L-Università ta' Malta
Faculty of Information &
Communication Technology

Individual Assigned Practical Task – Sentiment Analysis on Covid-19 Vaccine Uptake in EU

Karl Attard 203501(L)
B.Sc. (Hons) Artificial Intelligence

Study-unit: Individual Assigned Practical Task
Unit Code: ARI2201
Lecturer: Dr Claudia Borg

Google Drive Link for Code:
https://drive.google.com/drive/folders/1tzdrOhlWNYNhF07Il2ps0jnws4Jzuteu?usp=sharing

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as “the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

We, the undersigned, declare that the assignment submitted is our work, except where acknowledged and referenced.

We understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected and will be given zero marks.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Karl Attard

Student Name Signature

ARI2201 **Sentiment Analysis on Covid-19 Vaccine Uptake in EU**

Course Code Title of work submitted

14/06/2021

Date

Table of Contents

Introduction

The proposed idea for the IAPT is to perform sentiment analysis on a corpus of tweets related to covid-19 vaccine uptake in Europe. The main objective for this project is to compare what people are feeling about the vaccines before and after blood-clots were concluded as a rare side effect to a particular vaccine. Hence, I will be analysing how this phenomenon has affected people's thoughts and emotions with regards to the covid-19 jab, focusing only on three countries, Great Britain, Italy and Spain.

In this project, only the Twitter API was used. Therefore, the first thing which was done was to grasp knowledge on this API to allow me to collect relevant tweets from Twitter. Then, I used existing Sentiment Analysis library to determine whether a tweet is either positive, neutral, or negative. From these results obtained, I will then be analysing the relation of the sentiment to the vaccine uptake for the countries. This will be achieved through bar charts, word clouds and a pie chart (will be discussed further below).

Every project has their own challenges and limitations. One main limitation that this project has is that it considers only tweets in English. Therefore, this can be improved in future work to collect tweets in any language whatsoever and translate them to English (since Sentiment Analysis models work with English text).

Design/Summary of Functionality

Flowchart

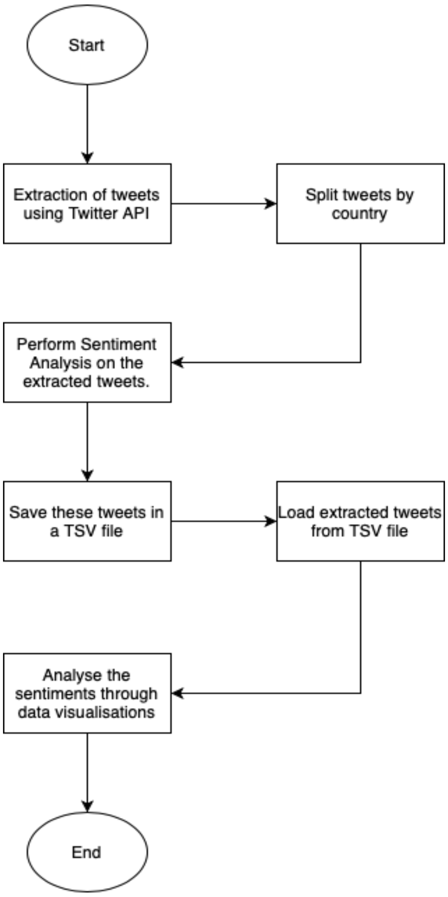

Start
Extraction of tweets using Twitter API
Split tweets by country
Perform Sentiment Analysis on the extracted tweets.
Save these tweets in a TSV file
Load extracted tweets from TSV file
Analyse the sentiments through data visualisations
End

Figure 1: Flowchart Showcasing Workflow Overview of Project

A flowchart is a graphical representation of the workflow or process of a system in a diagrammatic way.

Figure 1 illustrated above highlights the workflow of this project. First, it was required to extract tweets using twitter API which are related to covid-19 vaccines. These were then split by countries, Great Britain, Italy, and Spain respectively. Afterwards, sentiment analysis tool was used to determine the sentiment of each tweet and were then saved as a TSV file. Finally, these TSV file were loaded into another script of code to perform analysis on these results.

Moreover, a google drive link for the code implementation as well as all unfiltered and filtered tweets can be found here:

https://drive.google.com/drive/folders/1tzdrOhlWNYNhF07Il2ps0jnws4Jzuteu?usp=sharing

Background Research

Prior to development, it was imperative to conduct research on the tools which will be exploited throughout the whole project.

The first thing which was done was to research about Twitter API, and how it can be used. To get access to Twitter's API [1], one needs to initially have an account. Then, it was required to apply for a developer's account to make a request to the API. Once approved, a standard project was created with an associated developer app. This will provide me with a set of credentials that will allow me to authenticate with the Twitter API. The credentials were stored inside a JSON file to avoid regenerating them each time.

After having access to the API, it was time to research on tools and libraries that will permit me to use Twitter's endpoints to extract tweets from Twitter. Tweepy [2] is the library which I have used throughout my whole project to search and retrieve tweets. This library is simple to use that enables users to authenticate with Twitter's API using generated credentials as well as to connect with their endpoints, such as the search endpoint.

Additionally, data was collected into two distinct ways; from an existing corpus of tweets and searching for tweets using Twitter API. The standard Twitter API application only allows users to search for tweets that were posted in the previous seven days. Therefore, I had to research other ways of collecting tweets prior to the blood clot fears news, and hence, I used an existing corpus containing tweets related to covid-19 [3]. This dataset contains a dataset of tweets organised into folders, where each folder corresponds to a date the tweets were posted. Each folder had a filtered version of the dataset organised into a TSV file, containing Tweet ID, date, time, language and country code. Therefore, using this dataset, I could collect tweets related to the covid-19 vaccines before and after the blood-clot news emerged (18th March 2021).

There are many libraries to perform sentiment analysis using Python. The most widely known and used is NLTK's in-built Sentiment Analyser [4]. This library takes in a sentence and determines the polarity sentiment of that tweet, ranging from -1 to 1, i.e., negative to positive respectively. Finally, sentences were denoted by either positive, neutral, or negative according to their polarity score.

Sentiment analysis is a widely known natural language processing tool used both for academic purposes as well as industry purposes. There are many projects out there that utilises the NLP tool in order to understand people's emotions and thoughts on a particular subject. For instance, one can perform sentiment analysis on the stock market to grasp investor's opinions on a specific stock since sentiment sometimes may correlate with future stock price. A similar project [5] to mine was carried out by another individual, however, in this project sentiments were not calculated in terms of countries and dates were also not important. On the other hand, the project which I have conducted considers tweets sentiments of different countries at different dates to assess people's emotions after the blood clot news was concluded.

Lastly, visualising the data collected is essential for a good comparative study. Hence, for this project, it was decided to go with bar charts to compare sentiments for each country. Percentages were used as the values for the bar chart for a fairer study. This is because some countries had more collected tweets than other, therefore, using percentages would normalise values. Then, a pie chart was implemented to have an overview of sentiments combining all countries and using word cloud to output the most frequent words used inside tweets. The former was used to have a general overview of sentiments in these countries combined whilst the latter was used to identify what were the reasons behind any shifting of sentiment pre and post thrombosis news (further discussed later).

Implementation and Testing

Jupyter Notebooks were the chosen platform using Python 3 for the development of both the extraction and sentiment analysis of tweets as well as visualisation analysis. The final version for this project includes two Jupyter notebook files, one that performs extraction of tweets from the GitHub repository, searches for tweets using Twitter API and performs sentiment analysis using NLTK library. On the other hand, the other Jupyter notebook is used to read all saved TSV files (contains tweet's text, polarity score and sentiment) and process them in such a way to generate data visualisation which will be analysed accordingly.

Extracting, Parsing and Performing Sentiment Analysis

Using GitHub Repository

With regards to the GitHub repository containing covid-19 tweets, the following approach and decisions were taken.

Prior to development of code, a dataset containing these tweets was created. This dataset created is split into two parts, pre and post blood clot news. As previously mentioned, this GitHub dataset contains folders, where each folder corresponds to a date the tweets were posted. Hence, I could split the dataset into two distinctive parts since I knew the dates they were posted. Therefore, I set a cut-off date (18th March 2021) when the EMA concluded that blood-clots were a rare side effect of a particular vaccine [6].

Moreover, since Tweepy library will be used to get tweets by their id, it was first required to authenticate with the Twitter API via generated credentials. Then, I implemented an algorithm which reads through each file in the current dataset folder path and creates three separate data frames for each country. The next step would be that of getting those tweets which are relevant to covid-19 vaccines by first getting the tweet's text for each tweet id using Tweepy's 'get_status()' method.

```
tweet_text = api.get_status(tweet).text
```

A list of predefined keywords is initialised and if the tweet text contains any of these keywords, then it is relevant. This tweet would then be parsed by removing hyperlinks, '@' symbols, emojis and '\n'. Eventually, the algorithm will create another three-separate filtered data frame for each country.

Using Twitter API Search Endpoint

With regards to searching for tweets using Twitter's API, the following approach and decisions were taken.

Firstly, a dictionary of locations was created where each key corresponds to the location whilst each value corresponds to the geographic location of that location (obtained from Google Maps). Moreover, a set of keywords was created that will be used as a search query.

Then, the algorithm will search through each defined location, and for each location, it loops through every search keyword. Now, for every search keyword, the algorithm uses Twitter's search endpoint to get all tweets (max of 200 tweets for every search request) that matches the specified search keyword query. It is important to note that only tweets whose language is English were collected.

```
tweets=tweepy.Cursor(api.search,q=keyword,lang=language,geocode=value+",100
km").items(200)
```

Then, the algorithm creates a data frame containing all relevant filtered tweets for each country.

Sentiment Analysis

After having successfully collected relevant tweets which are filtered, it is time to perform sentiment analysis using NLTK library.

The first step is to initialise a copy of the Sentiment Analyser class.

```
sia = SentimentIntensityAnalyzer()
```

Then, a lambda function was implemented that takes in a sentence and calculate the polarity sentiment of that sentence.

```
calcPolarity = lambda x:sia.polarity_scores(x)['compound']
```

Hence, an additional column was initialised in each respective data frame for the polarity score of each tweet. Finally, another function was implemented that determines the sentiment (positive, neutral, negative) of each tweet based on every polarity score and subsequently, another column was created in each data frame that outlines the sentiment type for each polarity score previously generated.

```
def computeAnalysis(polarity):

        if polarity > 0:
                return "Positive"
        elif polarity == 0:
                return "Neutral"
        else:
                return "Negative"
```

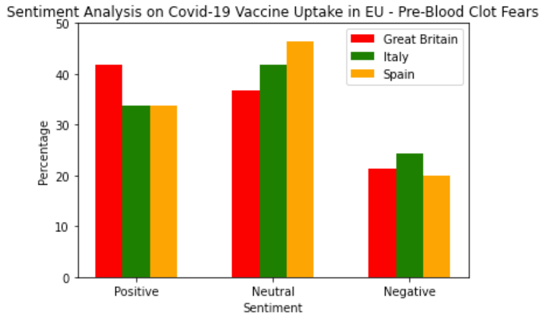
These data frames were then stored and saved in a TSV file and put in their respective folder.

Data Visualisation Analysis Graphs

All data has currently been collected and processed. Now, it is the time to visualise this data using appropriate graphs to analyse them. For this project, I have decided to go with three distinct visualisation graphs: a bar chart, pie chart and a word cloud.

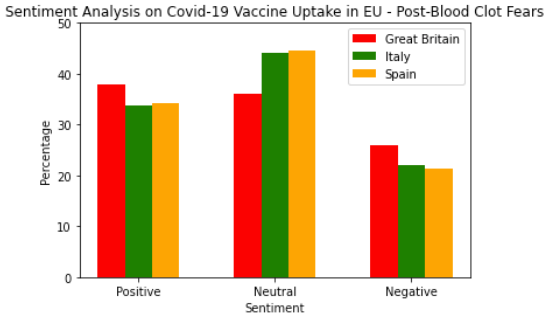
The first thing this jupyter notebook does is to read all saved TSV files from their respective folder and concatenate them into a single data frame for each country and for each time span. As previously mentioned, I will be analysing whether the thrombosis news had affected people's sentiments on the vaccines in general by country, and therefore, it is imperative to create a data frame that holds all tweets and sentiments for each country in the appropriate time span.

Now it is time to visualise the data using graphs. To create a bar chart, the Matplotlib library [7] was used. A bar chart is a chart that represent categorical data with bars proportional to the values they represent. For this visualisation method, I decided to normalise the frequency of each sentiment in a percentage to make it more accurate since tweets were not collected equally for all countries. Figure 2 and 3 below illustrates the bar chart for each country's sentiments pre and post blood clot news together with a table outlining each percentage value.



```
           Great Britain   Italy   Spain
Positive          41.74%  33.77%  33.67%
Neutral           36.82%  41.82%  46.48%
Negative          21.44%  24.42%  19.85%
```

Figure 2: Bar chart for Pre-Blood Clot News



```
           Great Britain   Italy   Spain
Positive           38.0%  33.83%  34.28%
Neutral           36.02%  44.14%  44.45%
Negative          25.98%  22.03%  21.27%
```

Figure 3: Bar chart for Post-Blood Clot News

Moreover, a pie chart and a word cloud were also created to further analyse covid-19 vaccine uptake in these countries. However, this time it was decided that countries were not of importance, hence, countries were combined to provide an overview of sentiments.

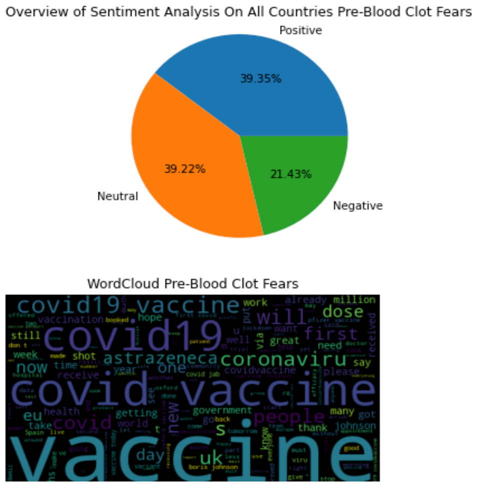


Figure 4: Pie Chart and Word Cloud pre-blood clot news

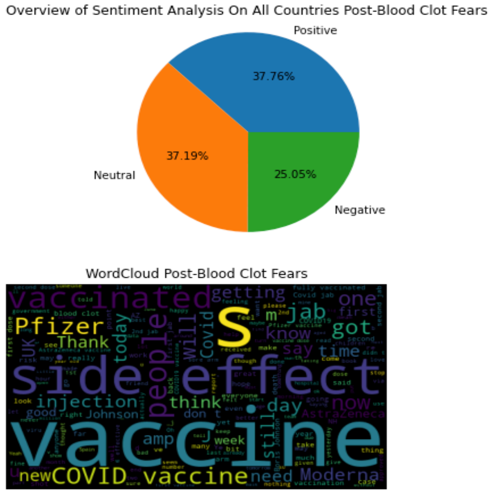
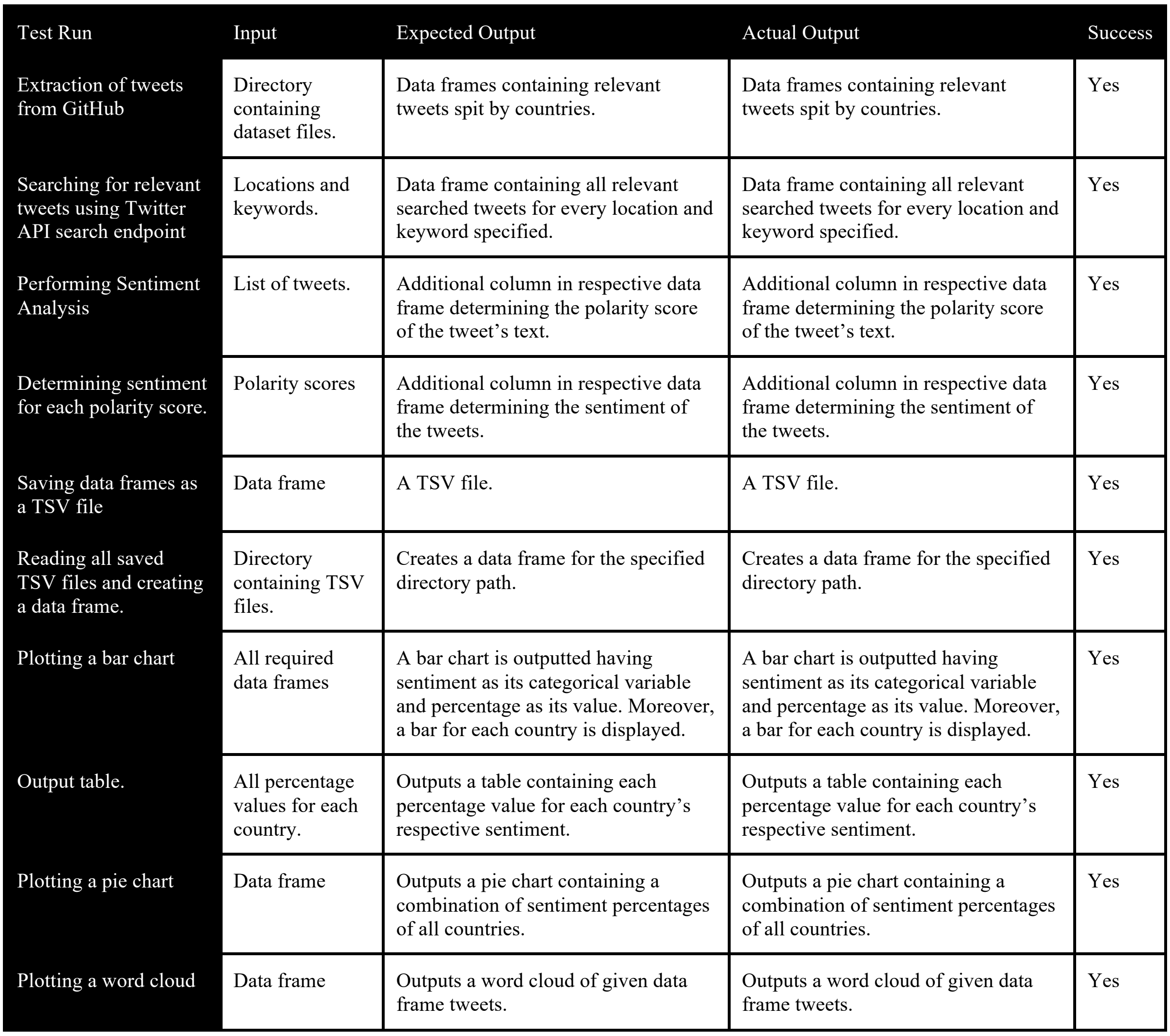


Figure 5: Pie Chart and Word Cloud post-blood clot news

Testing

The testing technique used for this project was White Box Testing, mainly focusing on the implemented algorithms highlighted in figure 6 below. Observably, the testing performed demonstrates that the project code works as intended.



Test Run	Input	Expected Output	Actual Output	Success
Extraction of tweets from GitHub	Directory containing dataset files.	Data frames containing relevant tweets spit by countries.	Data frames containing relevant tweets spit by countries.	Yes
Searching for relevant tweets using Twitter API search endpoint	Locations and keywords.	Data frame containing all relevant searched tweets for every location and keyword specified.	Data frame containing all relevant searched tweets for every location and keyword specified.	Yes
Performing Sentiment Analysis	List of tweets.	Additional column in respective data frame determining the polarity score of the tweet's text.	Additional column in respective data frame determining the polarity score of the tweet's text.	Yes
Determining sentiment for each polarity score.	Polarity scores	Additional column in respective data frame determining the sentiment of the tweets.	Additional column in respective data frame determining the sentiment of the tweets.	Yes
Saving data frames as a TSV file	Data frame	A TSV file.	A TSV file.	Yes
Reading all saved TSV files and creating a data frame.	Directory containing TSV files.	Creates a data frame for the specified directory path.	Creates a data frame for the specified directory path.	Yes
Plotting a bar chart	All required data frames	A bar chart is outputted having sentiment as its categorical variable and percentage as its value. Moreover, a bar for each country is displayed.	A bar chart is outputted having sentiment as its categorical variable and percentage as its value. Moreover, a bar for each country is displayed.	Yes
Output table.	All percentage values for each country.	Outputs a table containing each percentage value for each country's respective sentiment.	Outputs a table containing each percentage value for each country's respective sentiment.	Yes
Plotting a pie chart	Data frame	Outputs a pie chart containing a combination of sentiment percentages of all countries.	Outputs a pie chart containing a combination of sentiment percentages of all countries.	Yes
Plotting a word cloud	Data frame	Outputs a word cloud of given data frame tweets.	Outputs a word cloud of given data frame tweets.	Yes

Figure 6: Test cases for every algorithm from extraction of tweets to data visualisations.

From the testing performed, it appears that the project is working as intended. The above white-box testing allowed me to test whether every algorithm functionality in my scripts is working as should be. Moreover, it can be deduced that this project can be further used to scale up the number of tweets collected and perform sentiment analysis on them in the future.

Evaluation and Critical Analysis

This project should serve as a useful study to gain an overview of the wider public opinion behind the covid-19 jabs.

The resultant graph after performing sentiment analysis can be evaluated to gain insight on how the blood clot news might have affected people's emotions. Moreover, countries will be compared with each other for a more comparative study.

First of all, a total of 50051 filtered tweets were collected for this project. Figures 2 and 3 above illustrate the sentiment analysis graph pre and post thrombosis news. From figure 2, it can be noticed that overall, every country had a positive sentiment towards covid-19 vaccines when blood clots were not an issue. On the other hand, in figure 3, a change in sentiments can be observed. For instance, Great Britain's positive sentiment dropped from 41.74% to 38.0% (-3.74%) and their negative sentiment increased from 21.44% to 25.98% (+4.54%). Moreover, Spain's negative sentiment also increased by 1.42% after blood clots were concluded as a rare side effect of a particular vaccine. On the other hand, Italy's sentiments slightly changed to the positive way, but very minimal.

Furthermore, figures 4 and 5 respectively continue highlighting that the overall people's opinions and emotions on the covid-19 vaccine uptake shifted to a negative way when the thrombosis news emerged. Before the thrombosis news, there were 39.35% and 21.43% of positive and negative tweets respectively. After the thrombosis news, positive sentiments dropped to 37.76% (-1.59%) whilst negative sentiments increased to 25.05% (+3.62%). To try and identify which phrases of words changed this sentiment, a word cloud was outputted together with the above pie chart. The word cloud illustrated in figure 4 shows positive words like "great", "thank" and "work", amongst others whilst the word cloud in figure 5 shows more negative words like "side effect", "blood clot" and "death", amongst others. However, it is important to note that the majority of sentiments in both timeframes is positive, hence, overall people have good feeling towards vaccine uptakes, and this can be seen in all the above figures, including both word clouds.

Strengths and Weaknesses

A significant advantage in this project is the diversity of data visualisations being used. This was done to fully understand people's sentiments and what keywords are being used to express themselves. This will lead to a better a sentiment analysis because this study can better understand the most frequent words being used within tweets. Moreover, another advantage to this study is that different countries were compared with each other, thus evaluating sentiments across different locations.

On the other hand, this study considers all the vaccines together. Therefore, doing the same research study but this time separating the AstraZeneca vaccine (this vaccine only had thrombosis reports) with the other vaccines can better understand the true sentiment feelings of people.

Improvements and Future Work

Potential improvements to the current project include that of collecting more tweets and ensuring that every tweet collected is relevant to the covid-19 vaccine, thus removing any noise data.

With continued research and development, this project could be further expanded in such a way to use other sentiment analysis libraries such as Text Blob [8] and Stanza [9]. Then, a mean polarity score would be taken across these libraries and determining the sentiment of this resultant. This is because different sentiment analysis tools produce different polarity scores for the same sentence (trained differently), therefore, taking the mean polarity score would be the most ideal. Hence, sentiments for every tweet will be determined better.

Moreover, another data visualisation graph can be added. For instance, plotting a line graph to identify trends in sentiments across different dates and perhaps analysing how trends change in key dates (ex: when a vaccine gets approval for administration purposes). There are many things one can do on this project to further analyse people's perceptions on covid-19 vaccines.

Conclusion

The primary objective for this project was to analyse whether overall vaccine sentiment shifted when the blood clot issue was concluded as a rare side effect for vaccines. Thousands of tweets from different locations were collected to perform sentiment analysis on. These were done analysed through several data visualisations graphs outputted.

Based on the results, it can be concluded that the majority of individuals have more positive sentiments rather than negative sentiments. However, sentiments shifted to the negative side when the EMA concluded that blood clots were a rare side effect of a particular vaccine. These evaluations were concluded from the statistics and charts outputted as well as the word cloud, which the latter shows the most frequent words from the extracted tweets.

Sentiment analysis can help governments of the respective countries make informative decisions regarding the vaccine rollout plan. However, it is imperative to continue analysing the publics thoughts on vaccines continuously to get instinct information from the public as time passes.

References

[1] Twitter, “Twitter API,” [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api.

[2] J. Roesslein, “Tweept Documentation,” 2009. [Online]. Available: https://docs.tweepy.org/en/latest/.

[3] J. M. Banda, “Covid19 Twitter,” [Online]. Available: https://github.com/thepanacealab/covid19_twitter.

[4] NLTK, “Sentiment Analysis,” [Online]. Available: http://www.nltk.org/howto/sentiment.html.

[5] K. Leung, “COVID-19 Vaccine — What’s the Public Sentiment?,” [Online]. Available: https://towardsdatascience.com/covid-19-vaccine-whats-the-public-sentiment-7149c9b42b99.

[6] E. M. Agency, “COVID-19 Vaccine AstraZeneca,” 18 March 2021. [Online]. Available: https://www.ema.europa.eu/en/news/covid-19-vaccine-astrazeneca-benefits-still-outweigh-risks-despite-possible-link-rare-blood-clots.

[7] Matplotlib, “Matplotlib: Visualization with Python,” [Online]. Available: https://matplotlib.org/stable/index.html.

[8] TextBlob, “TextBlob: Simplified Text Processing,” [Online]. Available: https://textblob.readthedocs.io/en/dev/.

[9] Stanza, “Sentiment Analysis,” [Online]. Available: https://stanfordnlp.github.io/stanza/sentiment.html.

Appendix

Appendix A: Diagrams

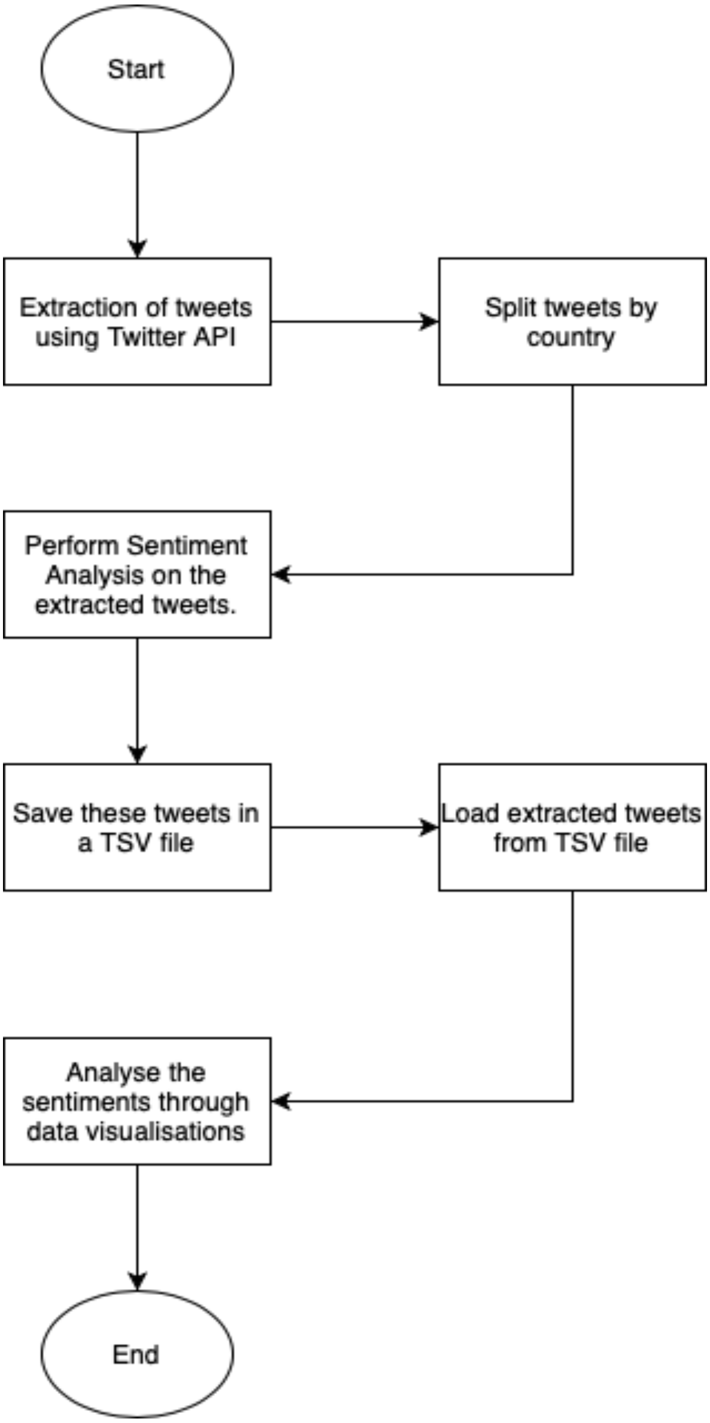
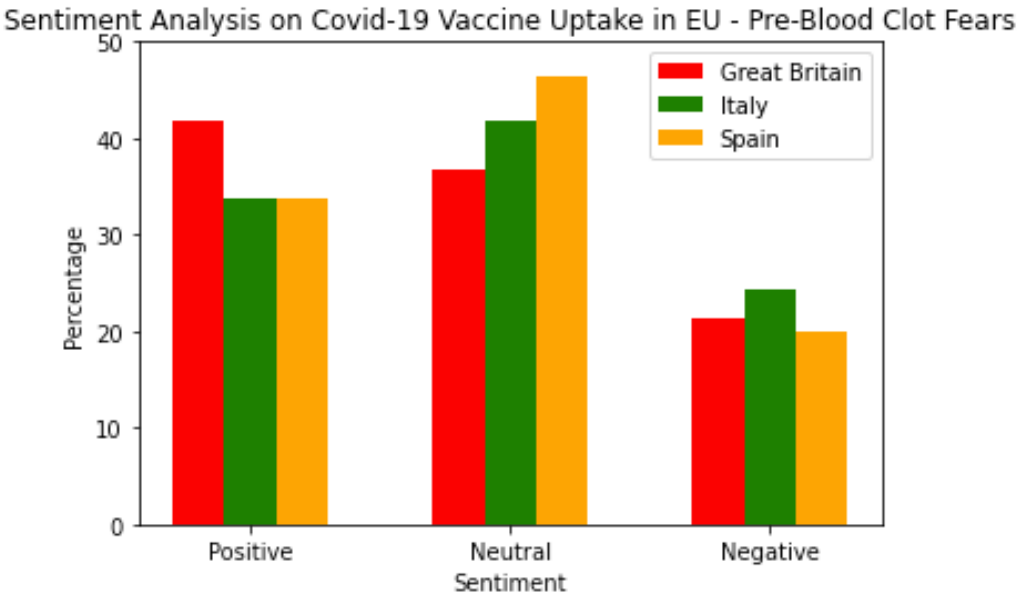
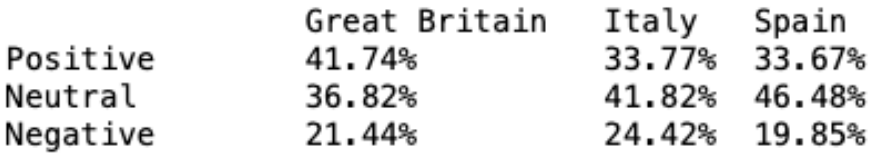

Start
Extraction of tweets using Twitter API
Split tweets by country
Perform Sentiment Analysis on the extracted tweets.
Save these tweets in a TSV file
Load extracted tweets from TSV file
Analyse the sentiments through data visualisations
End

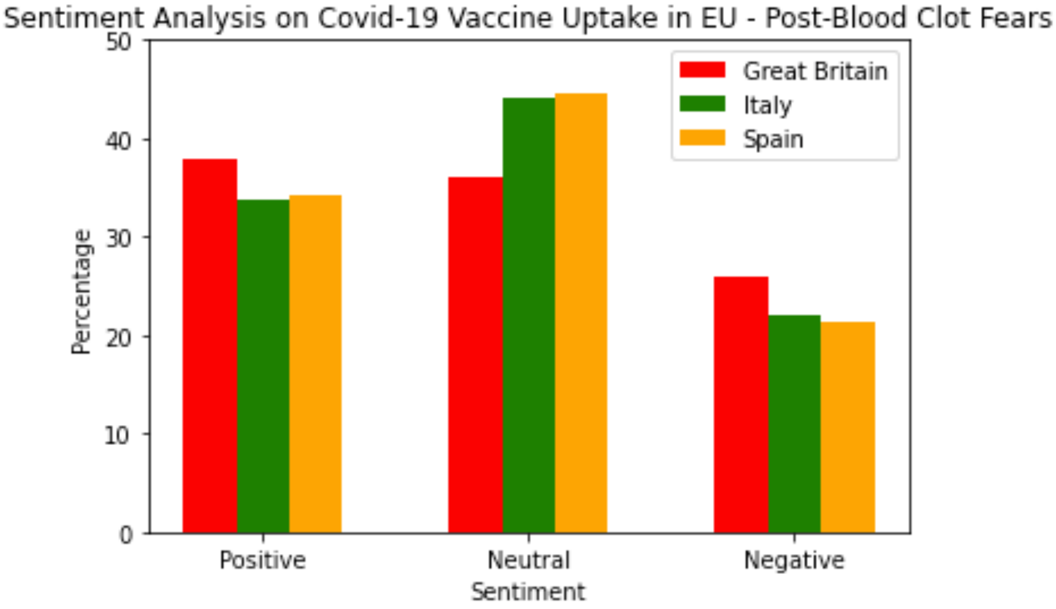
Figure 1

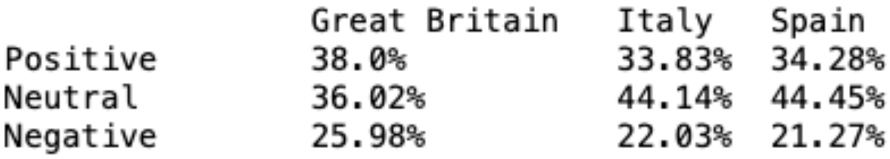




```
              Great Britain    Italy   Spain
Positive      41.74%           33.77%  33.67%
Neutral       36.82%           41.82%  46.48%
Negative      21.44%           24.42%  19.85%
```

Figure 2





```
              Great Britain    Italy   Spain
Positive      38.0%            33.83%  34.28%
Neutral       36.02%           44.14%  44.45%
Negative      25.98%           22.03%  21.27%
```

Figure 3

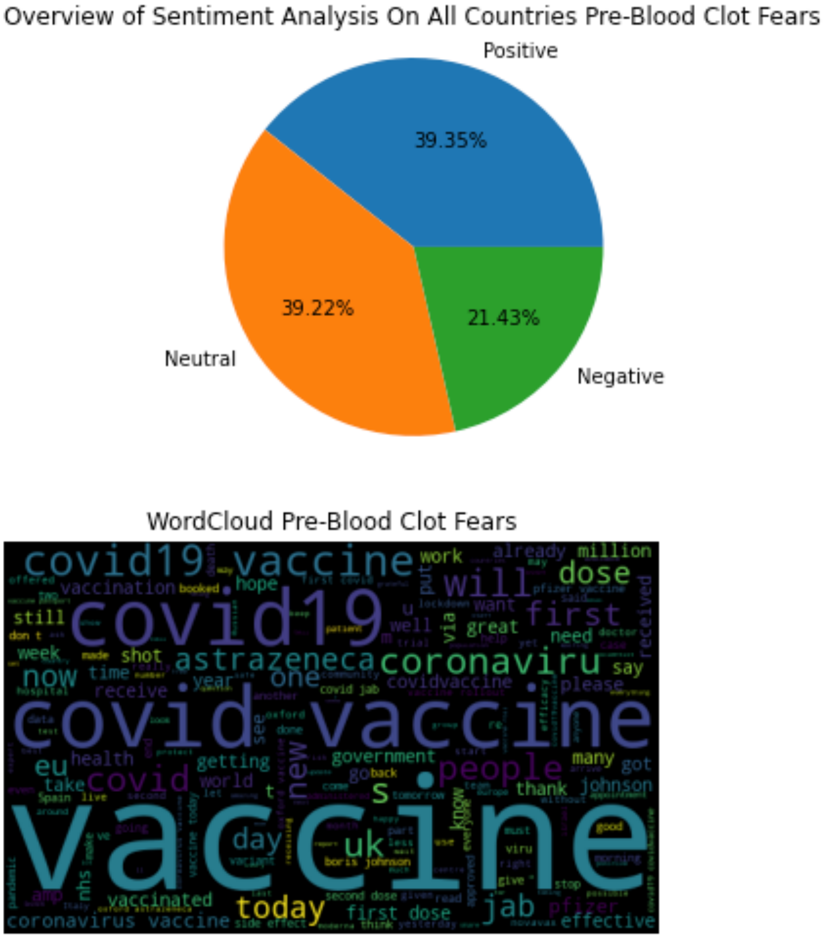


Figure 4

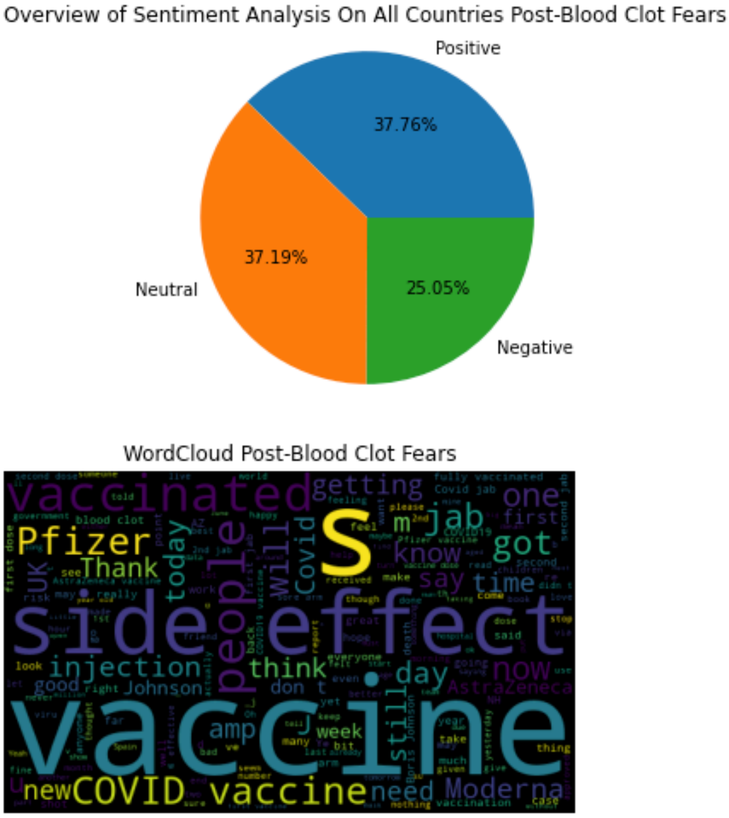


Figure 5