratios. From Table 38 we can see, compared to not closing , when the government have the 1st level of school closing, the odds of having more positive sentiment twitters (positive or neutral sentiment versus negative sentiment) is multiplied 1.117 times, given all other variables held constant. Compared to not closing , when the government have the 1st level of workplace closing, the odds of having more positive sentiment twitters is multiplied 0.715 times, given all other variables held constant. In other words, when the odd ratios is larger than 1, the odds of having more positive sentiments twitter is higher compared to the baseline. Otherwise, it will be lower compared to the baseline. Restrictions on gathering has the highest odd ratios, indicating when there is restrictions on gathering, the odds of having more positive sentiments will largely increase compared to no restriction gathering. The odd ratios for workplace closing and close public transport and vaccinations for most of the levels are smaller than 1, indicating compared to no measurement policy, introducing any levels of these types of policy will lead to less positive sentiment on Covid-19 vaccine. Introducing the government policy such as school closing, close public transport, gatherings restrictions, stay at home requirements will lead to more positive sentiment on covid-19 vaccination compared to the without related without measurement policies. Based on ordinary logistic regression result, different government policies will have different effects on the twitter sentiments. The effects of different levels of the different policies are different as well.

Both multiple regressions based on the sentiment score and ordinary logistic regression based on the sentiment label show that different types of government policies affect the sentiment differently. When the Covid -19 related government policies are stricter, people sentiment varies as well. Some policies might bring more positive sentiments after introducing them at a small level. Some policies might affect the sentiments positively after reaching certain levels. Therefore, based on the evidence discovered in this session, this hypothesis does not hold.

Chapter 6: Conclusion and limitation

This study focuses on investigating the COVID-19 vaccine-related twitters sentiments. The research questions are built and tested by different techniques and different hypotheses. Sentiments differences across different brands of vaccines for different country categories are analyzed. This paper used five datasets to analyze the sentiments and exam the hypothesis, including Pfizer vaccine-related twitters (from December 12th 2020 till June 23rd , 2021), Astrazen vaccine-related twitters (13th August, 2020 till June 23rd , 2021), COVID-19 vaccine-related twitters (from August, 2020 till June 23rd , 2021), Covid-19 statistic dataset (end of February, 2020 to beginning of July, 2021), and COVID-19 related government policy dataset (January 2020 to beginning of July, 2021). This research's findings could help governments, public health officials, and policymakers better understand people's attitudes towards vaccines. This session will first discuss the sentiment analysis findings. Second, five different hypotheses will be discussed separately. The last part will discuss the limitations of this paper and suggestions for further research.

Sentiment analysis

This paper uses unsupervised and supervised sentiment classification techniques to further the sentiment analysis on three Covid-19 vaccination twitters datasets. The VADER lexicon-based method is used to obtain the polarity of the sentiments. The Naïve Bayes and Support Vector machine methods are performed for each dataset. The results show that the Naïve Bayes model's accuracy for all datasets ranges between 70% to 74%. SVM machine method outperforms NB technique in all datasets. The accuracy of SVM models for Pfizer dataset is 86.7% and for Covid-19 vaccine, dataset is nearly 95%. Due to computation limitations, the sentiment classification for Covid-19 vaccine dataset based on Naïve Bayes method needs to split the dataset into four equal-length samples. The accuracy for all the sample sets is below 73%. When applying the SVM, the accuracy of one of the four samples is approximately 92%.

Hypothesis 1

First hypothesis is *people's sentiments on different brand of vaccine are different.*

This hypothesis uses the two sample t test, ANOVA and Chi-test methods to test whether people hold different sentiments on Pfizer and Astrazen brands vaccine. All the results show there is a statistical difference between different brands of the vaccine.

Shamrat, etc.' s paper published in June 2021 analyzed people's sentiment on three different brands of COVID-19 vaccines, including Pfizer, Moderna, and AstraZeneca vaccines based on twitters data. In their paper, they use KNN classification algorithm to classify the sentiment. Hypothesis 1 results are in line with Shamrat, etc.' s paper's results. The results show generally, people have higher positive sentiment towards Pfizer and Moderna vaccine compared to the AstraZeneca vaccine. This hypothesis and Shamrat, etc.' s paper's findings could help the government improve the vaccination brand strategy to increase the trust and acceptance of the public towards COVID-19 vaccine.

Hypothesis 2

The second hypothesis is *people in the developed countries are more positive towards COVID - 19 vaccination than people in the developing countries.*

This hypothesis use graph visualization, two sample t-test, one -way ANOVA and ordinary logistic regression method on Covid-19 vaccine dataset. These results show differences in people's sentiment on Covid-19 vaccine between developing countries and developed countries. People in developing countries are more positive towards COVID-19 vaccination than people in developing countries. This is against the original hypothesis derived based on Van Essen's research in 2003.

Yoda and Katsubama examined people's willingness to receive COVID-19 vaccination in Japan in recent research in 2021. They used descriptive statistics and chi-square test to perform the analysis. Their results show 65.7% of participants are willing to be vaccinated. Neumann-Böhme, etc.' s research in 2020 shows people's willingness to be vaccinated in European countries such as the UK, Netherlands, and Denmark, are all above 70% (some countries reach 80%). These two papers show people's willingness to be vaccinated in developed countries instead of their attitudes towards the COVID-19 vaccine. At a certain standard people's willingness to be vaccinated also reflects their sentiments towards vaccination. More positive attitudes towards vaccines could partly explain the high willingness to be vaccinated. Both researches show people's willingness to be vaccinated in developed countries is not low.

Chen, etc.'s research in 2021 investigated people's willingness to accept COVID-19 vaccine among Chinese adults based on a cross-sectional survey. Their results show 83.8% of participants are willing to receive COVID-19 vaccine. 76.6% of participants believe vaccination would be beneficial to their health. Compared with the results obtained from the other two research, people's willingness to receive COVID-19 vaccine is relatively higher than people's willingness in developed countries such as Japan, UK, or Netherlands. China belongs to the developing country category. This could be an indication to support the finding of hypothesis 2 based on this research. However, it might be biased if we only check on developing countries based on one paper's findings. Due to the lack of published research on the topic of this hypothesis. It is hard to find other studies to evaluate the accuracy of this hypothesis.

Hypothesis 3

The third hypothesis is *people's sentiments are more positive on COVID-19 vaccine when COVID-19 confirmed cases or death increase.*

Multiple linear regression and ordinary logistic regression methods are applied to test this hypothesis. The COVID-19 vaccine twitter dataset and the Covid-19 general statistics dataset are merged based on the date and location in this hypothesis. The total confirmed cases, new confirmed cases, new death and total death are used to understand their effects on people's attitude toward vaccine. The results show the relationship is indeed positive. In other words, people indeed hold more positive attitudes on vaccines when the COVID-19 confirmed cases or death increase.

Niu's research in July 2021 investigated the opinions and sentiments of COVID-19 vaccination-related tweets between August 2020 and June 2021 from Japanese Twitter users. They performed sentiment analysis by using Amazon Web Service and generated the correlation between sentiments and the number of deaths, infections and vaccinations. Their results show before and after the first vaccination in Japan. The correlations of sentiment with death, infection and vaccination changed significantly. They found death and infection significantly correlated (0.69) with negative sentiments before the first vaccination in Japan. However, we still cannot find enough evidence to further support our findings based on their finding.

Shim etc.'s research in June 2021 analyzed the COVID-19 vaccine sentiments among Korean public response. Their results show after the increase in the number of confirmed cases, the negative tweets are prominent. This does not match with the findings in this hypothesis. This might be because they used only tweets in Korean from the period 21st February 2021 until 22nd March 2021. The data is only limited to one country and for one month period. It might not be possible to be generalized. Due to the limited published papers related to this hypothesis, we cannot find evidence to support or against this hypothesis.

The finding based on this hypothesis could help governments better design the Covid-19 vaccination strategy, such as the supply of vaccinations that reacted to the change of confirmed cases or deaths due to COVID-19. Society and government can also better respond to the public and help people to gain trust in the vaccination during the COVID-19 peak period.

Hypothesis 4

The fourth hypothesis is *With the increased number of vaccinations, people's sentiments on COVID-19 vaccine tend to be more positive.*

Similar to hypothesis 3, we use the multiple linear regression and ordinary logistic regression method and COVID-19 vaccine twitter dataset and the Covid-19 general statistics dataset to investigate this hypothesis. The results shows the hypothesis 4 holds true, which means people indeed are more positive towards vaccine when the number of vaccination (injected) increases. Niu, etc.'s research in 2021 found the correlation between vaccination cases and positive sentiment is 0.532. The correlation between vaccination cases and negative sentiment is 0.575. Their finding indicates there is a relationship between the number of vaccinations and people's sentiment. We cannot find other studies to further this hypothesis.

Hypothesis 5

The 5th hypothesis is *people are more optimistic about COVID-19 vaccination when the government policies are stricter, such as closedown schools, etc.*

COVID-19 vaccine twitter dataset and government policy dataset are used in this hypothesis. In this hypothesis, nine policies are used to investigate their effects on people's sentiment. These policies includes: school closing, working place closing, cancel public event, restriction on gathering, closing public transport, stay at home requirements, restrictions on internal

movements, international travel controls and vaccine policy. Similar to the previous hypothesis, the multiple linear regression and ordinary logistic regression method are performed to examine the hypothesis. The outcome shows different government policies have different effects on the people's sentiments. There is not enough evidence to show that stricter policies lead to more positive sentiments. We cannot find other studies to further support the finding from this hypothesis due to limited research performed under this hypothesis.

Gupta and Kumar's research in 2020 studies the sentiments of Indian citizens regarding the nationwide lockdown policy. Their results show the majority of Indian citizens support the national lockdown policy. Their research indicates people might be positive towards the government policy during COVID-19 lockdown period. However, due to the lack of more related research on this hypothesis, we cannot provide further evidence to show the effects of government policies on people's sentiments towards COVID-19 vaccination.

Ali, etc.'s study in 2021 investigated Public's sentiment in US on COVID-19 vaccine by Twitters information. They suggested governments and vaccine manufacturing companies need to proactively make the policies to inform the society the reason behind rapid development of COVID-19 and necessary take the vaccines through social media platforms. This could help to increase people's trust on vaccine and increase people's positive attitudes towards vaccine.

Limitation and future research

Limitation in data and computation power

This study uses the twitters vaccine-related data from August 2020 to June 2021 (for Pfizer dataset from December 2020 to June 2021). Some dates contain missing values which makes it hard to capture the entire daily sentiment changes. Besides, due to the limitation of the computer memory, Naïve Bayes on the COVID-19 vaccine dataset need to be split into several samples, which will reduce the accuracy of the model. The location variables are not available for all the twitters. Age or gender information is not available on all twitters. This demographic information could help to better understand people's sentiments, which requires us to study further.

Future research to deeper understand the positive and negative topic about Covid-19 vaccine.

This paper only focused on understanding people's sentiment on covid-19 vaccination based on twitters data. After understanding people's sentiment, other questions could rise such as what topic brings the positive sentiment on Covid-19 vaccine and what topics brings negative sentiment on Covid-19. To understand such a question, the topic clustering and topic modeling could be applied to perform deeper analysis. Topic clustering has two common algorithms, which are the hierarchical clustering algorithm and the k-means algorithm. The basic idea of K-means clustering is the first chosen fixed number (k) of clusters. A random point for each cluster is selected and acting as the "centroid" of the algorithm. Then every document is allocated to the nearest centroid. The documents are grouped in the cluster where the distance between the document and the cluster is minimized compared to other clusters (Kwartler, 2017). k-Means clustering requires first to determine the number of clusters. However, we might not know the number of clusters at the beginning. Hierarchical clustering solves this predefined problem. It first merges the nearest clusters in hierarchical clustering. The centroids of the clusters determine the similarity. The clusters with the largest similarity have the smallest distance, which can be merged into one hierarchical cluster. There are two types of hierarchical clustering, agglomerative hierarchical clustering, and Divisive hierarchical clustering. Agglomerative hierarchical clustering first assigns each observation to one individual cluster. Then the nearest pair of clusters are merged into one cluster. This process is repeated until only one cluster is left. On the opposite, the Divisive hierarchical clustering starts with one single cluster, which contains all the observations. Then the farthest observations in the cluster are spitted into the separate cluster. This step is repeated until each cluster contains only one observation. The agglomerative hierarchical clustering and K-means clustering are most commonly applied in document clustering tasks. However, agglomerative hierarchical clustering performs better but slower than K-means (Karypis, Kumar & Steinbach, 2000).

Topic modeling is not a new technique. However, there are only a few researchers optimizing the categorizing method for research papers (Asmussen & Møller, 2019). Most of the topic modeling-related researches are for newspapers, tweets, web contents, books. For example, Jacobi, Van Atteveldt, and Welbers proposed Latent Dirichlet Allocation (LDA) in newspapers in 2016, and Guo et al. compared LDA with dictionary-based analysis based on Tweets posts in 2016. These researches applied Topic modeling because it efficiently reduced the time required.

Topic modeling aims to answer two questions. One is how to decide whether the chosen word belongs to a specific topic instead of others. The other is to identify a particular topic's frequency within a specific document (Kwartler, 2017). The most popular topic modeling method is Latent Dirichlet Allocation (LDA). LDA can automatically generate the topics based on the occurrence of the words (Jacobi, Van Atteveldt & Welbers, 2016). LDA was introduced by Blei, Ng, and Jordan in 2003. Based on them, it is the simplest method for topic modeling. LDA automatically searches the concealed group of words that belong to the same topic. The Dirichlet distribution is commonly used to learn multivariate probability distributions. LDA checks the words in each document and calculates the multi-word probability distributions between the words in the documents and obtains the pattern of the topics. Blei, Ng, and Jordan's research show the topics obtained by LDA can provide a clear representation of the documents (2003). LDA uses the "Bag of Words" method. Every word is treated as a distinctive feature of the specific document. In this case, the order of the words, the grammar, and the sentence's meaning are not considered by using "Bag of Word" approach. Therefore, the most related topic of a document is the most frequent words that appeared in the document (Kwartler, 2017). Due to this reason, the topic obtained by LDA usually not represent the full meaning of the text. However, it gives a general overview of the themes of the documents (Jockers & Mimno, 2013).

The above mentioned methods and literatures could be useful for researchers to deeper understand the reason behind people's sentiment by analyzing the related topics. The results could further help the government or researchers to better understand or react to the public.

References

- Ali, G. G., Rahman, M. M., Hossain, A., Rahman, S., Paul, K. C., Thill, J. C., & Samuel, J. (2021). Public perceptions about covid-19 vaccines: Policy implications from us spatiotemporal sentiment analytics. *Available at SSRN 3849138*.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).
- Ahmed, M. E., Rabin, M. R. I., & Chowdhury, F. N. (2020). COVID-19: Social media sentiment analysis on reopening. arXiv preprint arXiv:2006.00804.
- Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127, 511-520.
- Aschwanden, C. (2020). The false promise of herd immunity for COVID-19. *Nature*, 587(7832), 26-28.
- Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), 93.
- Bellegarda, J. R. (2005). Latent semantic mapping [information retrieval]. *IEEE Signal Processing Magazine*, 22(5), 70-80.
- Bilder, C. R., & Loughin, T. M. (2014). *Analysis of categorical data with R*. Chapman and Hall/CRC.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Chen, M., Li, Y., Chen, J., Wen, Z., Feng, F., Zou, H., ... & Sun, C. (2021). An online survey of the attitude and willingness of Chinese adults to receive COVID-19 vaccination. *Human Vaccines & Immunotherapeutics*, 17(7), 2279-2288.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- del Rio-Chanona, R. M., Mealy, P., Pichler, A., Lafond, F., & Farmer, J. D. (2020). Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy*, 36(Supplement_1), S94-S137.
- Dhingra, C. (2021). Analysis of twitter data for India and China Covid-19 Vaccination.

Ding, X., Liu, B., & Zhang, L. (2009, June). Entity discovery and assignment for opinion mining applications. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1125-1134).

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5), 533-534.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics (pp. 841-847).

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2), 1-41.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

Graffigna, G., Palamenghi, L., Boccia, S., & Barello, S. (2020). Relationship between citizens' health engagement and intention to take the covid-19 vaccine in italy: A mediation analysis. *Vaccines*, 8(4), 576.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, 93(2), 332-359.

Gupta, P., Kumar, S., Suman, R. R., & Kumar, V. (2020). Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter. *IEEE Transactions on Computational Social Systems*.

Gupte, A., Joshi, S., Gadgul, P., Kadam, A., & Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies*, 5(5), 6261-6264.

Gut, A. (2013). *Probability: a graduate course (Vol. 75)*. Springer Science & Business Media.

Harrell, F. E. (2001) *Regression Modeling Strategies*. New York: Springer-Verlag.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3), 464-472.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1), 89-106.

Jindal, N., & Liu, B. (2006, August). Identifying comparative sentences in text documents. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 244-251).

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6), 750-769.

Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, May). A comparison of document clustering techniques. In TextMining Workshop at KDD2000 (May 2000).

Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011, December). Twitter trending topic classification. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 251-258). IEEE.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

McKibbin, W., & Fernando, R. (2020). The economic impact of COVID-19. *Economics in the Time of COVID-19*, 45.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).

Neergaard L. and Fingerhut H. (2020,December). AP-NORC Poll: Only Half in US Want Shots as Vaccine Nears. Available online: <https://apnews.com/article/ap-norc-poll-ushalf-want-vaccine-shots-4d98dbfc0a64d60d52ac84c3065dac55>

Neumann-Böhme, S., Varghese, N. E., Sabat, I., Barros, P. P., Brouwer, W., van Exel, J., ... & Stargardt, T. (2020). Once we have it, will we use it? A European survey on willingness to be vaccinated against COVID-19.

Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering (Vol. 1, No. 1, pp. 61-67).

Niu, Q., Liu, J., Nagai-Tanima, M., Aoyama, T., Masaya, K., Shinohara, Y., & Matsumura, N. (2021). Public Opinion and Sentiment Before and at the Beginning of COVID-19 Vaccinations in Japan: Twitter Analysis. *medRxiv*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Quintana, I. O., Klein, C., Cheong, M., Sullivan, E., Reimann, R., & Alfano, M. The evolution of vaccine discourse on Twitter during the first six months of COVID-19.

Rajput, R., & Solanki, A. K. (2016). Review of sentimental analysis methods using lexicon based approach. *International Journal of Computer Science and Mobile Computing*, 5(2), 159-166.

Ritonga, M., Al Ihsan, M. A., Anjar, A., & Rambe, F. H. (2021, February). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. In IOP Conference Series: Materials Science and Engineering (Vol. 1088, No. 1, p. 012045). IOP Publishing.

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus pandemic (COVID-19). *Our world in data*.

Ruiz-Frau, A., Ospina-Alvarez, A., Villasante, S., Pita, P., Maya-Jariego, I., & de Juan, S. (2020). Using graph theory and social media data to assess cultural ecosystem services in coastal areas: Method development and application. *Ecosystem Services*, 45, 101176.

Scheffe, H. (1999). *The analysis of variance* (Vol. 72). John Wiley & Sons.

Shamrat, F. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 23(1), 463-470.

Shiller, R. J. (1995). Conversation, information, and herd behavior. *The American economic review*, 85(2), 181-185.

- Shim, J. G., Ryu, K. H., Lee, S. H., Cho, E. A., Lee, Y. J., & Ahn, J. H. (2021). Text Mining Approaches to Analyze Public Sentiment Changes Regarding COVID-19 Vaccines on Social Media in Korea. *International Journal of Environmental Research and Public Health*, 18(12), 6549.
- Singh, P. K., & Husain, M. S. (2014). Methodological study of opinion mining and sentiment analysis techniques. *International Journal on Soft Computing*, 5(1), 11.
- Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841-842).
- Sv, P., Ittamalla, R., & Deepak, G. (2021). Analyzing the attitude of Indian citizens towards COVID-19 vaccine—A text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.
- Szilagyi, P. G., Thomas, K., Shah, M. D., Vizueta, N., Cui, Y., Vangala, S., & Kapteyn, A. (2021). National trends in the US public's likelihood of getting a COVID-19 vaccine—April 1 to December 8, 2020. *JAMA*, 325(4), 396-398.
- Tan, A. H. (1999). *Text Mining: promises and challenges*. South East Asia Regional Computer Confederation, Singapore.
- Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.
- Van Essen, G. A., Palache, A. M., Forleo, E., & Fedson, D. S. (2003). Influenza vaccination in 2000: recommendations and vaccine use in 50 developed and rapidly developing countries. *Vaccine*, 21(16), 1780-1785.
- Ward, J. K., Alleaume, C., Peretti-Watel, P., Seror, V., Cortaredona, S., Launay, O., ... & Ward, J. (2020). The French public's attitudes to a future COVID-19 vaccine: The politicization of a public health issue. *Social science & medicine*, 265, 113414.
- Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., & Yao, H. (2019). Research on topic detection and tracking for online news texts. *IEEE Access*, 7, 58407-58418.
- Yoda, T., & Katsuyama, H. (2021). Willingness to receive COVID-19 vaccination in Japan. *Vaccines*, 9(1), 48.

Appendix

Appendix 1 :Covid- 19 government policy.

Name	Description
CountryName	Country name
CountryCode	Country code
RegionName	Region name
RegionCode	Region code
Jurisdiction	National- total or state total
Date	2020 Jan 01 till 2021 Feb
C - containment and closure policies	C1-C8
E - economic policies	E1-E4
H - health system policies	H1-H7
M - miscellaneous policies	M1

<https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md#containment-and-closure-policies>

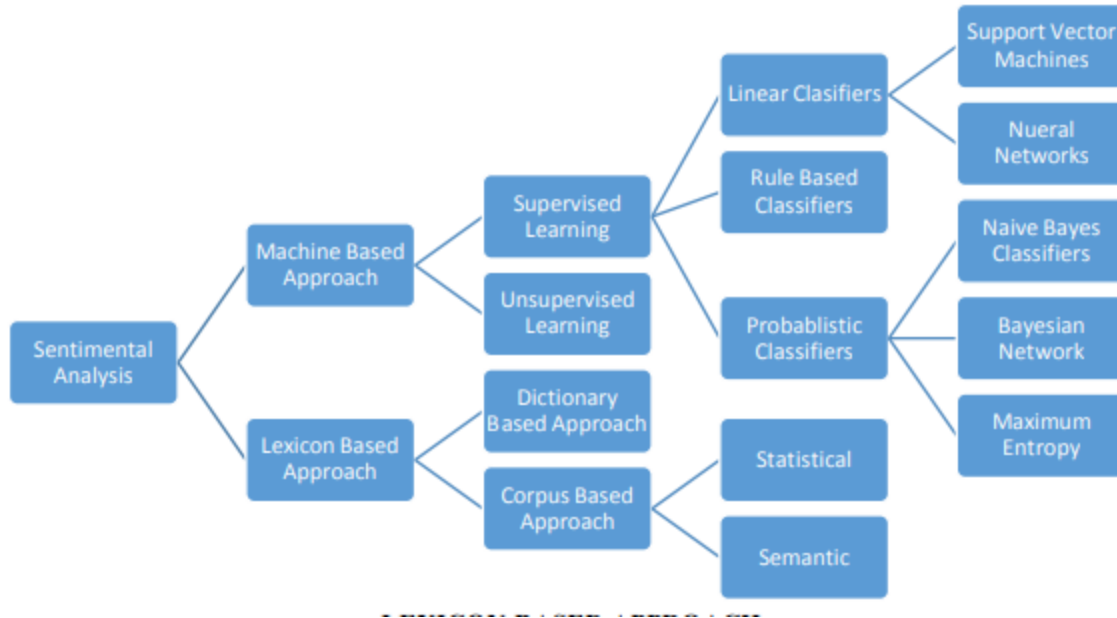
Appendix 2: Covid- 19 general statistics dataset.

No	Variables	No	Variables
1	iso_code	31	new_tests_smoothed_per_thousand
2	continent	32	positive_rate
3	location	33	tests_per_case
4	date	34	tests_units
5	total_cases	35	total_vaccinations
6	new_cases	36	people_vaccinated
7	new_cases_smoothed	37	people_fully_vaccinated
8	total_deaths	38	new_vaccinations
9	new_deaths	39	new_vaccinations_smoothed
10	new_deaths_smoothed	40	total_vaccinations_per_hundred
11	total_cases_per_million	41	people_vaccinated_per_hundred
12	new_cases_per_million	42	people_fully_vaccinated_per_hundred
13	new_cases_smoothed_per_million	43	new_vaccinations_smoothed_per_million
14	total_deaths_per_million	44	stringency_index
15	new_deaths_per_million	45	population
16	new_deaths_smoothed_per_million	46	population_density
17	reproduction_rate	47	median_age
18	icu_patients	48	aged_65_older
19	icu_patients_per_million	49	aged_70_older
20	hosp_patients	50	gdp_per_capita
21	hosp_patients_per_million	51	extreme_poverty
22	weekly_icu_admissions	52	cardiovasc_death_rate
23	weekly_icu_admissions_per_million	53	diabetes_prevalence
24	weekly_hosp_admissions	54	female_smokers
25	weekly_hosp_admissions_per_million	55	male_smokers
26	new_tests	56	handwashing_facilities
27	total_tests	57	hospital_beds_per_thousand
28	total_tests_per_thousand	58	life_expectancy
29	new_tests_per_thousand	59	human_development_index
30	new_tests_smoothed		

Appendix 2:

III. SENTIMENT CLASSIFICATION TECHNIQUES

The SC techniques, shown in Fig. 2:



<https://ijcsmc.com/docs/papers/February2016/V5I2201636.pdf>

Appendix 3:

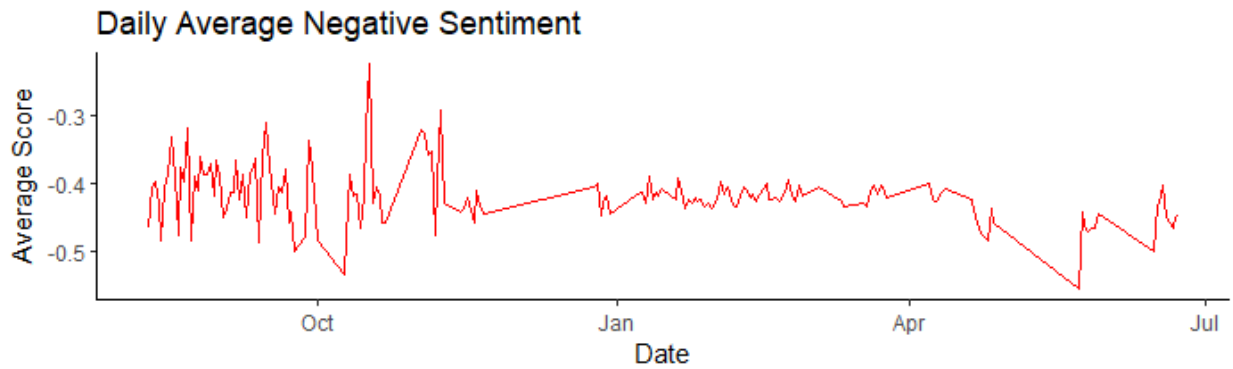
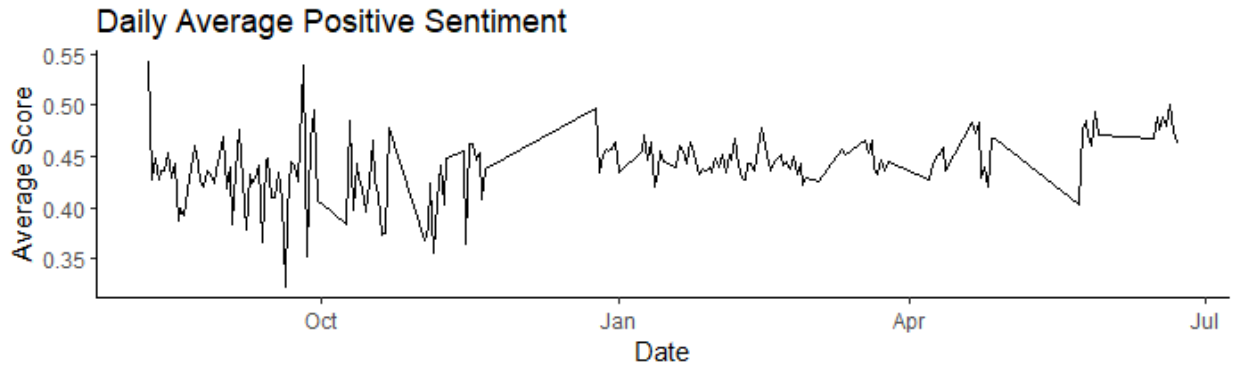


Figure 1 : The daily average positive sentiment score and negative sentiment score based on COVID vaccine

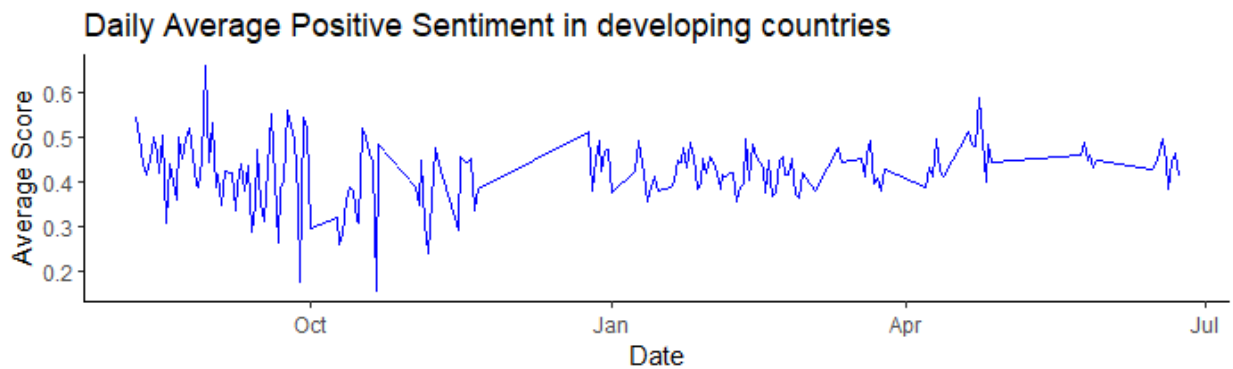
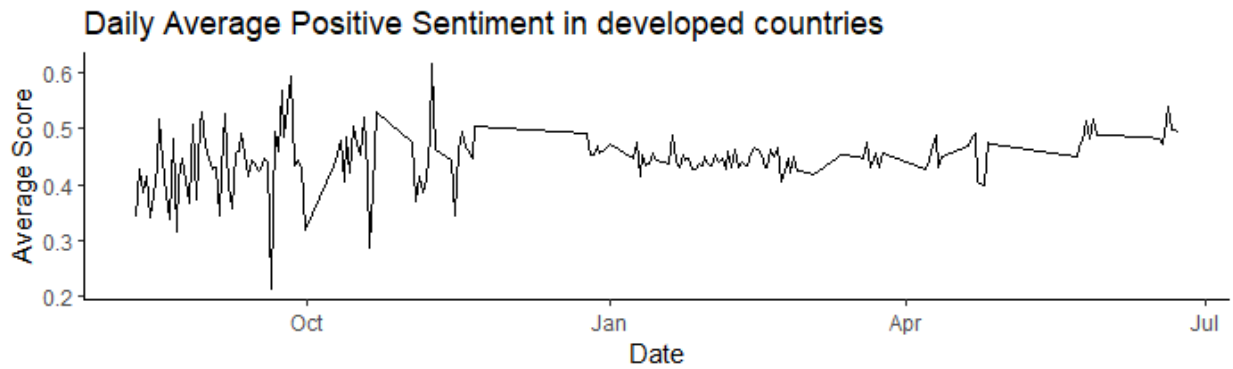


Figure 2: The daily average positive sentiment score of developed and developing country on Pfizer COVID vaccine

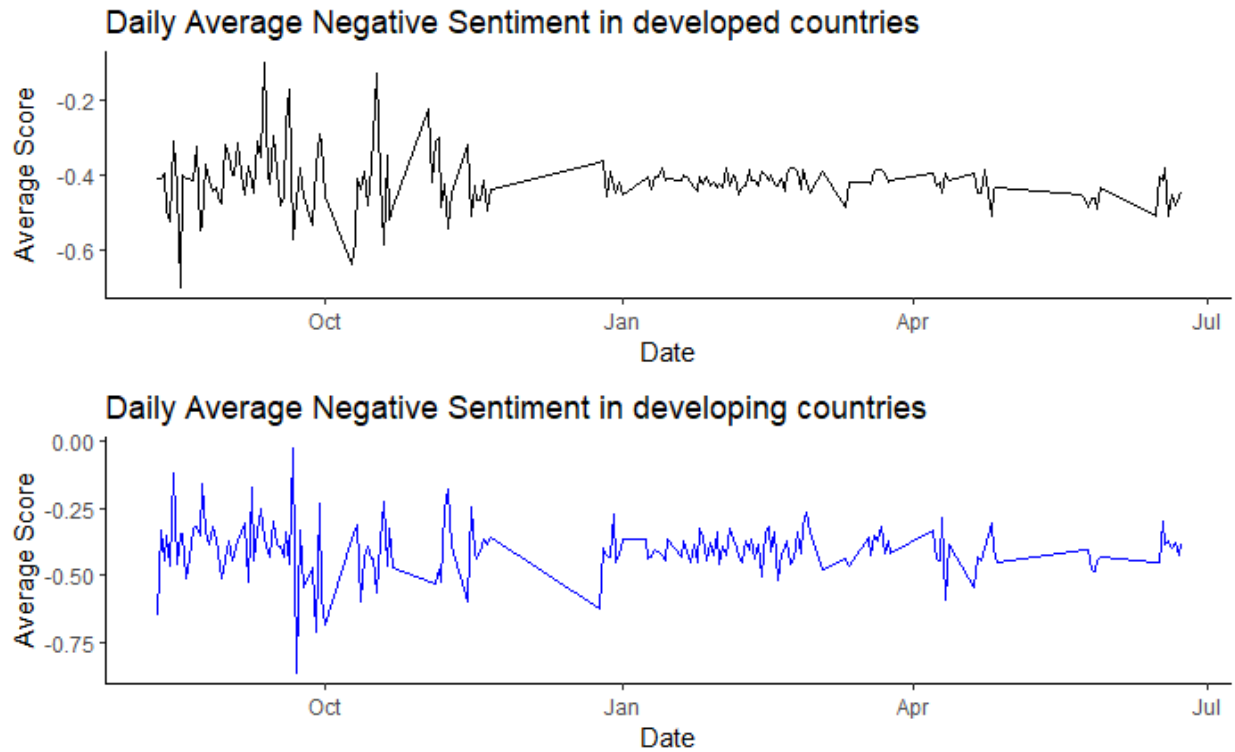


Figure 3: The daily average negative sentiment score of developed and developing country on Pfizer COVID vaccine