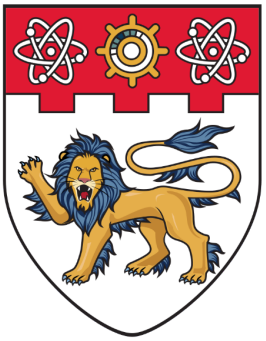


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CZ4034 Information Retrieval

Group **23** Submission

Name	Matric number
ERIC CHUA JEE HON	U2020527J
GOH HONG JUN DAMIEN	U1822722E
KELVIN WONG WAI LEONG	U1922962J
ONG ENG HAO	U1921623C
TRAN HIEN VAN	U1920891J
TRUONG QUANG DUC	U1840050G

1. Introduction **4**
1.1. Background 4
1.2. Objective 4
1.3. Method 4

2. Crawling **6**
2.1. Methodology of crawling data 6
Question 1:
a. How you crawled the corpus (e.g., source, keywords, API, library) and stored it 6
2.2. Analysis of crawled data 9
Question 1:
b. What kind of information users might like to retrieve from your crawled corpus 9
Question 1:
c. The numbers of records, words, and types (i.e., unique words) in the corpus 9

3. Indexing and querying **9**
3.1. Indexing 9
3.2. Querying 11
3.3. UI for querying of crawled data and indexing new data 13
Question 2:
a. Build a simple web interface for the search engine 13
Question 2:
b. A simple UI for crawling and incremental indexing of new data 17
Question 2:
c. Write five queries, get their results, and measure the speed of the querying 17
3.4. Innovations for enhancing indexing and ranking 19
Question 3 19
3.4.1. Tweet Sorting 19
3.4.2. Vaccine types and Sentiment Filter 20
3.4.3. Visualisation for statistics from tweets 20
3.4.4. Spell Checking 20

4. Classification: **21**
4.1. Classification Approaches 21
Question 4 & 5: 21
a. Motivate the choice of your classification approach in relation with the state of the art 21
1) BERT-CNN 21
2) LSTM 21
3) Roberta 22
4) Ensemble 22
b. Discuss whether you had to preprocess data (e.g., microtext normalization) and why 22
c. Build an evaluation dataset by manually labelling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80% 26
d. Provide evaluation metrics such as precision, recall, and F-measure and discuss results Discuss performance metrics, e.g., records classified per second, and scalability of the system 27
BERT-CNN 28

LSTM 29
Roberta 30
Ensemble 31
4.2. Summary 32
5. Conclusion **33**
6. Contribution **34**
7. Submission Link **35**
Appendix **36**

1. Introduction

1.1. Background

On 31 December 2019, WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China. Many cases later, the official name of COVID-19 was issued by WHO on 11 February 2020. COVID-19 has affected millions of lives and despite widespread immunization against COVID-19, there is still a sizable group of people who reject vaccination (anti-vaxxers) or people who are against certain types of vaccines such as Covaxin, Moderna, Sputnik or Sinovac for various reasons. It's critical for vaccine firms and government agencies to understand what people are worried about when it comes to being vaccinated against current and future pandemics. Individuals may also wish to know what the public thinks about a specific vaccine or what is currently trending about a specific vaccine topic. However, reading through many netizens' thoughts to gain a sense of the public's opinion on immunisation or to find out what is trending is not practical. As a result, an information retrieval system is required that not only displays search results but also predicts netizen sentiment on a topic and displays overall public sentiment.

1.2. Objective

The purpose of this assignment is to create and build an information system that does sentimental analysis of netizens' opinions of COVID-19 vaccinations. This information retrieval system can help shed light for future vaccine companies to discover what customers want, such as transparency in data for future pandemic vaccines. We will focus on vaccination firms such as Covaxin, Moderna, Sputnik and Sinovac in this assignment.

1.3. Method

Figure 1. Twitter search results returning latest results in both German and English

To gain insights of the data, the team has chosen Twitter which is one of the most popular social media platforms. Twitter has always been a platform for netizens to express their opinions, feelings and emotions. However, the search results that we use on Twitter can show results of other languages other than English (as shown in Figure 1).

To overcome this problem, the team has standardised the tweets that are in the information retrieval system to only English tweets.

To build the information retrieval system, we crawled the data from Twitter about general COVID-19 vaccination and vaccines from companies such as Pfizer, Moderna, Sinovac, Sputnik and Covaxin. For indexing and querying, we then used Apache Solr, a platform to index the crawled data to provide search results and spell correction. Displaying of the search results and spell correction is powered by REACT - a JavaScript library for building user interfaces. The system also displays a visual representation of the most common words for the search result with WordCloud.

To determine the public sentiment of a tweet, the team first combined the manually labelled sentiments from both Kaggle and a portion of the crawled tweets to train a few models using LSTM, Bert+CNN, RoBERTa and Ensemble. Then, we used the trained model to label the sentiment of both the latest 7 days tweets (limitations of TwitterAPI) and the crawled tweets. The system then uses the labelled tweets to build a sentiment distribution graph for better illustrating the general sentiments.

2. Crawling

2.1. Methodology of crawling data

Question 1:

a. How you crawled the corpus (e.g., source, keywords, API, library) and stored it

All the data is crawled from Twitter using Twitter API (v2). We managed to crawl a total of ***10288*** tweets, consisting of either covid-19 vaccination tweets or vaccine companies. To ensure our crawled tweets do not contain other languages other than English, we included “lang: en” at the end of every query. Besides that, we had also filtered out retweets, as they cause duplicates in the data.

```
query = 'covid19 vaccine OR covid vaccine OR Coronavirus Vaccine OR coronovirus vaccine lang:en'
covid_vaccine_data = crawl_data(query, limit=15000)
covid_vaccine_data
```

Figure 2. Crawling general covid-19 vaccine

We first crawled the general covid-19 vaccine tweets with the query of “covid19 vaccine OR covid vaccine OR Coronavirus Vaccine OR cornovarius vaccine lang:en" with keywords being “vaccine”. Different forms of the word such as “covid19 vaccine” or “Coronavirus vaccine” are also being queried.

```
query = 'Pfizer/BioNTech vaccine OR Pfizer-BioNTech vaccine OR Pfizer vaccine OR BioNTech vaccine OR pfizer lang:en'
new_pfizer_data = crawl_data(query, 20000)
new_pfizer_data
```

Figure 3. Crawling Pfizer data

Then, we continued with crawling of vaccine companies tweets such as Pfizer with the query: “Pfizer/BioNTech vaccine OR Pfizer-BioNTech vaccine OR Pfizer vaccine OR BioNTech vaccine OR Pfizer lang:en” with keyword being “pfizer”

To make sure that we are able to crawl more data, we use different synonyms for a word. For example, both “CoronaVac” and “Sinovac” were queried because they have the same meaning but of different spelling. Additionally, since twitter API’s query is not case-sensitive, we omitted all the letter case permutation queries. For example, the API returned a tweet by “YourFlawedView” that contains “PFIZER” but not “Pfizer” after querying “Pfizer”:

“I am beside myself PEOPLE IF YOU FOLLOW THE SCIENCE ON THE PAPERS PFIZER JUST RELEASED YOU KNOW THEY ARE TRYING TO KILL YOU I guess if you get the 4th you deserve the outcome”

	tweet_id	user_id	username	text	like_count	retweet_count	location	creation_date_time
0	1503289026140569601	814398812030529537	IndianJPharmSci	Valneva now expects recommendation on COVID-19...	0	0	Mumbai, India	2022-03-14 08:36:48+00:00
1	1503288967160123392	1003681978972024833	Phleming1	@p_duval @olivierveran second booster 😭😭 \nhttp...	0	0	Rhône-Alpes, France	2022-03-14 08:36:34+00:00

Figure 4. Samples of crawl tweets

We have used the same method for all the vaccine companies (Covaxin, Moderna, Sputnik, Sinovac) and finally, combined both vaccine tweets and vaccine companies tweets.

	tweet_id	user_id	username	text	like_count	retweet_count	location	creation_date_time
2	1502897767315767298	1396115471703805953	CovidvaxDEL	North delhi has at least 2721 new slots availa...	0	0	New Delhi, India	2022-03-13 06:42:04+00:00
32	1502844917907070978	1396115471703805953	CovidvaxDEL	NE delhi has at least 13 new slots available b...	0	0	New Delhi, India	2022-03-13 03:12:04+00:00
34	1502843408498724871	1396115471703805953	CovidvaxDEL	NE delhi has at least 11 new slots available b...	0	1	New Delhi, India	2022-03-13 03:06:04+00:00
35	1502843406716125184	1396115471703805953	CovidvaxDEL	North delhi has at least 1982 new slots availa...	0	1	New Delhi, India	2022-03-13 03:06:04+00:00
44	1502820762880966657	1396115471703805953	CovidvaxDEL	North delhi has at least 1686 new slots availa...	0	0	New Delhi, India	2022-03-13 01:36:05+00:00
...	...	...	...	...	...	...	...	...
2120	1500661533344342016	1396115471703805953	CovidvaxDEL	Central Delhi has at least 9058 new slots avai...	0	0	New Delhi, India	2022-03-07 02:36:05+00:00
2196	1500555850167242752	1396115471703805953	CovidvaxDEL	NW Delhi has at least 16404 new slots availabl...	0	1	New Delhi, India	2022-03-06 19:36:08+00:00
2197	1500554336057061377	1396115471703805953	CovidvaxDEL	NW Delhi has at least 12374 new slots availabl...	0	0	New Delhi, India	2022-03-06 19:30:07+00:00
2199	1500552833451495435	1396115471703805953	CovidvaxDEL	NW Delhi has at least 8424 new slots available...	0	0	New Delhi, India	2022-03-06 19:24:09+00:00
2299	1500424721472823296	1396115471703805953	CovidvaxDEL	East Delhi has at least 3000 new slots availab...	0	1	New Delhi, India	2022-03-06 10:55:04+00:00

80 rows × 8 columns

Figure 5. An example of a tweet made by bot/spammer

After combining the two sets of data, the team looked through the tweets to identify if there were any bots/spammers. If there is, we will manually remove the tweets from the data. This step is essential because repeating tweets will incur bad performance on our model when doing sentiment analysis. One such example of bots generated tweets is from the user “CovidvaxDEL” (shown in Figure 5).

	C	D	E	F	G	H
1	username	text	like_count	retweet_count	location	creation_date_time
2	chovue	sigh youre the one in the dark If you truly dont want to research it thats on you	0	0	No Location	13/3/2022
3	chovue	jnj is based on adenovirus tech Covaxin is based off polio's technology of a trad	0	0	No Location	13/3/2022
4	CaptainXe	I don't care about covaxin No do your own work Present your own sources I will	0	0	No Location	13/3/2022
5	chovue	lol see my point exactly no buddy I am talking about covaxin vaccine SIGH go dc	0	0	No Location	13/3/2022
6	Nii_Anue_	Iâ€™m with you all the way Covaxin is the foot in the door OCGN baby steps	0	0	No Location	13/3/2022
7	Harlempe	3 So let's not conflate use of a whole virus w mismatched strains Yes use of a wh	0	0	No Location	13/3/2022
8	FireTruck1	Is Covaxin approved by EU UK AUS CAN etc	0	0	No Location	13/3/2022
9	FullTimeD	The studies coming out on vaccines and grey brain matter are scary Why would	1	0	No Location	13/3/2022
10	chovue	Like I SAID follow the money MRNA shots are garbage and they just obtained bi	0	0	No Location	13/3/2022
11	TimesMai	+91 88266 04947 AIIMS Booster dose no for trial of Nasal Spray Send WhatsApp F	2	0	No Location	13/3/2022
12	Kiresa1	Safe and effective is Covaxin Explain why we donâ€™t have this REAL vaccine ir	1	0	Denver	13/3/2022
13	pmcdunn	Medicago seems to be something in between NovaVax and Covaxin Spike is in	0	0	Toronto, Ontario	13/3/2022
14	Turrialba2	He even knows COVAXIN is the best	0	0	No Location	13/3/2022
15	NewIndia	covaxin Mettupalayam Gh Issues Tokens To Check Covid Vaccine Wastage Coim	0	0	Ghaziabad, India	13/3/2022

Figure 6. A screenshot of the combined dataset in csv file

The combined tweets are then stored in a csv file (shown in Figure 6) and each tweet (record) corresponds to a line.

2.2. Analysis of crawled data

Question 1:

b. What kind of information users might like to retrieve from your crawled corpus

There are several information users can use to retrieve from the crawled corpus. For example, to understand the public's opinion on a certain vaccine company, users can search for example, "Pfizer" in the search bar. Users can also filter the text, based on popular tweets using like_count/retweet count to retrieve the top comments. Besides that, users can also filter to see the different types of comments made on different vaccine types.

Question 1:

c. The numbers of records, words, and types (i.e., unique words) in the corpus

Our combined data has a total of ***10288*** tweets crawled using Twitter API (v2). All the crawled tweets are in English language. The total number of words in our datasets is ***264777***. Among those, we have ***17055*** unique words. Beside that, our dataset has 8 columns namely, ***tweet_id, user_id, username, text, like_count, retweet_count, location and creation_date_time.***

3. Indexing and querying

3.1. Indexing

For the project, Apache Solr was used as the platform for querying the data. Solr is written in Java and runs as a standalone full-text search server within a servlet container such as Jetty. It uses Lucene Java search library for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it usable from many popular programming languages. Solr's external configuration allows it to be tailored to many types of applications without Java coding, and it has a plugin architecture to support more advanced customization.

After retrieving and preprocessing data from Twitter, we post them into Solr for indexing. The field names in Solr are as follows:

```
<field name="created_at" type="pdates"/>
<field name="creation_date_time" type="pdates"/>
<field name="id" type="string" multiValued="false" indexed="true" required="true" stored="true"/>
<field name="like_count" type="plongs"/>
<field name="location" type="text_general"/>
<field name="retweet_count" type="plongs"/>
<field name="sentiment" type="text_general"/>
<field name="source" type="text_general"/>
<field name="spellcheck" type="text_general" multiValued="true" indexed="true" stored="false"/>
<field name="text" type="text_gen_stem"/>
<field name="tweet_id" type="plongs"/>
<field name="user_desc" type="text_general"/>
<field name="user_id" type="plongs"/>
<field name="username" type="text_general" multiValued="false"/>
<field name="verified" type="booleans"/>
```

Figure 7. Field names in Solr

In order to index the crawled data, there are configurations needed to be made in the "managed-schema" file prior to the posting of the data to the core. Because the focus is mainly on the tweets itself (and not other parameters like Location due to increase in awareness in privacy), we set the tweets to have field type of "text_gen_stem" and the following filters:

1. *StopFilterFactory*
2. *LowerCaseFilterFactory*
3. *SnowballPortalFilterFactory*
4. *NGramFilterFactory*

For every document, Solr will index it by removing all the stopwords, making all words lowercase, stemming by Porter rules, and applying Ngram for phrased query.

```
"tweet_id":[1502900173445296136],
"user_id":[489184831],
"username":"chovue",
"like_count":[0],
"retweet_count":[0],
"location":["No Location"],
"creation_date_time":["2022-03-13T00:00:00Z"],
"text":"sigh one dark truly want research I baby sitter I already tell need follow fukin money research covaxin sheesh",
"sentiment":["NEUTRAL"],
"id":"06b28adc-d73c-47f0-9cfa-45a74da1ae5f",
"_version_":1728832429988249600},
```


Figure 8. Sample tweet data

Once the data has been successfully crawled using the Twitter API, details, such as fields and types, of the tweets were selected and added to Solr altogether. Figure 7 shows the details of every tweet that were recorded. Specific fields of the tweets would later then be used for querying data based on the user's input.

3.2. Querying

Querying can also be done in Solr once the data has been successfully indexed and added. As Solr supports communication via its REST API, later, our system could get search results for queries via GET request and update data to Solr for indexing via POST request.


```
http://localhost:8983/solr/CZ4034/select?indent=true&q.op=OR&q=text%3Amoderna

{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "q":"text:moderna",
      "indent":"true",
      "q.op":"OR",
      "_":"1649001446774"}},
  "response":{"numFound":2190,"start":0,"numFoundExact":true,"docs":[
      {
        "tweet_id":[1502730298819293187],
        "user_id":[16631823],
        "username":"MaryKRe",
        "like_count":[0],
        "retweet_count":[0],
        "location":["Ann Arbor, MI"],
        "creation_date_time":["2022-03-12T00:00:00Z"],
        "text":"I need update asap pfizer booster pfizer shot dose increase moderna booster moderna shot dose increase pfizer booster modernas increase
        "sentiment":["NEUTRAL"],
        "id":"9f1c027c-211b-4d15-a279-f7d6db72b09f",
        "_version_":1728832431039971385},
```


Figure 9. Sample querying in Solr

Figure 9 shows an example of how the result of a tweet will be returned when the word "**moderna**" has been queried. More examples will be shown in Question 2c and appendix.

For special features such as spell checks, “solrconfig.xml” and “managed-schema” files have to be edited. The schema has 2 different fields to store the tweet content: "text" and "spellcheck". The "text" field is used for constructing index for tweet content retrieval while "spellcheck" is used for constructing index for spell check so that the system will suggest based on misspelt word input by users. The “spellcheck” field’s type is the same as “text_gen_stem” except that we removed the stemming and NGram filter to avoid the suggestions being in stemmed and NGrams form. Besides, in “solrconfig.xml”, we have added “/spell” requestHander that is able to return a spell error in the query. This handler uses *solr.IndexBasedSpellChecker* to make suggestions for a misspelt query. More examples are added in the Appendix section.

Figure 10. Example of spell check for word “pfizrr”

http://localhost:8983/solr/CZ4034/spell?indent=true&q.op=OR&q=hello&spellcheck=true


```
{
  "responseHeader":{
    "status":0,
    "QTime":6},
  "response":{"numFound":0,"start":0,"numFoundExact":true,"docs":[]
  },
  "spellcheck":{
    "suggestions":[
      "hello",{
        "numFound":1,
        "startOffset":0,
        "endOffset":5,
        "origFreq":9,
        "suggestion":[{
            "word":"hell",
            "freq":44}]}],
    "correctlySpelled":false,
    "collations":[]}}
```


Figure 11. Example of spell check for word “hello”

It is noted that the suggestion will be based on the frequency of similar words in the corpus. For example in Figure 11, if a user queries “hello”, although our search engine still returns related tweets, the system will also suggest another word (“hell”) which is more frequent than the initial query in the database.

3.3. UI for querying of crawled data and indexing new data

Question 2:

a. Build a simple web interface for the search engine

The UI that was used to design the web interface for the search engine created was mainly done in ReactJS. ReactJS was used as the platform allowed for both frontend and backend development to be done together. Along with ReactJS, Flask was also used to display graphs to show visualisations of the tweet details.

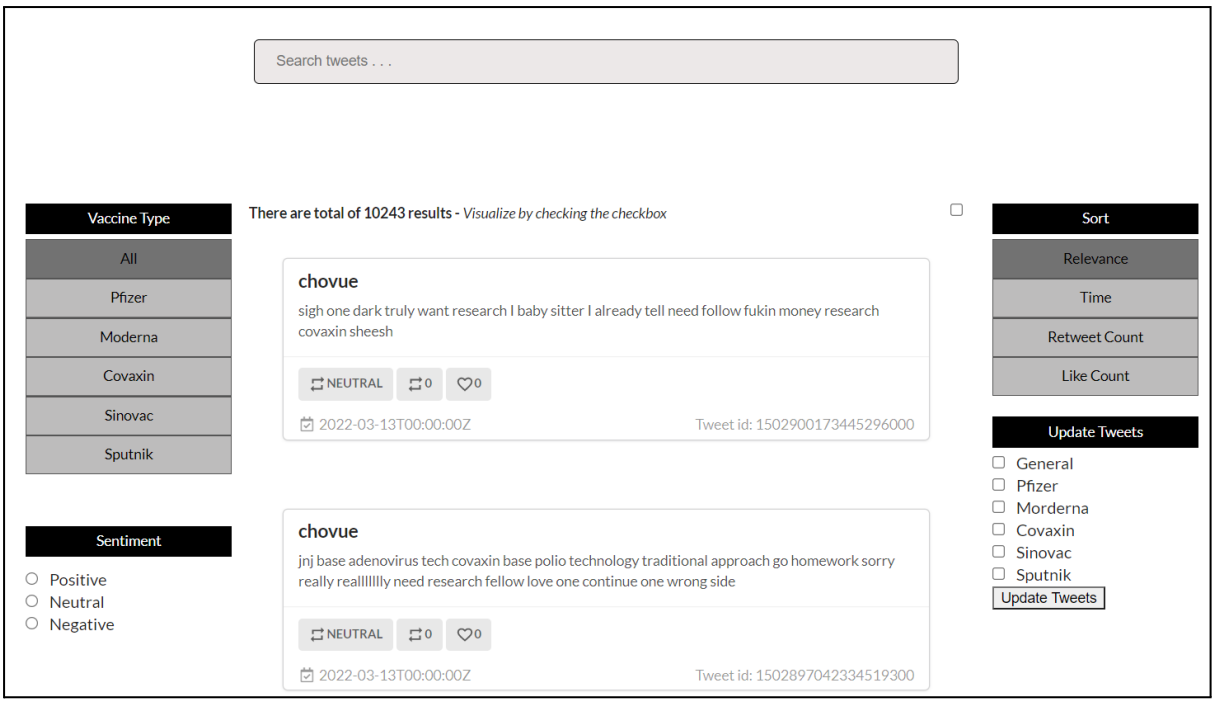


Figure 12. UI of main page

Figure 12 shows the main UI for the web search engine that was designed. The main page consists of a **search bar**, **vaccine category filter**, **sentiment selection section**, **visualisation checkbox**, **display section**, **sorting section** and a **tweet update section**.

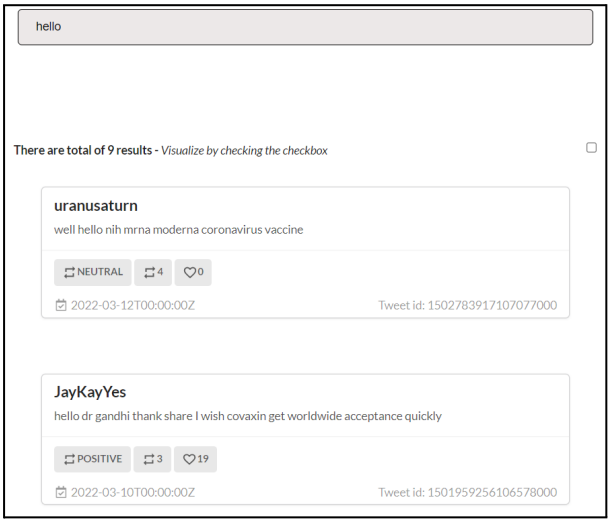


Figure 13a. Search Bar function

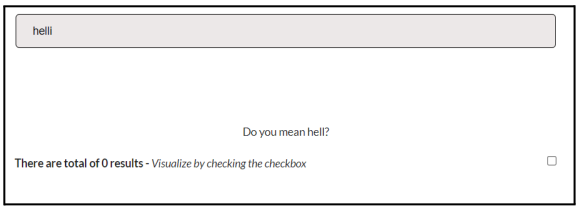


Figure 13b. Search Bar function

The **search bar** allows for users to search for tweets depending on what the user types in the search bar as shown in Figure 13a. The search function also includes a spell check which will clarify what the user is trying to search for if the users spell a word wrongly as seen in Figure 13b. The **display**

section will then display the tweets that users have searched for and the tweets will also include some information about the tweets as well.

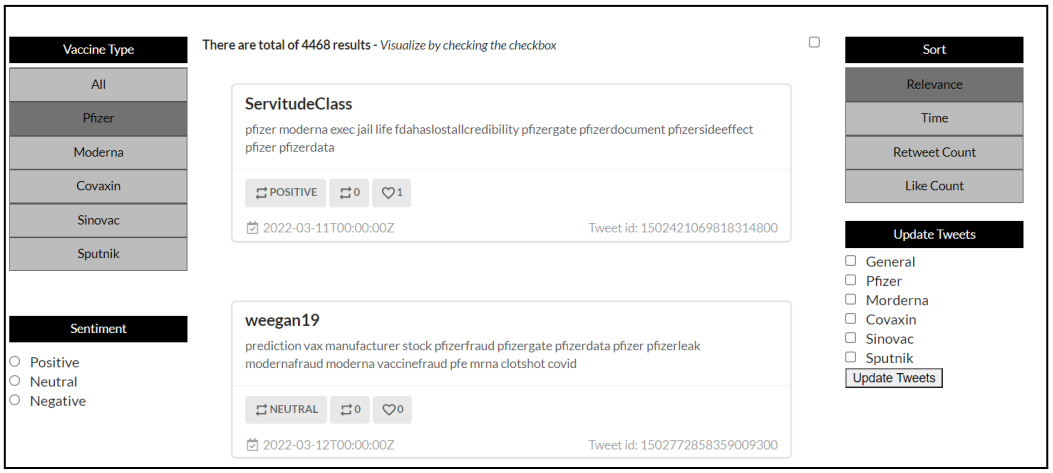


Figure 14. Vaccine category function

A **category filter** on the left was included to filter tweet results based on the different vaccine types chosen by the user. Figure 14 shows an example of when the “Pfizer” vaccine type was selected which will result in all the tweets containing that vaccine type to be returned.

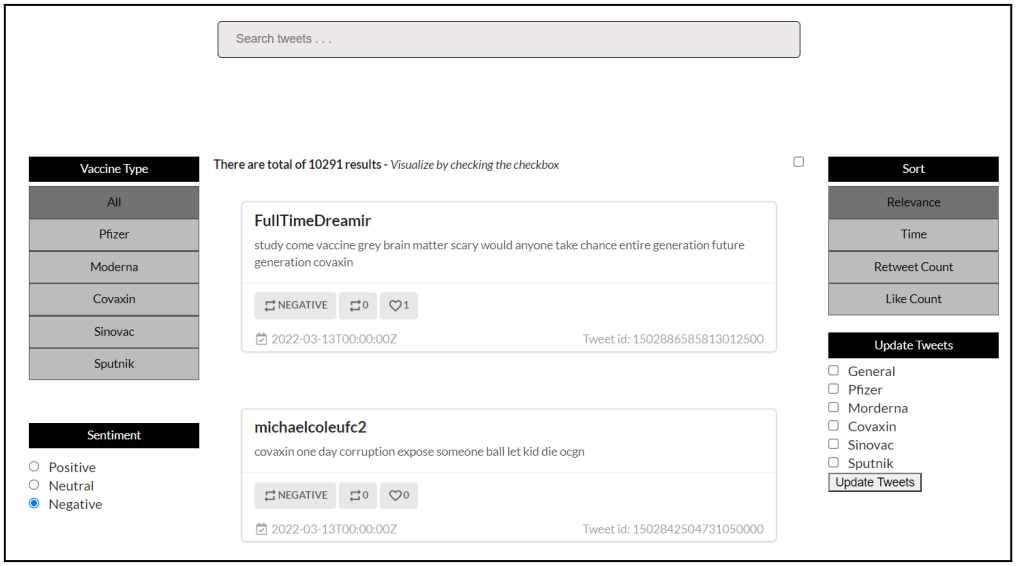


Figure 15. Sentiment selection function

A **sentiment selection section** was added for users to choose the type of sentiment of tweets to be shown. Figure 15 shows the Negative sentiment being selected and hence the tweets that will be returned will have a negative sentiment.

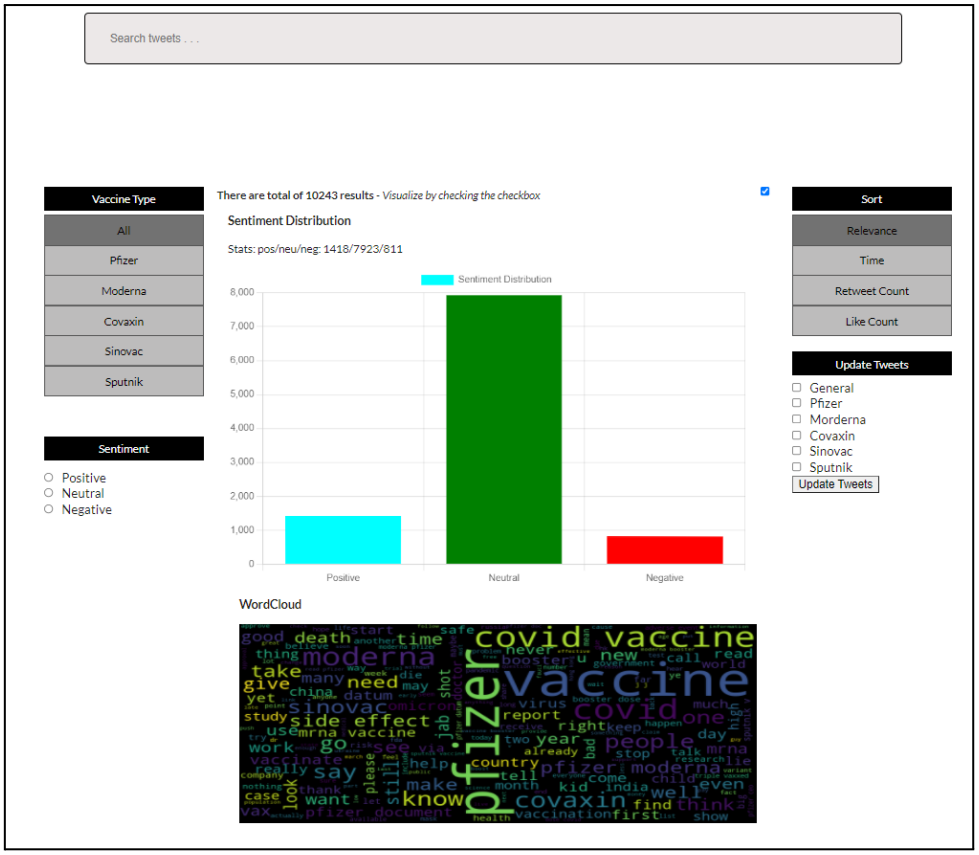


Figure 16. Visualisation checkbox function

The **visualisation checkbox** when ticked, will display the sentiment distribution as well as the word cloud of the tweet results that was queried to visualise the results. When the checkbox is unticked, the main page of the search engine will be shown back again.

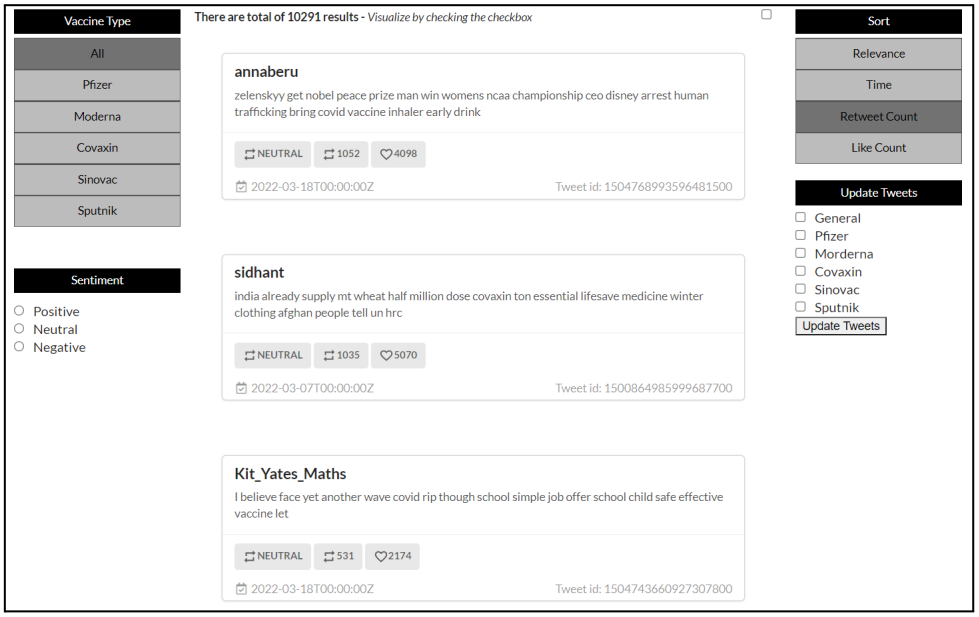


Figure 17. Sorting function

The **sorting section** allows users to sort tweets based on the options the user selects. Figure 17 shows an example where the user decided to sort the tweet results based on the number of retweets the tweets has.

Question 2:

b. A simple UI for crawling and incremental indexing of new data

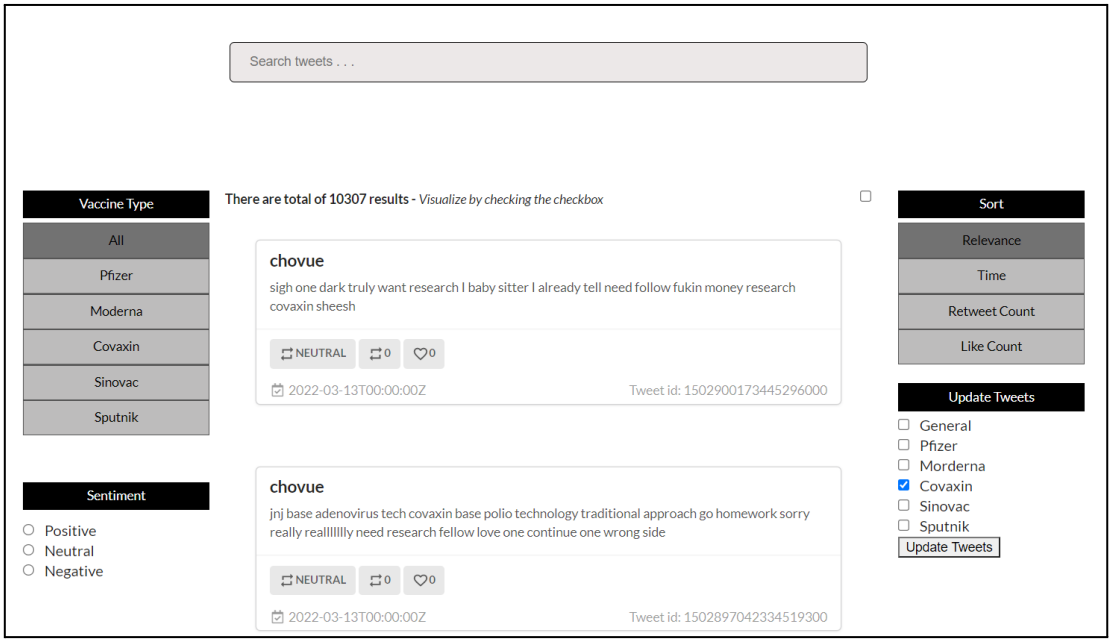


Figure 18. Update tweets function

An additional **tweet update section** was implemented to allow users to update the tweet database based selection of the vaccine type. By checking the checkbox, user can crawl latest tweets (in past 7 days as limited by Twitter API access) from checked vaccine types. The update function will fetch the latest tweets related to such vaccines, classify tweets into 3 different sentiments and then index them into Solr. After finishing indexing, the website will be reloaded and return tweet results with updated tweets. Figure 18 shows the number of results increasing after “Covaxin” related tweets was selected to be added to the database.

Question 2:

c. Write five queries, get their results, and measure the speed of the querying

The table below shows five queries, their results, and the speed of the querying.

Query	Time taken/ ms	#Results	Top Tweet	
“anti”	1	216	“pfizer corrupt criminal two faced selfish opportunist bribe salary instruct destructive antidemocratic fascist antimlk jr antiimmigration antius antiworld anti anti hidden terrorist couple usa world putin”	"responseHeader":{ "status":0, "QTime":1, "params":{ "q":"text: anti", "indent":"true", "q.op":"OR", "_":"1649338548349"}}, "response":{"numFound":216,"start":0,"numFoundExact":true,"docs":[
“vaccine”	1	4637	“pfizer covid vaccine side effect reveal official document covidvaccine covaxin coronavirusupdate corona covid vaccinesideeffect vaccinepassports vaccine vaccination vaccineswork vaccinemandate pfizerdocument pfizer fda”	"responseHeader":{ "status":0, "QTime":1, "params":{ "q":"text: vaccine", "indent":"true", "q.op":"OR", "_":"1649338548349"}}, "response":{"numFound":4637,"start":0,"numFoundExact":true,"docs":[
“booster”	1	750	"pfizer booster j j booster coca cola booster aromat booster I I come andizi"	"responseHeader":{ "status":0, "QTime":1, "params":{ "q":"text: booster", "indent":"true", "q.op":"OR", "_":"1649338548349"}}, "response":{"numFound":750,"start":0,"numFoundExact":true,"docs":[
“covid”	1	3328	"please take vaccine rather get covid vaccine card without vaccine offer country cdc covid card usa nhs covid pass uk canada covid pass australia digital covid certificate eu digital covid certificate dm"	"responseHeader":{ "status":0, "QTime":1, "params":{ "q":"text: covid", "indent":"true", "q.op":"OR", "_":"1649338548349"}}, "response":{"numFound":3328,"start":0,"numFoundExact":true,"docs":[
“side effect”	2	493	"I think list side effect occur I think list side effect interest monitor yes possible side effect covaxin likely many"	"responseHeader":{ "status":0, "QTime":2, "params":{ "q":"text: side AND text: effect", "indent":"true", "q.op":"OR", "_":"1649338548349"}}, "response":{"numFound":493,"start":0,"numFoundExact":true,"docs":[

Table 1. Results and time taken for five queries

3.4. Innovations for enhancing indexing and ranking

Question 3

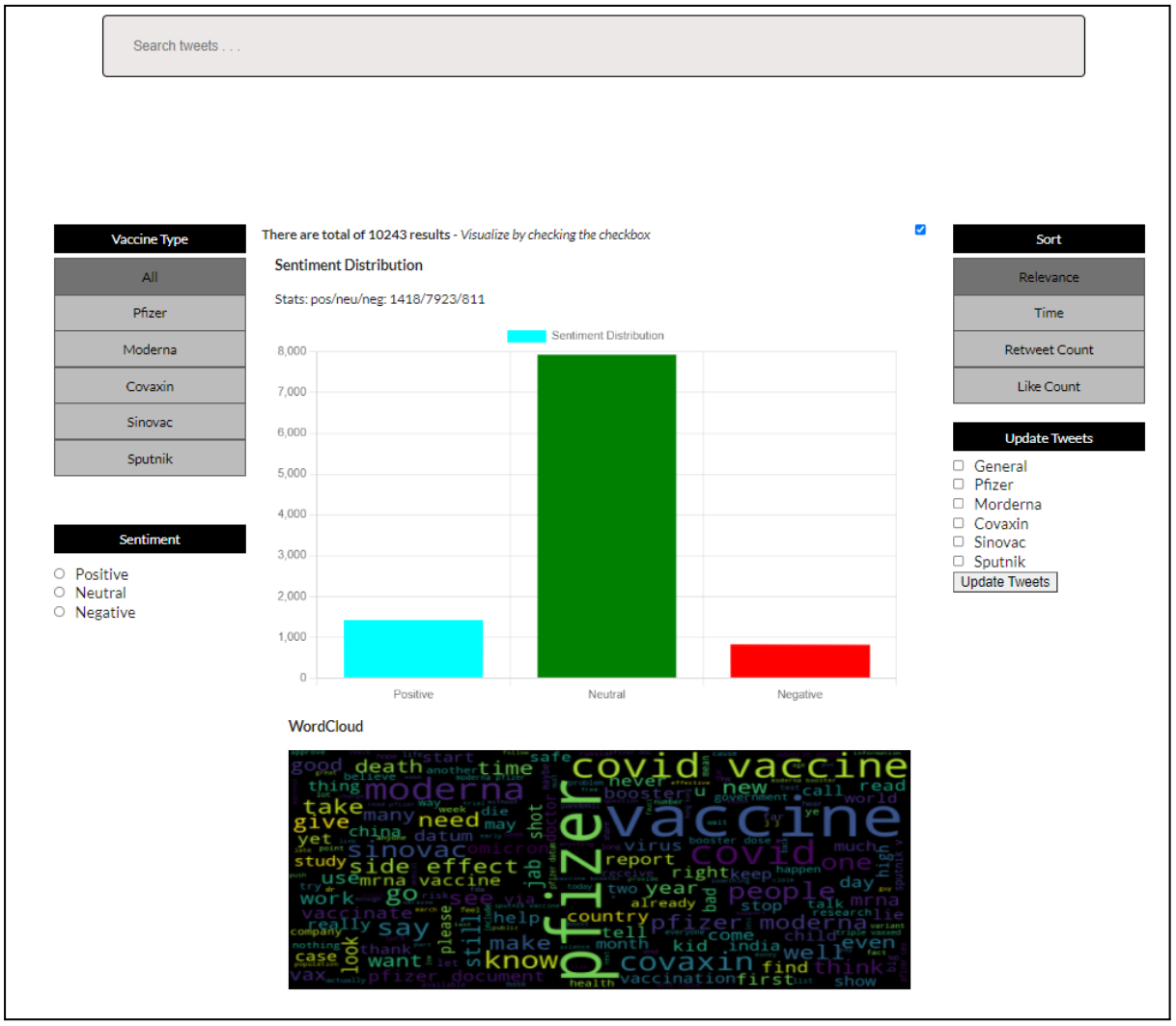


Figure 19. UI of visualisation page

Often, we will want to understand the sentiments of the netizens towards a certain topic and what are the current trends for the given topic for further analysis or inference. To achieve this, the team has built some innovations for enhancing indexing and ranking for the users to produce a better search result experience.

3.4.1. Tweet Sorting

There are 4 different types of sorting: Relevance, Time, Retweet Count, and Like Count as shown in Figure 19.

- Sort by Relevance: This is the default setting when a user first visits this website without adjusting any other parameters. The tweet results are returned by our search engine which will rank the documents based on their similarity with the query.
- Sort by Time: As Twitter is a social media, therefore, it is important to capture the latest reaction of people. Users could also utilise this sorting type to see the latest tweets after crawling new data.
- Sort by Like Count: It is necessary to mention that a high number of likes means a more popular post and may reflect the reaction of a large part of users.
- Sort by retweet count: For Twitter users, besides number of likes, there is a number for retweets, which is another way to measure popularity. A larger number of retweets could show that the post is more relevant and concerned by more people.

3.4.2. Vaccine types and Sentiment Filter

As users may be interested in some types of vaccine and would like to search for specific types, we implemented a Vaccine Categories filter. Here, we listed our main vaccine types that we have in our database. Users could search a query or even filter sentiment and sort by different criterias for a particular vaccine as shown in Figure 14.

For a search query, the user is able to view tweet results based on their sentiment by selecting a radio button for that specific sentiment: Positive, Negative or Neutral as shown in Figure 15. The sentiment of each tweet is classified by our model and stored in our database before the system is deployed except the newly crawled tweets which will be predicted in real time and indexed to the database.

3.4.3. Visualisation for statistics from tweets

After searching for the search query, the user is able to visualise the statistics by clicking the checkbox next to the “Sort” section. The UI will display graphical representation from query results: Sentiment Distribution and WordCloud as shown in Figure 20. Users can see a bar chart showing the number of Positive, Negative, and Neutral tweets for this particular search query. Furthermore, we also implemented “WordCloud” - a graphical representation of word frequency from all the tweets that give a greater prominence to words that appear more frequently.

3.4.4. Spell Checking

The application utilises the suggestion function provided by Solr. The system will still return results for misspelt words if available. However, based on similar but correct-formed words in the corpus, the system will ask the user by “Do you mean {correct-spelt word}” as shown in Figure 13b. One thing to

note is that since Solr provides suggestions based on index created and words are lowercase when building index, the UI will display lowercase words.

4. Classification:

4.1. Classification Approaches

Question 4 & 5:

a. Motivate the choice of your classification approach in relation with the state of the art

The team researched a few methods to classify sentiments for the tweets. The four models are BERT-CNN, Lstm, RoBERTa, Ensemble. The team calculated the accuracy for each of the methods and used the highest accuracy model to classify our tweets in the system. The highest accuracy we got for the tweets is from the RoBERTa model.

Below is an overview of the models that we have used:

1) BERT-CNN

Bi-Directional Encoder Representations from Transformers (BERT) is state-of-the-art language model which can be fine-tuned based on context, which in this case, for the sentimental analysis and Convolution Neural Network (CNN) is a deep learning algorithm that is initially used to classify images. The team uses a BERT base model where a tweet is passed through 12 layers of transformer block to get the vector representation of the context. Then it is passed to the CNN for classification of whether a tweet is negative, neutral or positive.

2) LSTM

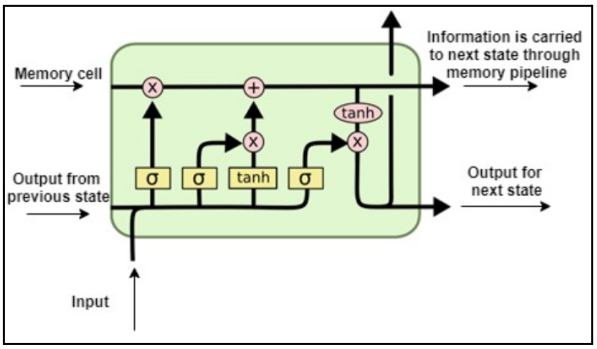


Figure 20. Architecture of Long Short Term Memory (LSTM)

LSTM is also known as long short term memory. It is a type of Recurrent Neural Network (RNN). It is very popular nowadays, especially in speech recognition, machine translation, sentiment analysis, and text classification. It can not only just process single data points but entire sequences of data. Many people have started using LSTM instead of RNN due to the model's excellent capability of memorising information.

From Figure 21, the memory cell serves as the purpose to help carry the information from the current point to another point. With this, it can remember data from the previous state to the next state, which is beneficial in helping us predict better.

3) Roberta

A robustly optimised technique for pretraining natural language processing (NLP) systems that improves on Google's self-supervised method, Bidirectional Encoder Representations from Transformers, or BERT. RoBERTa builds on BERT's language masking method, which teaches the system to predict purposely hidden content within otherwise unannotated language instances.

4) Ensemble

Ensemble model is based on the voting classifier: It takes the predictions from multiple other models and combines the results together to form the final prediction, which can produce a much more accurate result overall.

The models used for this voting ensemble are: Random Forest Classifier, C-Support Vector Classifier, and Logistic Regression. This is also paired with some hyper-parameters tuning to enhance the accuracy even further.

Random Forest is a model that is widely used for classification tasks. It builds decision trees on different samples and takes their majority vote for classification.

C-Support Vector Classifier is a supervised learning model used for classification as well. It was developed by Vladimir Vapnik, and it used a vectorized space to classify.

Logistic model is a statistical model that models the probability of an event taking place. This can be used to classify with a certain degree of precision.

b. Discuss whether you had to preprocess data (e.g., microtext normalization) and why

Before we can perform sentiment analysis on the tweets, the first step that we need to do is preprocessing. Preprocessing is an important step in text classification. It helps eliminate unwanted parts of the data, such as the numbers/punctuation and common words. Good preprocessing can help to achieve better performance on our model.

There are many different kinds of preprocessing tools packages available. Our team used a few of them, such as NLTK, Python Regular Expression (RE), Spacy and Emoji.

Some of the preprocessing techniques that we use:

* represents very important preprocessing technique

1) LowerCase*

Making the text all in lowercase, not only helps to reduce our vocabulary size, it also helps to maintain consistency of the vocabulary dictionary. For instance, ***Hello*** and ***hello*** will be categories as ***hello*** in the vocabulary dictionary.

```
data['cleaned_text'] = data['tweet_text'].apply(lambda x:x.lower())
```

Figure 21. Code on how to perform lowercase on the text

2) Remove punctuations and urls*

Punctuation and urls do not yield any useful information to the data, therefore, it is a good practice to remove it away.

```
def punc_clean(text):

    text= ' '.join(re.sub('(@[A-Za-z0-9_]+)|(&[A-Za-z0-9_]+)|([^0-9A-Za-z \t]) |(\w+:\/\/\S+)' , " ", text).split())
    text = re.sub('['+string.punctuation+']', '', text)
    return text
```

Figure 22. Code on how to remove punctuation as well as urls

3) Remove numbers (Optional)

Removing numbers is an optional process, it depends on the dataset and the problem statement that you're handling.

```
def remove_numbers(text):
    text = re.sub('[0-9]+', '', text)
    return text
```

Figure 23. Code on how to remove numbers

4) Remove stopwords*

Stopwords are those common words such as ***a/the/I/was/were***. All these common words don't add useful information to the model. However, for our problem statement there are some words we must not remove such as ***not/don't/no/cannot***. If we remove these few words, it can completely change the sentence. For example, ***"Pfizer is no good"***, if we remove ***"no"*** from the text, it will become ***"Pfizer is good"***, this will make our model predict positive instead of negative.

```
def remove_stopwords(text):
    stopwords_list = stopwords.words('english')
    # Some words which might you dont want to
    whitelist = ["n't", "not", "no", "don't", "cannot"]
    words = text.split()
    clean_words = [word for word in words if (word not in stopwords_list or word in whitelist) and len(word) > 1]
    return " ".join(clean_words)

text = 'I am Kelvin. Pfizer vaccine not good'

print(f"Before removal of stopwords: {text}\n")
print(f"After removal of stopwords: {remove_stopwords(text)}")
```

```
Before removal of stopwords: I am Kelvin. Pfizer vaccine not good

After removal of stopwords: Kelvin. Pfizer vaccine not good
```

Figure 24. Code on how to remove stopwords and don't remove words that are in the white_list. Sample text to show before and after removal of stopwords

5) Lemmatization*

Lemmatization helps to convert the words to root form, this will greatly help to reduce our vocabulary size. For example, ***goes*** and ***go*** both have a similar meaning, the only difference is that goes is plural and go is singular form. Therefore, there is no point having two words but with different forms. As such, we apply lemmatization on the dataset.

```python
def space(text):
    doc = nlp(text)
    return " ".join([token.lemma_ for token in doc])

text = 'I not goes go to school'

print(f"Before Lemmaztization: {text}\n")
print(f"After Lemmatization: {space(text)}")
```

```
Before Lemmaztization: I not goes go to school

After Lemmatization: I not go go to school
```

Figure 25. Code on how to apply lemmatization on the text using Spacy package, and sample text to show the before and after lemmatization

6) Convert emoji to meaningful words (Optional)

Emoji can add sentiment to a text, for instance the emoji has been converted to word. Smiling face with heart eyes, this can help to imply it is a positive comment.

```python
text = "after getting pfizer dose 😍"
text = emoji.demojize(text, delimiters=("", ""))
text
```

```
'after getting pfizer dose smiling_face_with_heart-eyes'
```

Figure 26. Code on how to convert emoji to meaningful words using the emoji package. Sample text to show after demojize text

	tweet_id	sentiment	tweet_text	cleaned_text
0	1.387628e+18	3	Please when offered a vaccine don't wait for o...	please offer vaccine do not wait one other the...
1	1.392383e+18	3	RT @aminamnzr: "High- and upper-middle income ...	rt high upper middle income country represent ...
2	1.392313e+18	2	2) Starting May 24, Lyft will begin providing ...	start may lyft begin provide ride code cover...
3	1.392235e+18	1	RT @g10dc: @BSimonward Who the heck (unless ex...	rt heck unless extremely vulnerable would take...
4	1.391884e+18	3	RT @Health4AllAmer: Today found out husband of...	rt today find husband staff member school get ...
...	...	...	...	...
8562	1.500000e+18	1	People have stopped taking it after reading ab...	people stop take read side effect year later b...
8563	1.500000e+18	3	Loughborough Hospital have a Covid-19 walk-in ...	loughborough hospital covid walk in vaccine cl...
8564	1.500000e+18	2	It's a clone this one Nkosi yam If it knows ab...	clone one nkosi yam know pfizer report still v...
8565	1.500000e+18	1	Massachusetts pharma company sues Moderna and ...	massachusetts pharma company sue moderna pfize...
8566	1.500000e+18	2	Now add Pfizer and Moderna please	add pfizer moderna please

8298 rows × 4 columns

Figure 27. Data before and after of preprocessing process

The dataset we use for the classification problem is a combination of Kaggle + Huggingface (7.5k), and our own crawled 1k tweets (which we label ourselves). After preprocessing, we managed to get a total of 8298 rows (Shown in Figure 27). The cleaned dataset will then be used to train our model.

After that, to evaluate the performance of our model, we manually labelled another 1k tweet (from our own crawled data). This manually labelled data will then be used to test the performance of our model.

c. Build an evaluation dataset by manually labelling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80%

C	E	I	J	K
username	text	sentiment	sentiment 2	
TonyGriffonbrx	Not everyone that isn't you is controlled opposition in the same way that every event in the world from major catastrophe to the minutiae of our lives is a conspiracy or that EVERY death forever more is due to Covid19 vaccine injury Hope this helps	POSITIVE	POSITIVE	POSPOS
PPDCRO	We have been at the forefront of enhancing immunogenicity methodologies for three decades and our new research is no exception In a new blog learn how our expert approach streamlines vaccine development projects	POSITIVE	POSITIVE	POSPOS
JoyShiningSta	Russia called out the Swiss listed Liechtenstein as Putin is coming to terms w what his legacy is San Marino was given the Sputnik vaccine which allowed Russian tourism there is taking a diplomatic jab at San Marino bc they used to be friends hehe	NEUTRAL	NEUTRAL	NEUNEU
mybmcWardM	List of centers in the ward and related details about vaccination for tomorrow Time 9 am to 5 pm Covishield Dose 1St 2nd and Precaution dose Covaxin Dose 15 to 18 years	NEUTRAL	NEUTRAL	NEUNEU
LynnNMonteitl	I just love this Moderna is next Promise	POSITIVE	POSITIVE	POSPOS
sammy44231	pfizer mortality for under-65s is around 3 1 higher than in 2020 There have been around 120 400 more deaths from all causes than expected in the UK from the start of the pandemic to 31 December 2021 Of these 72 900 occurred in 2020 and 47 500 in 2021	NEGATIVE	NEGATIVE	NEGNEG
McDutchoven	Flynn chose to speak as Q messenger so did Mike They can all wear it because frankly the topic is so stupid that nobody gives a shit who did what in what order Racketeering doesn't depend on one person furthering the crime knowing what every 1 of their codefe	NEUTRAL	NEUTRAL	NEUNEU
c3_cc3	Why Pfizer Stock Jumped While the Market Stumbled Today NASDAQ	POSITIVE	NEUTRAL	POSNEU
Sue67374532	pfizer Keep reaching out for help Pray Jesus loves you and hears your cries for help Have daily calls with supportive friends family pastors vax injured support groups whoever will listen Lots of good suggestions here from people who care Also Try	POSITIVE	POSITIVE	POSPOS
djhahn13	pfizer Absolutely horrible man It's better that you are here then gone I can't imagine the struggle	NEGATIVE	NEGATIVE	NEGNEG
technikest	Marketing your business costs money A LOT of money sixmajor programa4por4 lapijamadadefabiola animalkingdom ourgameisback winner nachosinfonico gebzeescort pfizer florida asroma	NEUTRAL	NEUTRAL	NEUNEU
rapturousrat	pfizer The testimonials are mind-blowing and its effects are downright miraculous I hope you will try it and let us know what you experience Here's more info - - -	POSITIVE	POSITIVE	POSPOS
mymoneytips	Ocugen Latest News What OCGN Stock Investors Need to Know About the FDA's Covaxin Rejection entrepreneur forextips money signals tradingstrategy forexsignals invest trader fx	NEGATIVE	NEGATIVE	NEGNEG
chupakarma6(	If you donate for Ukraine all charges will be dropped no matter what you did We must fight for democracy and vaxination and make marriage between men and women illegal Having kids must be illegal Kids must be made in the health labs at Moderna and Pfizer t	NEGATIVE	NEGATIVE	NEGNEG
DoctorLunge	You're being melodramatic pseudouridine is found naturally all over the shop There are two stop codons in pfizer and three in Moderna most mRNAs have some redundancy	NEGATIVE	NEGATIVE	NEGNEG
BonVangUFO	UK coroners in two separate cases concluded individuals who received AstraZeneca's COVID-19 vaccine died from blood-clotting disorders caused by the vaccine which uses the same adenovirus technology as the Johnson Johnson COVID vaccine	NEGATIVE	NEGATIVE	NEGNEG
HappyClown_	Getting the population ready for the new improved 4th Pfizer jab that protects against Omicron due to be ready for the FDA in March I'd like to see a public record of the politicians jabs	POSITIVE	POSITIVE	POSPOS
gypsyvampry	The vaccine wasn't created to save you from the virus it was created to depopulate genocide This entire plandemic wasn't about your health it was about pushing their agenda The Great Reset and money How much did Pfizer and Moderna make from the vax	NEGATIVE	NEGATIVE	NEGNEG
4frespeech1	Pfizer Released List of Over 1000 Covid Vax-Injuries While World Is Distracted Vincent James Red Elephants LBRY via	NEUTRAL	NEUTRAL	NEUNEU
alembic	As someone with onset of tinnitus and other issues after 2 doses of the Pfizer vaccine I am glad to see that some attention and interest is finally being paid to this	NEUTRAL	NEUTRAL	NEUNEU
Michael35335(	This MFer murder my mother with his death jab back in the end of January 2021 My little brother is magnetic from the Moderna death Jab This piece of crap needs to be locked up and the blood money he's made be returned to the we people he has stolen it from	NEGATIVE	NEGATIVE	NEGNEG
TaylordforTasb	What does Pfizer have to hide bonniehenry	NEGATIVE	NEGATIVE	NEGNEG
KiwiLithium	I like how the original data from pfizer said it didn't but magically when we got vaccines all of a sudden they told us it did	NEGATIVE	NEGATIVE	NEGNEG
amyriscool	As I understand it China is using the same vaccine as Hong Kong The Pfizer and Moderna and J J are really the go-to	POSITIVE	POSITIVE	POSPOS
Michael75785(	I can do you one better As of about a week ago quadruple vaxxed with Moderna with doctor's encouragement and no Covid yet	POSITIVE	POSITIVE	POSPOS
OliveNewton8(	I take Covid seriously Just because I'm not vaccinated doesn't mean I don't I wear a mask I stayed home for 2 years I've never mocked anyone for getting the vaccine This guy got cocky and over zealous My point was always stay humble whether vaxxed or not	POSITIVE	POSITIVE	POSPOS
LEB_3_168	The more government promote covid vaccine the more I runaway	NEGATIVE	NEGATIVE	NEGNEG
MarkHubbard2	Even against Alpha per Oxford university calculator my survival probability was 99 9887 Why would I risk Pfizer Read my pinned post	NEGATIVE	NEGATIVE	NEGNEG
MAAWLAW	Public information about nirmatrelvir the new drug in 's combination therapy has been relatively limited and manufacturers have had to wait for licenses and Pfizer's original product to become available… COVID19 compliance procurement	NEUTRAL	NEUTRAL	NEUNEU
Emirati_News	10 598 doses of the COVID-19 vaccine administered during past 24 hours MoHAP	NEUTRAL	NEUTRAL	NEUNEU
Steve_Byra	How is this so hard for you to understand You asked which pharma company promotes masks I prvided links to Pfizer you then asked if they sold masks I said I bet J J does and provided a link	NEUTRAL	NEUTRAL	NEUNEU
bornwithnoleg	Why do you think Sinovac's vaccine is not as good just trying see whats missing compared to others	NEGATIVE	NEGATIVE	NEGNEG
HSI_CC_Unite	RE-POST Help keep students in the classroom on the playground by getting your children 5 yrs vaccinated against COVID19 Have your family SleeveUp for school get all their COVID-19 vaccines incl a booster Find a COVID-19 vaccine near you	POSITIVE	POSITIVE	POSPOS
Eddie_J_G	Claim 1200 people died in the 3 month Pfizer trial Verdict False Those 1200 people died in the first 3 months of the EUA	NEGATIVE	NEGATIVE	NEGNEG
ColinDotNet	I just saw a commercial for Oral Treatments for covid by pfizer cdc No funding They've got ads though for a treatment they can't buy	NEGATIVE	NEGATIVE	NEGNEG
SharonMcMer	Will you ask your TDs to stand in solidarity with those most at risk from Covid-19 by ensuring Ireland supports VaccineEquity Take action	POSITIVE	POSITIVE	POSPOS
Newsweek	Moderna Pfizer push for 4th doses despite only minor benefits in study	NEUTRAL	NEUTRAL	NEUNEU
MattFrank	For anyone looking for the possible reason for San Marino being included This NYT article from last year shows the close ties they had including sharing vaccines I'm guessing Russia isn't happy they aren't repaying what it sees as favors	NEUTRAL	NEUTRAL	NEUNEU
xMonsieurPlus	I really hope you get better soon brother I also hope your voice and the ones of many people in your situation gets heard Sadly pfizer has ruined the lives of way to many young people but there is nobody out there to take care of them now or even to admit it ❤️🙏	NEGATIVE	NEGATIVE	NEGNEG
cityoflondon68	Poorly You read the list of known side effects of this jab released that Pfizer wanted sealed for 75 years yet Poorly🤕🤦	NEGATIVE	NEGATIVE	NEGNEG
snufflelove	In what may be his response to the inconvenient discovery of a key gene sequence of the COVID-19 virus in a 2016 patent bearing his name Moderna CEO Stephane Bancel reveals what many have suspected all along that SARS-CoV-2 is a result of a lab leak	NEGATIVE	NEGATIVE	NEGNEG
WUp2022	pfizer The FDA is crooked No recall on metal on metal hip implants so consumers could learn of the poisoning they could get Also how did Biomet keep their name off the legal commercials for these devices That's what I had and didn't know it Lost support tissue t	NEGATIVE	NEGATIVE	NEGNEG
doc_holliday5(	If you dig further I bet you'll find that those same agencies were also paid to blocked or shadow ban news or posts regarding alternatives withheld from Americans like Covaxin a traditional non mRNA vax available since last May STILL not available	NEGATIVE	NEGATIVE	NEGNEG
Stashk311	It's time to give covaxin a chance It's been delayed enough Thanks for speaking about it in the past Dr Hildreth	POSITIVE	POSITIVE	POSPOS
BHPanimalwa	And the United States has still made no move to suspend or cancel the Covid19 Vaccine Patents or enabled any vitally needed technology transfer to enable vaccine manufacturing despite the transparently self-serving global distribution claims of	NEGATIVE	NEGATIVE	NEGNEG

Figure 28. Excel file of the manually labelled 10% of the collected data

Figure 28 shows a portion of our manually labelled tweets. We first have the classification team to manually label the sentiments, then two of the classification team members to re-label tweets that they did not label (first sentiment column is hidden to ensure a valid labelling). To ensure the manually labelled data is not biased, we use Kappa score as metrics to evaluate the inter annotator agreement score.

		Judge 2 Relevance			
		POSITIVE	NEUTRAL	NEGATIVE	Total
Judge 1 Relevance	POSITIVE	141	24	2	167
	NEUTRAL	1	401	2	404
	NEGATIVE	1	11	444	456
	Total	143	436	448	1027

Table 2. Results for the total number of matches between Judge 1 and 2

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$$P(A) = \frac{141 + 401 + 444}{1027} = 0.96007$$

$$P(E) = \frac{167 + 143}{1027 + 1027}^2 + \frac{404 + 436}{1027 + 1027}^2 + \frac{456 + 448}{1027 + 1027}^2 = 0.38373$$

$$Kappa = (0.96007 - 0.38373) - (1 - 0.38373) = 0.93522$$

Figure 29. Calculation of Kappa score

The team got P(A) = 0.96007 and P(E) = 0.38373 (shown in Figure 29) after computing the relevance from Table 2.

After discounting the P(E) from the P(A), the team got the Kappa score of 0.93 (as seen in Figure 29) which is above 0.8 and therefore there is a good inter-annotator agreement.

d. Provide evaluation metrics such as precision, recall, and F-measure and discuss results Discuss performance metrics, e.g., records classified per second, and scalability of the system

The team has provided precision, recall, and F-measures for each individual model (BERT-CNN, LSTM, RoBERTa and Ensemble). The precision value tells us of all tweets that were labelled as a

sentiment (e.g. Neutral), how many actually shared the same sentiment (e.g. Neutral). The recall value tells us of all tweets that are truly Negative (or Neutral or Positive), how many did we label correctly. The F-measure then balances both precision and recall to a number.

1. BERT-CNN

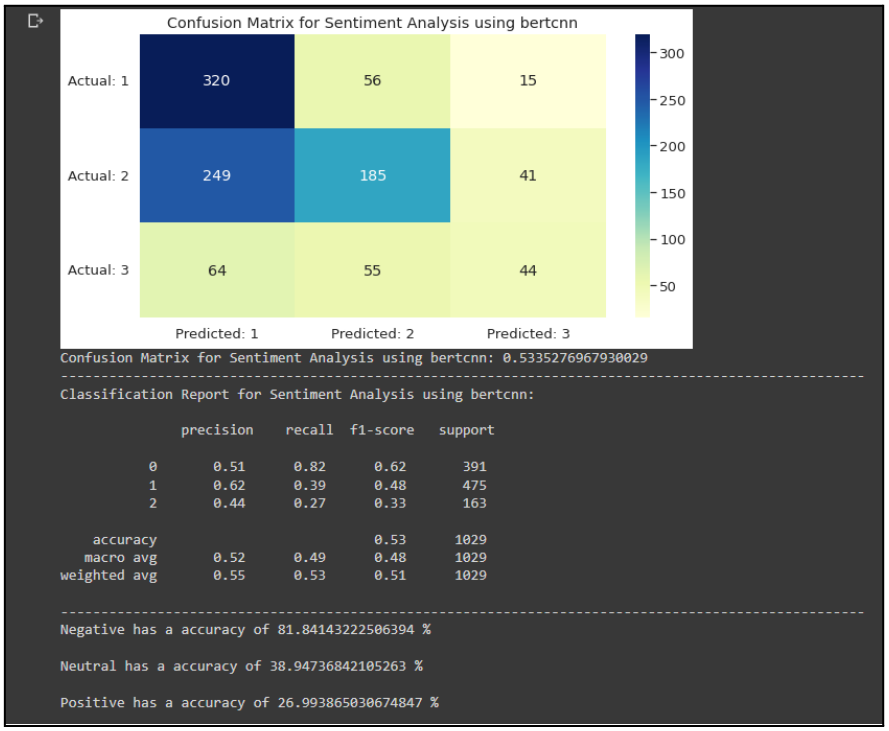


Figure 30. Confusion matrix and report for BERT-CNN

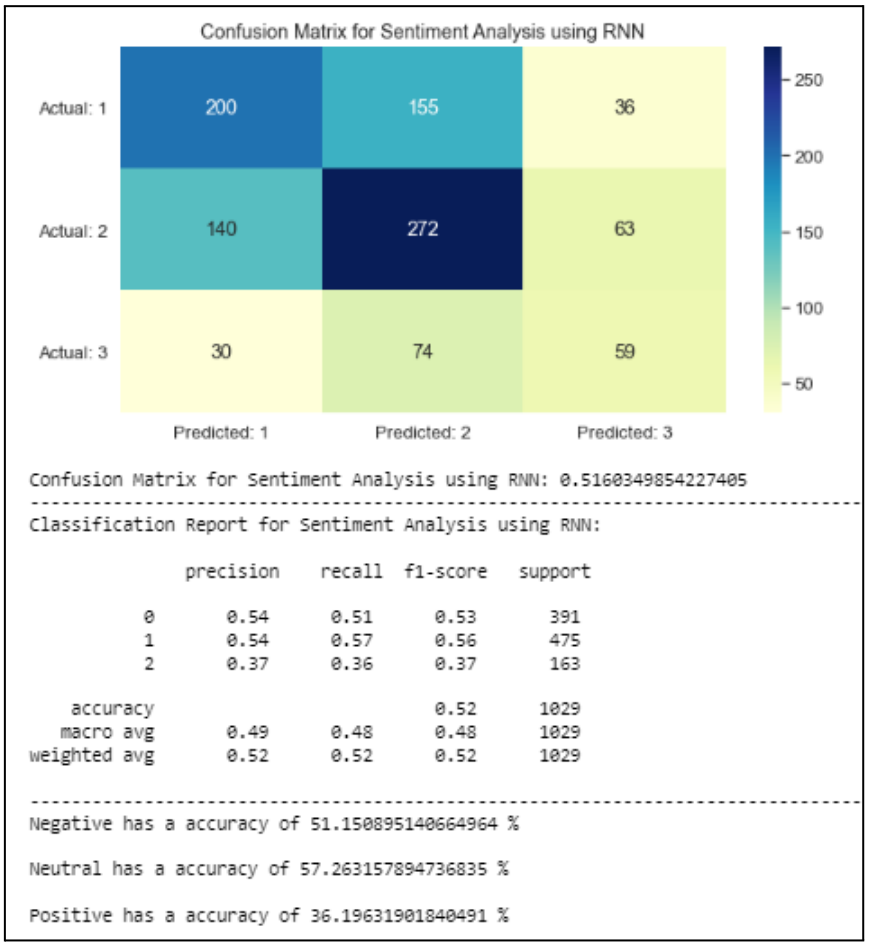
```
[243] start_time = time.time()
      for i in test_y_text:
        pred_lst.append(int(LABEL.vocab.itos[predict_sentiment(model,tokenizer, i)]))
      print('Total predictions', len(pred_lst))
      end_time = time.time()

      elapsed_time = end_time - start_time
      print('Elapsed time: ', elapsed_time)
      print('Classification per second', len(pred_lst)/elapsed_time)

      Total predictions 1029
      Elapsed time:  19.973747730255127
      Classification per second 51.51762272642143
```

Figure 31. Time taken for the BERT-CNN model to classify

In the Classification Report section (Figure 30), 0, 1 and 2 represent Negative, Neutral and Positive respectively. We can note that the algorithm that predicts and classifies Negative is liberal as precision is 0.51 and recall is 0.82, meaning that it classifies a lot of non-Negative as Negative sentiment. The precision and recall scores of Neutral and Positive are not high, therefore, the model is not able to accurately predict both sentiments. Since the classification per second is higher than most of the

models listed here (51), therefore, when there are more tweets to be classified, it should not have much impact.

2. LSTM




```
Confusion Matrix for Sentiment Analysis using RNN: 0.5160349854227405
--------------------------------------------------------------------------------
Classification Report for Sentiment Analysis using RNN:

              precision    recall  f1-score   support

           0       0.54      0.51      0.53       391
           1       0.54      0.57      0.56       475
           2       0.37      0.36      0.37       163

    accuracy                           0.52      1029
   macro avg       0.49      0.48      0.48      1029
weighted avg       0.52      0.52      0.52      1029

--------------------------------------------------------------------------------
Negative has a accuracy of 51.150895140664964 %

Neutral has a accuracy of 57.263157894736835 %

Positive has a accuracy of 36.19631901840491 %
```

Figure 32. Confusion matrix and report for LSTM

```python
# Calculate the time of how fast it predict
import time
y_pred = []
start_time = time.time()

for text in test_cleaned['cleaned_text']:
    sentiment = predict_sentiment(text, token)
    y_pred.append(sentiment)

print("Total predictions", len(y_pred))
end_time = time.time()

elapsed_time = end_time - start_time
print(f"Elapsed time: {elapsed_time}")
print(f"Classification per seconds: {len(y_pred)/elapsed_time}")
```

```
Total predictions 1029
Elapsed time: 51.394134521484375
Classification per seconds: 20.021740021127226
```

Figure 33. Time taken for the BI-LSTM model to classify

In the Classification Report section in Figure 32, 0, 1 and 2 represent Negative, Neutral and Positive respectively. The precision and recall scores for Negative and Neutral are about the same 54-57%. However, the precision and recall for positive is very low 37%. Therefore, the model is not able to accurately predict all sentiments. Since the classification per second is high (20), therefore, when there are more tweets to be classified, it should not have much impact.

3. Roberta

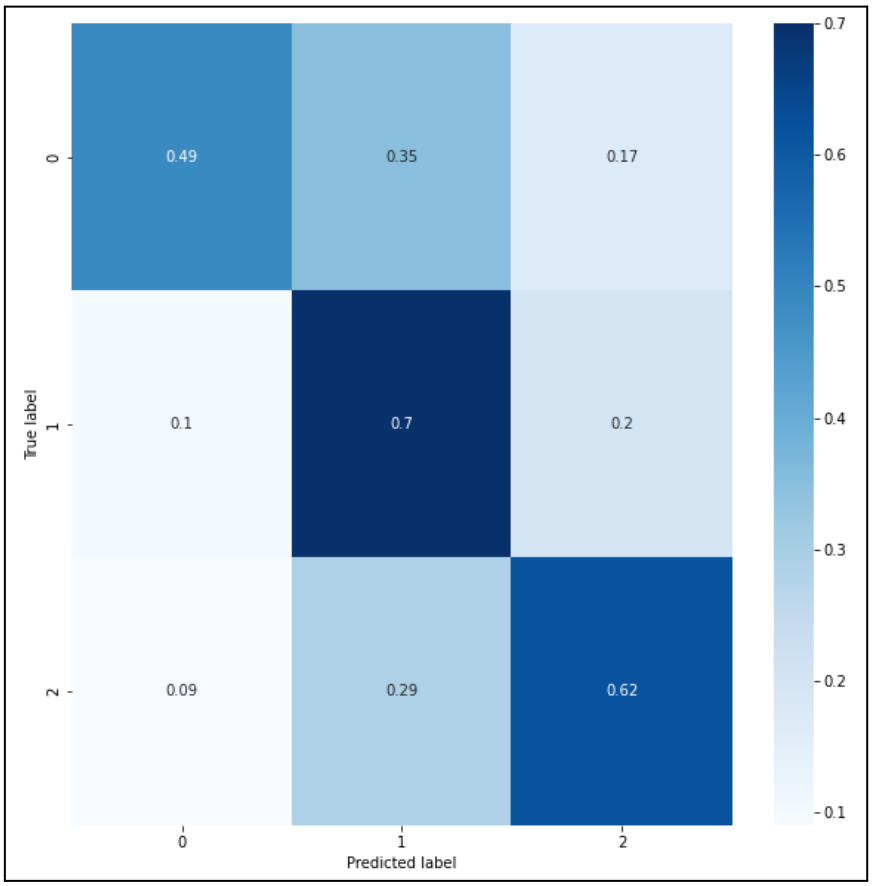


Figure 34. Confusion matrix for RoBERTa

```
Confusion Matrix for Sentiment Analysis using <keras.engine.functional.Functional object at 0x7f74540eac50>: 0.6343042071197411
-----------------------------------------------------------------------------------------------------
Classification Report for Sentiment Analysis using <keras.engine.functional.Functional object at 0x7f74540eac50>:

              precision    recall  f1-score   support

           0       0.61      0.78      0.68       391
           1       0.72      0.53      0.61       475
           2       0.54      0.61      0.57       163

    accuracy                           0.63      1029
   macro avg       0.62      0.64      0.62      1029
weighted avg       0.65      0.63      0.63      1029

-----------------------------------------------------------------------------------------------------
Negative has a accuracy of 77.87610619469027 %

Neutral has a accuracy of 53.06122448979592 %

Positive has a accuracy of 61.224489795918366 %
```

Figure 35. Confusion report for RoBERTa

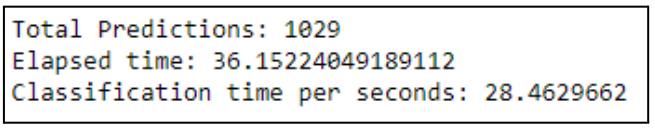
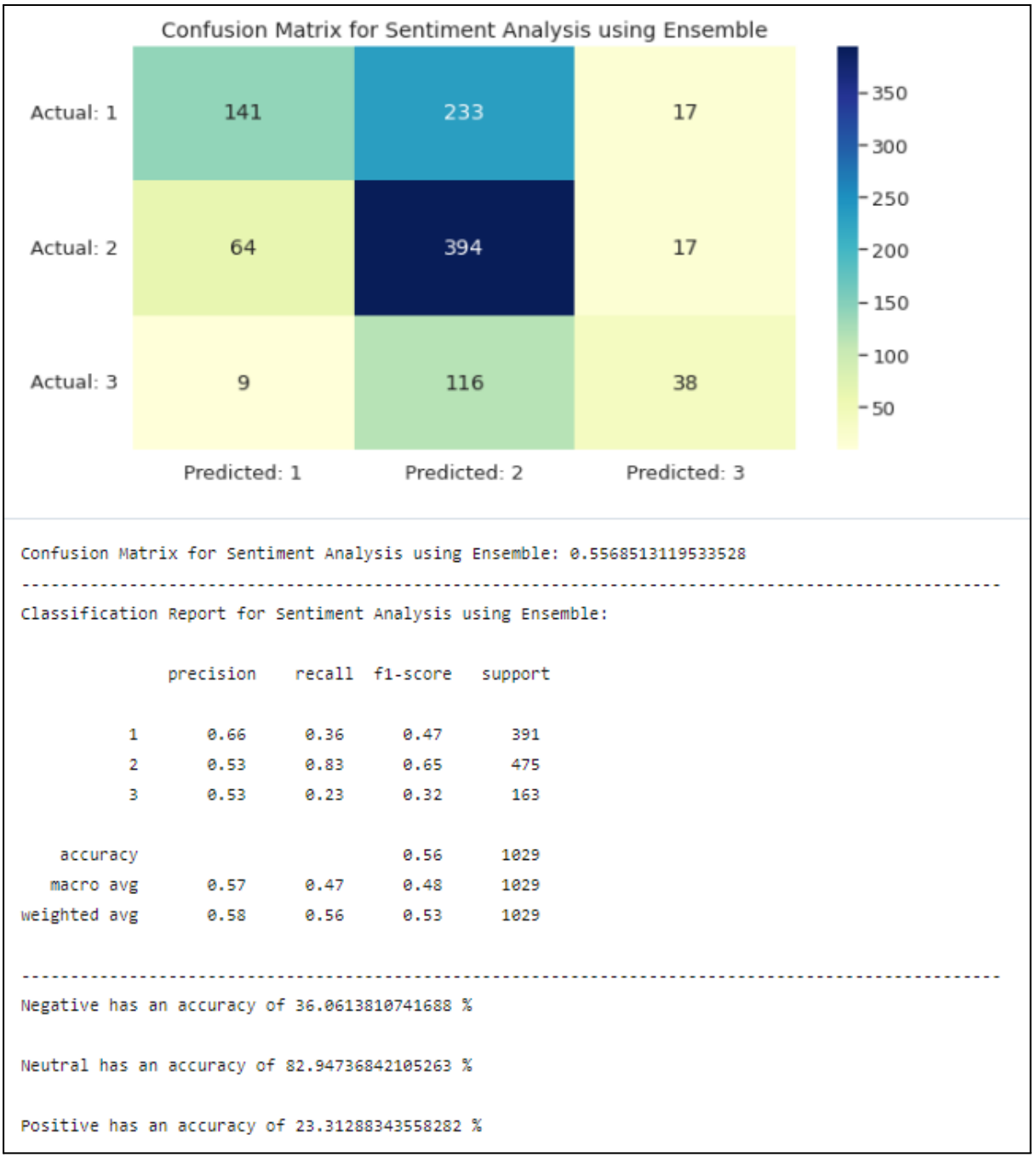


Figure 36. Time taken for the RoBERTa model to classify

In the confusion report in Figure 34, 0, 1 and 2 represent Negative, Neutral and Positive respectively. We can note that the algorithm that predicts and classifies Negative is liberal as precision is 0.61 and recall is 0.78, meaning that it classifies a lot of non-Negative as Negative sentiment. Overall, the precision, recall and F1 scores are not particularly high but the statistics compared to other models demonstrates that RoBERTa is the best.

4. Ensemble



Confusion Matrix for Sentiment Analysis using Ensemble: 0.5568513119533528

Classification Report for Sentiment Analysis using Ensemble:

	precision	recall	f1-score	support
1	0.66	0.36	0.47	391
2	0.53	0.83	0.65	475
3	0.53	0.23	0.32	163
accuracy			0.56	1029
macro avg	0.57	0.47	0.48	1029
weighted avg	0.58	0.56	0.53	1029

Negative has an accuracy of 36.0613810741688 %

Neutral has an accuracy of 82.94736842105263 %

Positive has an accuracy of 23.31288343558282 %

Figure 37. Confusion matrix for Ensemble model

```
start_time = time.time()
Y_pred = clf.predict(X_data)
end_time = time.time()
elapsed_time = end_time - start_time
print('Number of predictions: ', len(Y_pred))
print('Elapsed time: ', elapsed_time)
print('Classification per second: ', len(Y_pred)/elapsed_time)
```

```
Number of predictions:  1029
Elapsed time:  572.8769388198853
Classification per second:  1.79619728125157
```

Figure 38. Time taken for the Ensemble model to classify

In the confusion report in Figure 37, 1, 2 and 3 represent Negative, Neutral and Positive respectively. We noted that the algorithm that predicts and classifies Neutral is liberal as precision is 0.53 and recall is 0.83, meaning that it classifies a lot of non-Neutral as Neutral sentiment. The precision and recall scores of Negative and Positive are not high, therefore, the model is not able to accurately predict both sentiments. The classification per seconds is lower than most of the models listed here (1), this could be due to the extra computation needed to run the different models during the ensemble learning.

4.2. Summary

From the above figures, we can see that the Roberta model clearly outperforms the rest of the model with its stable precision, recall as well as F1-score. The model will be used for sentiment analysis of the text corpus and later will classify newly crawled tweets before indexing them into Solr.
The obtained results can be found in the table below.

Model	Avg. Precision	Avg. Recall	Avg. f1-score	Time spent to predict the manually labelled data	Number of tweets classified per seconds
BERT-CNN	0.53	0.49	0.48	19.97s	51
LSTM	0.49	0.48	0.48	51.39s	20
Roberta	0.62	0.64	0.62	36.15s	28
Ensemble	0.57	0.47	0.48	572.88s	1.34

Table 3. The obtained results of all the models used

These obtained results were not very high. There are a few possible explanations:

1. Incorrect Sentiment Classification

Our task is to determine whether a tweet is positive, neutral or negative using sentimental analysis. As a result, the sentiment classification must be credible and reliable. However, the classification process is very reliant on the individuals performing it.
For example, the tweet "***I'm vaxxed and boosted with Pfizer and not had COVID*** 👍" should have been logically labelled as 'positive'; however, in Kaggle, this was labelled 'negative'.

2. Insufficient Dataset

The size of our classification dataset is too small. As a result, we are unable to create a comprehensive word dictionary. With a small vocabulary dictionary, we might predict incorrectly especially when dealing with data that has never been seen before.

Through this classification problem, we understand that sentiment analysis on tweets is not an easy task. There are a lot of things we need to consider such as whether the labelled tweets are credible or not. Thus, it is always good to find datasets that have been inter-annotated.

5. Conclusion

The team presented an information retrieval system that is able to provide results depending on search queries, selected filters and sort type. When there is a spelling mistake in the query, the system is able to provide a spell correction suggestion. In addition, the information retrieval system is able to provide a sentiment analysis to the tweet and display both the topic's sentimental distribution as well as the most common words pertaining to the queried topic. Even though the deep learning models we implemented do not achieve (or even come close to) the accuracy of competitions,we believe that this work will help vaccine companies and government entities better understand what the public's concerns and sentiments are in order to increase vaccination rates in the event of another pandemic. We discovered that transparency is also a highly regarded principle. People look out for companies that are open about their vaccines because openness allows for easy knowledge, which makes people happy.

6. Contribution

Subteam	Initial members	Final members
Frontend	Damien, Duc, Eric	Van, Damien
Backend - Solr	Eng Hao, Kelvin, Van	Eng Hao, Kelvin, Van
Backend - Flask	Van	Van
Classification	All	Eng Hao, Kelvin, Eric, Duc

Tasks	Subtasks	Done by the end of	Member who claimed to do	Member who did	Status by submission
Data collection	Crawl data	Week 8 - YES	Kelvin, Eng Hao	Kelvin, Eng Hao	Done
	Clean data	Week 8 - YES	Kelvin, Eng Hao	Kelvin, Eng Hao	Done
	Remove bots, duplicate - ensure data quality	Week 8 - YES	Kelvin, Eng Hao	Kelvin, Eng Hao	Done
Solr as backend for searching	Config for preprocess data - stemming, tokenize	Week 8 - YES	Van, Eng Hao	Van, Eng Hao	Done
	Config for spell checking	Week 8 - YES	Van	Van	Done
	Config for phrase query	Week 8 - YES	Van	Van	Done
	Config for updating data without duplicates	Week 10 - YES	Van	Van	Done
Flask Backend	Crawling new tweets script	Week 10 - YES	Kelvin	Kelvin	Done
	Visualisation:	Week 11 - YES	Van, Damien	Van, Damien	Done

	WordCloud, charts				
	Integrate classifier	Week 11 - NO	Van	Van	Done
Frontend	Create static UI elements	Week 8 - NO	Damien, Duc, Eric	Van, Damien	Done
Frontend-backend integration	Integrate Solr, Flask, ReactJS	Week 11 - YES	Frontend and backend team	Van	Done
Classification task	Data collection + preprocess	Week 10 - YES	Van	Van	Done
	Manual labelling	Week 11 - YES	Classification team	Classification team	Done
	Inter-annotator	Week 11 - YES	Eng Hao, Kelvin	Eng Hao, Kelvin	Done
	BERT + CNN	Week 10 - YES	Eng Hao, Van	Eng Hao	Done
	LSTM	Week 10 - YES	Kelvin	Kelvin	Done
	RoBERTa	WEEK 10 - NO	Eric	Eric	Done
	Ensemble	WEEK 10 - NO	Duc	Duc	Done
Report		Week 12 - YES	All	All	Done
Presentation		Week 12 - YES	All	All	Done
Video		Week 12 - YES	All	Duc, Eric	Done

7. Submission Link

1. Youtube video: https://youtu.be/3MM8Nxhb4WQ
2. A Dropbox (or Google Drive) link to a compressed (e.g., zip) file with crawled text data, queries and their results, manual classifications, automatic classification results, and any other data for Questions 3 and 5:
   Folder file:
   https://drive.google.com/drive/folders/136niIsgfWjNIU74QJGT0-qbfNL3hMhzd?usp=sharing
   Zipped file:
   https://drive.google.com/file/d/1dlvHt4gEA0bx6u0vntvJdZ56px8HhCZw/view?usp=sharing
   The files inside the zipped folder are as followed:

- automatic_classification_labelled_tweets: text classified by our model
- crawled_text_corpus_data: all the crawled text data from Twitter for our system
- manually_classified_training_data: manually classified data used for training the model
- manually_classified_validation_dataset: manually classified data as evaluation dataset
- preprocessed_kaggle_data_for_question4: Additional pre-processed data from Kaggle to be used for training the model

3. A Dropbox (or Google Drive) link to a compressed (e.g., zip) file with all your source codes and libraries, with a readme file that explains how to compile and run the source codes

   Source code: https://drive.google.com/file/d/1QQZiR6Ke9KtHLg-t_tDTJYaimxgwaq6Y/view?usp=sharing

   Weights: https://drive.google.com/file/d/1aKA5P4t7iA-gcCgOFpIXIWKYIz3i3e3Y/view?usp=sharing

   In case above zip files cannot be opened:

   - The drive folder (weights + notebooks) for model parts is there: https://drive.google.com/drive/folders/10kl7DZ4wr5qKvBDq1dzzEfFfbJiuGdbB?usp=sharing
   - GitHub repo (with readme.md): CZ4034: Tweet IR for Vaccine Reaction

Appendix

1. Solr Search results examples:

Query	Solr results

pfizer	http://localhost:8983/solr/CZ4034/select?indent=true&q.op=OR&q=text%3A%20pfizer
	{
	"responseHeader":{
	"status":0,
	"QTime":10,
	"params":{
	"q":"text: pfizer",
	"indent":"true",
	"q.op":"OR",
	"_":"1649338548349"}},
	"response":{"numFound":4489,"start":0,"numFoundExact":true,"docs":[
	{
	"tweet_id":[1502421069818314754],
	"user_id":[1162912958944428032],
	"username":"ServitudeClass",
	"like_count":[1],
	"retweet_count":[0],
	"location":["unhoused"],
	"creation_date_time":["2022-03-11T00:00:00Z"],
	"text":"pfizer moderna exec jail life fdahaslostallcredibility pfizergate pfizerdocument pfizersideeffect pfizer pfizerdata",
	"sentiment":["POSITIVE"],
	"id":"a0bc8b13-6aff-43a6-a6e5-b56f2a37fffc",
	"_version_":1728832431171043401},
	{
	"tweet_id":[1502772858359009284],
	"user_id":[23446256],
	"username":"weegan19",
	"like_count":[0],
	"retweet_count":[0],
	"location":["US"],
	"creation_date_time":["2022-03-12T00:00:00Z"],
	"text":"prediction vax manufacturer stock pfizerfraud pfizergate pfizerdata pfizer pfizerleak modernafraud moderna vaccinefraud pfe mrna clotsh
	"sentiment":["NEUTRAL"],
	"id":"7ed342c5-e85d-428c-812e-d95427f41656",
	"_version_":1728832431024242724},

moderna	http://localhost:8983/solr/CZ4034/select?indent=true&q.op=OR&q=text%3A%20moderna
	{
	"responseHeader":{
	"status":0,
	"QTime":6,
	"params":{
	"q":"text: moderna",
	"indent":"true",
	"q.op":"OR",
	"_":"1649338548349"}},
	"response":{"numFound":2200,"start":0,"numFoundExact":true,"docs":[
	{
	"tweet_id":[1502730298819293187],
	"user_id":[16631823],
	"username":"MaryKRe",
	"like_count":[0],
	"retweet_count":[0],
	"location":["Ann Arbor, MI"],
	"creation_date_time":["2022-03-12T00:00:00Z"],
	"text":"I need update asap pfizer booster pfizer shot dose increase moderna booster moderna shot dose increase pfizer booster modernas increase
	"sentiment":["NEUTRAL"],
	"id":"9f1c027c-211b-4d15-a279-f7d6db72b09f",
	"_version_":1728832431039971385},
	{
	"tweet_id":[1502505805970817026],
	"user_id":[1492559136592207880],
	"username":"maria2019xz3",
	"like_count":[0],
	"retweet_count":[0],
	"location":["No Location"],
	"creation_date_time":["2022-03-12T00:00:00Z"],
	"text":"wrong good oh might miss profit moderna hi moderna hi moderna blood hand investor enjoy bloody dripping hand profit make moderna stock
	"sentiment":["NEUTRAL"],
	"id":"68a5009e-3db9-4a58-8ee2-e4424ab902aa",
	"_version_":1728832431141683207},
	{

sinovac	http://localhost:8983/solr/CZ4034/select?indent=true&q.op=OR&q=text%3A%20sinovac
	{
	"responseHeader":{
	"status":0,
	"QTime":1,
	"params":{
	"q":"text: sinovac",
	"indent":"true",
	"q.op":"OR",
	"_":"1649338548349"}},
	"response":{"numFound":807,"start":0,"numFoundExact":true,"docs":[
	{
	"tweet_id":[1502954945712033792],
	"user_id":[1231486724087222272],
	"username":"Transit_Jam",
	"like_count":[0],
	"retweet_count":[0],
	"location":["Hong Kong"],
	"creation_date_time":["2022-03-13T00:00:00Z"],
	"text":"nursing home provide sinovac delay week guy could biontech give doctor say unsuitable sinovac antisinovac propaganda misinformation work
	"sentiment":["NEUTRAL"],
	"id":"f068c287-0a4b-4b96-8bd8-fc38a34c5c87",
	"_version_":1728832431583137 37},
	{
	"tweet_id":[1502421191956611077],
	"user_id":[1225820672],
	"username":"Amine19O",
	"like_count":[1],
	"retweet_count":[0],
	"location":["No Location"],
	"creation_date_time":["2022-03-11T00:00:00Z"],
	"text":"make make sinovac come I aware obligation sort yo vaccinate sinovac death hk among vaccinated come number low considering sinovac ineffe
	"sentiment":["NEGATIVE"],
	"id":"eb1c0d10-8bd2-41e0-b490-9e7782ab26f4",
	"_version_":1728832431599911012},

2. Solr Spellcheck and suggestions examples:

Query	Solr spellcheck results
vaccime	http://localhost:8983/solr/CZ4034/spell?indent=true&q.op=OR&q=vaccime&spellcheck=true


```
{
  "responseHeader":{
    "status":0,
    "QTime":4},
  "response":{"numFound":0,"start":0,"numFoundExact":true,"docs":[]
  },
  "spellcheck":{
    "suggestions":[
      "vaccime",{
        "numFound":8,
        "startOffset":0,
        "endOffset":7,
        "origFreq":0,
        "suggestion":[{
            "word":"vaccine",
            "freq":4106},
          {
            "word":"vacccine",
            "freq":1},
          {
            "word":"vaccigen",
            "freq":11},
          {
            "word":"vaccinea",
            "freq":1},
          {
            "word":"vaccinee",
            "freq":2},
          {
            "word":"vacciner",
            "freq":1},
          {
            "word":"vaccines",
            "freq":4},
          {
            "word":"vaccin",
            "freq":7}]}],
    "correctlySpelled":false,
    "collations":[]}}
```

modernq	http://localhost:8983/solr/CZ4034/spell?indent=true&q.op=OR&q=modernq&spellcheck=true


```
{
  "responseHeader":{
    "status":0,
    "QTime":6},
  "response":{"numFound":0,"start":0,"numFoundExact":true,"docs":[]
  },
  "spellcheck":{
    "suggestions":[
      "modernq",{
        "numFound":8,
        "startOffset":0,
        "endOffset":7,
        "origFreq":0,
        "suggestion":[{
            "word":"modern",
            "freq":9},
          {
            "word":"moderna",
            "freq":2037},
          {
            "word":"modernad",
            "freq":1},
          {
            "word":"modernas",
            "freq":64},
          {
            "word":"morderna",
            "freq":2},
          {
            "word":"xmoderna",
            "freq":1},
          {
            "word":"modena",
            "freq":3},
          {
            "word":"modrna",
            "freq":1}]}],
    "correctlySpelled":false,
    "collations":[]}}
```

covacin	http://localhost:8983/solr/CZ4034/spell?indent=true&q.op=OR&q=covacin&spellcheck=true


```
{
  "responseHeader":{
    "status":0,
    "QTime":3},
  "response":{"numFound":0,"start":0,"numFoundExact":true,"docs":[]
  },
  "spellcheck":{
    "suggestions":[
      "covacin",{
        "numFound":2,
        "startOffset":0,
        "endOffset":7,
        "origFreq":1,
        "suggestion":[{
            "word":"covaxin",
            "freq":1362},
          {
            "word":"covaccine",
            "freq":1}]}],
    "correctlySpelled":false,
    "collations":[]}}
```