C1730333

Final Report

Cardiff University School of Computer Science and Informatics

ANALYSING COVID-19 SENTIMENT ON TWITTER IN THE UNITED KINGDOM

CM3203 - One Semester Individual Project

Author: Akansha Garg

Supervisor: Steven Schockaert

Contents

Contents 2
Table of Figures 3
Abstract 4
Acknowledgements 5
Introduction 6
Background 7
Twitter API 7
Tweepy 7
Sentiment Analysis 8
NRCLex 8
Textblob 9
Natural Language Toolkit 9
Data Analysis 10
Power BI 10
Pointwise Mutual Information 10
Related Work 11
Project Aim 12
Approach 13
Creating a Tweet Database 13
Selecting keywords 13
Tweepy 13
Sentiment Analysis 15
Data Analysis 17
Creating a word dictionary 17
Calculating Pointwise Mutual Information 18
Results 19
Summary of Results 29
Evaluation of Results 29
Evaluation of collecting tweets. 29
Evaluation of sentiment analysis tools 29
Evaluation of analysis of results 30
Conclusions 31
Future work 32

Reflections on learning 32
References 34

Table of Figures

Figure 0: a code snippet showing the method of running a Twitter Stream. 14
Figure 1: a code snippet to show the if else statement to extract the tweet text. 14
Figure 2: a code snippet of the checks made to ensure the tweet streamed matches required criteria. 15
Figure 3: a snippet of the excel file created after running the NRCLex sentiment tool. 15
Figure 4: a code snippet of the code to analyse sentiment and appending it to a dictionary.. 16
Figure 5: a code snippet of the method to analyse the tweets with the Textblob tool. 16
Figure 6: a code snippet of how the word dictionary was made. 17
Figure 7: A code snippet showing the method of calculating PMI scores. 18
Figure 8: A column chart to show the number of tweets collected per month 19
Figure 9: A pie chart to show the distribution of positive, negative, and neutral sentiment... 20
Figure 10: A Line Chart to Show Average Polarity of Joy, Positive, Anticipation, Trust and Surprise Sentiment Over Time 20
Figure 11: A heatmap showing average sentiment across the country. 21
Figure 12: A Line Chart to Show Average Polarity of Negative, Fear, Anger, Disgust and Sadness Sentiment Over Time 21
Figure 13: A bar chart to show the number of tweets by sentiment in February. 22
Figure 14: A bar chart to show the number of tweets by sentiment in March. 23
Figure 15: A bar chart to show the number of tweets by sentiment in April. 23
Figure 16: A bar chart to show the number of tweets by sentiment in May. 24
Figure 17: Word Cloud to show the most common words used in Negative Tweets. 244
Figure 18: Word Cloud to show the most common words used in Positive Tweets: 25
Figure 19: A bar chart to show the number of tweets tweeted by sentiment in Northern Ireland 25
Figure 20 : A bar chart to show the number of tweets tweeted by sentiment in Wales..... 266
Figure 21: A bar chart to show the number of tweets tweeted by sentiment in England..... 26
Figure 22 : A bar chart to show the number of tweets tweeted by sentiment in Scotland. ... 277
Figure 23: A table to show the top 20 words with the highest PMI value for Positive and Negative sentiment.. 278

Abstract

For over a year and half the world has been dealing with the Covid-19 pandemic, which has been described as the most challenging crisis humanity has had to face since World War II (bbc.co.uk, 2020). The pandemic has affected everyone and everything; the impacts have been long lasting and will change the way people will live their lives in the future. With new information being released almost every day in the United Kingdom regarding the rules and regulations, people's sentiment towards the pandemic is constantly fluctuating.

Twitter is a social networking site that was released in 2006 and has grown in popularity and is now one of the most popular. 'Twitter is what's happening and what people are talking about right now' is how the microblogging site describes itself (twitter.com, n.d.). With over 192 million daily users, around 6,000 tweets are posted every second (Lin, 2021). It is a free resource for the public to use to express their sentiment about any situation instantly. The social media platform allows its users to post a short message of up to 280 characters in the form of a 'tweet'. This project aims to analyse the sentiment that people in the United Kingdom have about Covid-19 using Twitter.

Acknowledgements

I would like to thank my supervisor Steven Schockaert for his support and guidance throughout this project. I would also like to thank my friends and family for their encouragement and motivation that has helped me complete this project during such unprecedented times.

The pandemic has been a difficult time for everyone and their families, but I would like to add a special thank you to the frontline workers that have continued to work tirelessly to help people and the country during this time.

Introduction

The Coronavirus pandemic, also referred to as the COVID-19 pandemic is an on-going global crisis. With over 161 million cases and 3.35 million deaths registered worldwide to date in May 2021 (JHU CSSE COVID-19 Data, 2019) this has been one of the deadliest pandemic's humanity has had to ever face.

The first case of the virus was discovered in December 2019 in Wuhan, China. The World Health Organisation (WHO) officially declared a pandemic on 11th March 2020 (WHO, 2019). Due to the nature of the virus, it spreads extremely quickly and easily through social contact and to prevent the spread of the virus many countries enforced a strict lockdown. The United Kingdom announced its first and official lockdown on 23rd March 2020 (Gov.uk, 2020). These new restrictions prevented people from meeting in person and greatly impacted the world's economy.

Along with a national lockdown many other regulations came into place that were unprecedented and new to people such as the mask mandate and social distancing. It became compulsory to wear a face covering in indoor settings such as shops, supermarkets, banks etc in April 2020 in the United Kingdom (GOV.UK, 2021). Social distancing was enforced in any public setting, which involved always maintaining a distance of at least 6 feet from people. The national lockdown also has a huge impact on businesses, since non-essential businesses and venues had to be closed. These restrictions and measures were put in place to reduce the spread of the infection and protect lives. However, people still had an opinion and one of the main outlooks to express them during the pandemic become through social media due to the 'stay at home' regulation.

Social media is no longer a new or unknown concept to people, it has grown rapidly and has a very large user consumption now. Social media has been used to track and analyse disease outbreaks in the past and Twitter is a very popular social media platform used to track analyse the data. Twitter is a social media microblogging site which allows users to express their thoughts and opinions and interact with people online through the format of a short message, referred to as a 'tweet'. The immense amount of usage and information that can be found on it is what makes it such an attractive tool to use for analysis. In fact, studies have found that through the tracking of tweet streams, flu outbreaks have been predicted 1-2 weeks ahead of the CDC's (Centres for Disease Control and Prevention) surveillance average, which is the national public health agency of the United States. The studies also found that were able to also track user's concerns during the Swine Flu pandemic in 2009 by looking at most commons terms and words found in the Twitter streams (Schmidt, 2012).

Over the last 18 months, the United Kingdom has gone through a multitude of regulations as mentioned and during the timeframe of this project the country went through its 3rd official lockdown, easing of restrictions including the increase of social contact, reopening of restaurants, a fast and steady vaccination role out and more (analysis, 2021) . All these changes will have had a significant impact on people and their sentiment regarding the coronavirus pandemic and this project aims on analysing that sentiment by utilising Twitter.

During the course of this project, it will be interesting to see the sentiment people have in the United Kingdom regarding covid-19 at this stage in the pandemic. Where being in lockdown and practicing social distancing is no longer new for people, the country has a steady vaccination roll out and a roadmap to the return of a 'normal' country has been presented by the government. From the analysis of the dataset, I am hoping to find out if the number of positive tweets and negative tweets will increase/decrease around certain milestone dates in the United Kingdom Covid Roadmap. For example: The number of positive tweets should increase around 12 April 2021 since non-essential retail shops, hairdressers, gyms etc. are scheduled to open (gov.uk, 2021).The analysis by location should help to uncover if the sentiment across Wales, England and Scotland will vary due to each country having different roadmaps. Through this study I aim to get a better understating of the public perception of Covid-19 in the United Kingdom and discover interesting and significant trends around the pandemic.

Background

In this section I will introduce and cover the most important concepts and software tools that I will use in this project and explain the function they provide.

Twitter API

The Twitter API lets you read and write Twitter data. It can be used to write tweets, get access to your Twitter timeline and in the case of this project it will be used to stream data on Twitter about Covid-19. To get access to the Twitter API a 'Twitter Developer' account must be created. After making an account unique authentication credentials are generated, these credentials enable you to stream data from Twitter. There are 4 unique credentials generated: 'API key', 'API secret key', 'Access Token', 'Access Token Secret', with these credentials you are able to utilise the various Twitter streaming APIs available.

Tweepy

There are a number of different Twitter Streaming API's available for different programming languages. For my project I chose 'Tweepy' which is a python library used to access the Twitter API. I choose Tweepy as python is a programming language I am comfortable and well versed in using.

Tweepy uses your unique Twitter API credentials to access and stream Twitter. The tweets streamed are returned in a JSON format 'JavaScript Object Notation'. Each tweet JSON has over 150 attributes associated with it for example the actual text of the tweet, the number of favourites it has, the date it was created at, the user that tweeted it and so on. For my project the following attributes are the most important:

- *'text':* contains the actual text of the tweet.
- *'extended_tweet':* if tweet is over 140 character, contains the actual text of those tweets.
- *'retweet_status'*: checks if the tweet is a retweet.
- *'place'*: if the tweet has been geo tagged the text from the tag.
- *'created_at':* the time and date of when the tweet was tweeted.

Sentiment Analysis

Sentiment analysis is the process of using natural language processing (NLP) and machine learning to determine the emotions and attitude of a piece a text, usually by assigning a polarity score that will state if the text is positive, negative, or neutral in sentiment. (lexalytics.com, n.d.)

Sentiment analysis has multiple applications for a broad range of industries and can provide invaluable information to organisations. One use case of sentiment analysis includes analysing customer feedback. A company can receive a large amount of feedback from multiple sources including surveys, online comments and reviews. Being able to run sentiment analysis on that data can allow companies to gauge the sentiment around certain products and improve in areas where there is a large amount of negative sentiment. Sentiment analysis on customer feedback allows for large amounts data to be processed and for companies to keep up to date and real time knowledge of their customer's sentiment.

Another use case includes using sentiment analysis to monitor a brand reputation on social media. Social media allows customers to post their true opinion on a company and its brand. Through sentiment analysis tools, all social media posts about a company on different platforms can be analysed and overall sentiment and real time public opinions about a company from customers can be revealed. This analysis can also be used to predict customer opinions and reactions which can aide in creating new products and services. (MonkeyLearn, 2020)

There are many different methods and approaches that can be taken when analysing sentiment. The approach I will be using for my project is a lexicon-based approach, this approach uses a pre-defined dictionary of words that have a sentiment polarity assigned to them. The algorithm breaks down its input text into words to match a polarity value to each word in the input, the sum of the polarity values for the whole sentence is the sentiment assigned to the input. The reason for choosing this approach is because the text input I will be using are tweets, which mostly contain common words that should be present in a lexicon sentiment analysis dictionary and therefore should be able to provide an accurate sentiment analysis of the tweets. Another common approach to sentiment analysis includes machine learning. This method first trains an algorithm on a training dataset by presenting it with inputs and their expected outputs before applying the algorithm to the actual dataset. The machine learning algorithm improves with frequent uses and large datasets. This was one of the reasons I did not choose this approach as my project was on a short timeframe and finding a training dataset and collecting my real tweet dataset would have been difficult in the time frame. (Devika M D, 2016)

NRCLex

One of the sentiment analysis tools available and one of the tools I will be using to analyse the tweets dataset is NRCLex. NRCLex is described as an affect generator, this refers to the fact that not only does it find positive and negative sentiment but also is able to associate text to 8 different emotions. It is based on the NLTK library's WordNet synonyms sets and the National Research Council Canada (NRC) affect lexicon. It is able to measure the emotional affect from

a body of text and assign it a score between 0 and 1. It was created by Mark M. Bailey as a PyPi Project and is MIT-approved (Anon., 2020). It has 27,000 words approximately that can help it assign a score, for 2 sentiments and 8 different emotions:

1. fear
2. anger
3. anticipation
4. trust
5. surprise
6. positive
7. negative
8. sadness
9. disgust
10. joy

NRCLex is a lexicon-based sentiment analysis tool. This means that there is a dictionary of words with a preassigned sentiment score. When a text is processed it is tokenized, which means it is broken down into individual word components referred to as tokens. Each token is matched with an available word in the dictionary to find its sentiment and score. After each token has been an analysed the scores are processed through an algorithm that outputs the relevant score for the different sentiments and emotions for the text. (Roul, 2021)

Textblob

Textblob is the second sentiment analysis tool that I will use to analyse the tweets dataset. Textblob is a Python Library used for processing textual data. It can do a host of different processes on text including part of speech tagging, noun phrase extraction, spelling corrections and Sentiment analysis (textblob.readthedocs.io, n.d.). When running sentiment analysis on the text, Textblob uses natural language processing to assign the text with a sentiment. The sentiment is divided into two scores, a polarity score and a subjectivity score. The polarity score is between 1 and -1 where a score between 1 and 0.1 is positive, a score between -0.1 and -1 and score of 0 is neutral. The subjectivity score is given between 0 and 1 where 0 is very objective and 1 is very subjective.

Textblob is also a lexicon-based sentiment analysis tool. Similar to NRCLex, Textblob also has its own dictionary of words that have been assigned a subjectivity score and a polarity score. Once the text has been tokenized, each token is given both polarity and subjectivity score. It then finds an average of each score and returns it to provide an overall score for the text. (planspace.org, 2015)

Natural Language Toolkit

The Natural Language Toolkit or NLTK is a python library which contains a suite of text processing libraries for classification, tokenization, stemming and more. (NLTK.org, n.d.). I will be using multiple libraries from NLTK. One of the tools I will use is the NLTK tokenizer which breaks down a string into substrings, in the case of my project I will be breaking down the tweets into individual words.

I will also make use of the NLTK Lemmatizer. Lemmatization reduces the transformed words into its root word referred to as the 'Lemma'. A Lemma is the dictionary form of a set of words. For example, 'runs', 'running', 'ran' are all forms of the word run, therefore making run the lemma of all the words (datacamp.com, 2018). Running a lemmatization tool of text will return the actual word used. For my project I will first tokenize all the words in a tweet and then lemmatize them to be able to count the true number of times a word appears and not have to worry about duplicates that could be found if the tweets had not been lemmatized.

Another library available in NLTK includes stop words, stop words are English words that do not add important meaning to a sentence and can be ignored and removed without risking losing the meaning of the sentence. Stop words are commonly used words such as 'the', 'a' 'in' etc. (GeeksforGeeks, 2020). For my project I will be making a use of the NLTK library throughout. For example, when analysing the tweets for Pointwise Mutual Information (PMI) and Word Clouds, I will first tokenize the tweets, then lemmatize each word and remove stop words.

Data Analysis

Power BI

Power BI is a business intelligence tool made by Microsoft, which allows for data sources to be connected and transformed and then allows you to visualize and analyse your data to gain relevant insights. (docs.microsoft.com, 2021). I will be using PowerBI in my project to create visualisations such as Heatmaps, Word Cloud, Bar charts etc using the Tweets Dataset.

Pointwise Mutual Information

Pointwise Mutual Information (PMI) is a measure of association Using PMI you are able to quantify the likelihood of co-occurrence of two properties. For my project I will be calculating a PMI score for the likelihood of a word in a positive tweet and the likelihood of a word in a negative tweet. The formula computes the (log) probability of the co-occurrence of the two events scaled by the product of the individual probability of the occurrence of each independent event. The formula is as follows:

$$PMI\ (a,b) = \log(\frac{P(a,b)}{P(a)P(b)})$$

when 'a' and 'b' are independent and do not form a unique concept, the joint probability will be equal to the product of their independent probabilities I.e., when the ratio will equal 1 and hence the log of the ratio will be 0. In my project I will use the PMI formula to investigate the association between words and tweet sentiments. For a word to form a unique concept with either a positive or negative tweet the joint probability of the word and tweet sentiment must be high, resulting in a log of a value higher than 0 (Alto, 2020). Which would mean a word would be found more often than expected in a positive or negative tweet, suggesting a possible association.

Related Work

Sentiment analysis on social media and in particular on Twitter is not a novel idea and has been researched and written about extensively. The huge amount of data and information being available easily and for free on social media sites are what make them a popular choice for researchers.

The COVID-19 pandemic, despite only being declared last year, has had studies and research conducted on it including many sources of literature published on sentiment analysis of COVID-19 using Twitter data. In this section, I will review some of the key literature that I have found, specifically the method of collecting the data, the tools used to analyse sentiment and finally the results and conclusions from the studies.

The article 'Twitter Sentiment analysis during COVID-19 Outbreak in Nepal' used the tweepy Twitter API to collect tweets that contained the hashtags '#COVID19' and '#Coronavirus' and to filter the tweet collection stream so that only tweets sharing their location in Nepal were saved. The study used Textblob to calculate and analyse the sentiment of the tweets. The dataset was collected for 10 days in May 2020, with a total collection of just over 600 tweets. Pokharel used Bar charts, pie charts and word clouds to display the results. One of the conclusions drawn mentions how the sentiment varies day to day but does not try to further delve into the reasons of varied sentiment, a factor that I will consider when evaluating my results. Some of the limitations mentioned in the study will also impact mine. However, the limitations caused from collecting data in 7-day period such as creating a small dataset will be alleviated as I will aim to collect my data over a 3-month period. (Pokharel, 2020)

'Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study' aimed to increase an understanding of the public awareness of COVID-19 pandemic trends and find themes of concern posted on Twitter. Tweets were collected by specifying keywords including '#covid_19', '#coronavirus', 'covid-19' and using the Twitter streaming API which is a Java application that connects to the real-time global Twitter stream. Data was collected between December 2019 and March 2020 and therefore focused on public perception at the beginning of the pandemic, with a total of 107,990 tweets collected. The data was analysed by scoping specific keywords and their frequencies and visualised for example through word clouds. Sentiment was analysed using the National Research Council (NRC) sentiment lexicon which analyses text for 10 different types of sentiments. The study found that over 70% of the tweets contained negative sentiment concluding that Twitter users had a negative outlook towards COVID-19 between their tweet collection periods. One of the limitations stated refer to the fact the analysis was done right at the beginning the pandemic, before it spread worldwide. In my project I collect tweets much further into the pandemic almost a year after it first began. (Boon-Itt, 2020)

'Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak' created a dataset of tweets with keywords including '#covid-19', '#StayHomeStaySafe', '#Lockdown', '#Quarantine' etc and filtered by using tweets that tagged locations from the top 10 most infected countries and the gulf region. There were over 50,000 tweets collected between June 21, 2020 and July 20, 2020. The tweets were collected through a RTweet Twitter API package for R programming

and the analysis tool used to find sentiment was the NRC emotion lexicon tool, which finds 10 sentiments in tweets. Data was analysed based on countries and sentiments with word clouds, bar charts and line graphs used to visualise the results. The study found that the countries that were almost infected had mostly a positive sentiment towards the pandemic and countries such as USA and Chile which had a balance between negative and positive tweets. Limitations included the chosen keyword filters not taking in account of misspelt words and only saving tweets tweeted in English. (Mohammad Abu Kausar, 2021)

The final study I will review is 'Sentiment Analysis on COVID-19 Twitter Data'. Which focused on the sentiment of people in India. In this study a dataset was collected using the GetOldTweets Python Twitter API with tweets being collected between November 2019 and May 2020 that had one or more of the three keywords 'corona', 'COVID' and 'COVID19'. A total of 140,000 tweets were collected and the dataset was separated by national states and by month. The tweets' sentiments were analysed by using the Textblob API. Insights and analytics were mostly found through frequency visualisation of bar charts and line graphs. The results showed over 60,000 tweets with positive sentiment over the whole collection period. The study delved deeper into each state and found which states were tweeting more frequently than others. There was also analysis done of sentiment on days when national lockdowns were declared, and it was observed that there was higher frequency of tweets during those announcements. (T. Vijay, 2020)

By reviewing the available literature, it has become clear that there are some common themes found in analysing Covid-19 Sentiment on Twitter but there is currently no literature available on the research I aim to focus on. There are many similarities between my projects approach and the approaches taken in the literature reviewed. For example, I will be using the Tweepy as my Twitter streaming API and analysing sentiment using the Textblob API. Similarly, the tools for analysis and visualisation will overlap. However, some of the key differences in my project is that it focuses on the sentiment of people tweeting in the United Kingdom and the dataset it collected over a 3-month period in which the country is in their third lockdown where vaccines are available and rolling out rapidly.

Project Aim

The aim of this project is to analyse the sentiment that people in the United Kingdom have about Covid 19 using Twitter. To achieve this aim a set of objectives must be accomplished which include:

1. Creating a tweet dataset with tweets that have been tweeted within the United Kingdom and are about the covid-19 pandemic.
2. Defining the sentiment of each of the tweets in the dataset.
3. Pre-processing the dataset for data analysis to find overall themes and insight on the sentiment of people in the United Kingdom.
4. Analysing the sentiment of the dataset.

Approach

Creating a Tweet Database

The goal of this project is to analyse Covid-19 sentiment in the United Kingdom on Twitter. To achieve this the first aim was to create a dataset of tweets that have been tweeted from within the United Kingdom and are related to the pandemic.

Selecting keywords

In order to narrow down the tweets a list of the following 12 keywords were chosen:

1. Coronavirus
2. Covid
3. Corona
4. Lockdown
5. Social distancing
6. Isolation
7. Tier
8. Face Masks
9. Vaccine
10. Pandemic
11. New normal
12. Boris Johnson

These words were mainly selected by utilising my own individual understanding and knowledge of the most frequently used words and phrases when discussing Covid-19 and the pandemic in the United Kingdom. I compared my list with online articles, however since the pandemic is unprecedented there was no correct way to find the top phrases and words.

The list also has keywords that are more tailored towards the United Kingdom's approach to fighting the Coronavirus. For example, the keyword 'Tier' refers to the restrictive measures put in place in December 2020 (GOV.UK, 2020) by the UK government to create multiple levels of restrictions in parts of the country dependent on the epidemiological indictors of that area. The keyword 'Boris Johnson' was used as being the Prime Minister of the United Kingdom. During the pandemic he addressed the nation with the updates and restrictions being put in place due to the coronavirus. Vaccinations have been a prominent conversation throughout the pandemic. From initial conversations regarding the curiosity and hope around when and if a vaccination for covid-19 will be created, through to the effectiveness and comparisons of the produced vaccines and its roll out scheme. For that reason, the keyword 'Vaccine' was also added to the keyword list.

Tweepy

The next step after identifying the keywords was to connect to the Twitter API using Tweepy and to stream tweets that matched my criteria and save them into an Excel file. To do this I first created a Stream Listener this requires authentication with 4 unique keys that are provided when creating a Twitter developer account.

The Tweepy stream listener allows multiple different filters to be applied to it. However, it treats every filter as an 'OR' and not as an 'AND' input. This meant that if I applied a filter for the keywords and a filter for the location the stream would return all tweets that were either tweeted in the United Kingdom, contained one of the keywords or both. This stream would not produce the desired results and to overcome this I created a stream that would return every tweet tweeted in the United Kingdom and that was tweeted in English then added a check for every incoming tweet to see if it contained any of the chosen keywords and if so that tweet's content could be saved into the Excel file.

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

myStreamListener = MyStreamListener()
myStream = tweepy.Stream(auth , myStreamListener)

try:
    #print("trying to connect to Twitter API ")
    myStream.filter( locations=[-6.38,49.87,1.77,55.81], languages = language)
except KeyboardInterrupt:
    print("Stopped")

finally:
    myStream.disconnect()
```

Figure 0: a code snippet showing the method of running a Twitter Stream.

Every tweet streamed has over 150 attributes presented in a JSON format. For the project not all the attributes were necessary. In fact, only 3 attributes were saved in the Excel file, 'text', 'place' and 'created_at'. However, the attributed 'retweeted status' and 'extended tweet' were used to make sure the content of tweet text was fully retrieved.

```
if hasattr(status, 'retweeted_status'):
    try:
        tweet =  status.retweeted_status.extended_tweet["full_text"]
    except:
        tweet =  status.retweeted_status.text
else:
    try:
        tweet = status.extended_tweet["full_text"]
    except AttributeError:
        tweet = status.text

try:
    tweet = unidecode(tweet)
except:
    print("unable to unidecode tweet")
```

Figure 1: a code snippet to show the if else statement to extract the tweet text.

The if-else statement above checked if the tweet streamed is a retweet. If the tweet was a retweet the text content from the retweet attribute would be stored at the tweet text. If the

tweet was not a retweet the content from the regular text attribute would be stored as the tweet text. Both scenarios also check to see if the tweet was more than 140 characters and if so, it stores the content from the extended tweet attribute.

Another point that I noted when writing the program is that not every tweet tweeted in the United Kingdom would a have geo tagged location. It was important to save tweets that had a geo tag so that when analysing the data later the dataset could be split by location. I therefore added another check after verifying that the tweet contained one of the keywords, which would make sure that the tweet had a geo tag. Once both these conditions were met the 'Tweet text', 'Location' and 'Time and Date' were all inserted into the Excel File.

```
if any(word.lower() in tweet for word in keywordlist ):
    keywordcheck = True
    #print("true word in  keywordlist")


if status.place is not None:
    #print ("location valid")
    # because of the streamer all tweets will be within the United Kingdom so only need to check for a geo tag
    locationcheck = True


if locationcheck and keywordcheck:
    #print("both true")
    excelinput = []
    tweet= str(tweet)
    print (">>> " + tweet)
    location = str(status.place.full_name)
    print (location)
    date = str(status.created_at)
    print (date)
    excelinput.append(tweet)
    excelinput.append(location)
```

Figure 2: a code snippet of the checks made to ensure the tweet streamed matches required criteria.

To stream and save the tweets, I would manually have to start and end the python program throughout the collection period. I collected tweets from 26 February 2021 till 11 May 2021 and accumulated over 30,000 tweets.

Sentiment Analysis

Once the program to stream and save tweets from Twitter was functioning, the next step was defining the sentiment of each of the tweets. I used two different sentiment analysis tools to analyse sentiment of the tweets, NRCLex and Textblob. The reason for choosing two different sentiment analysis tools was due to the different scores they both gave. The NRCLex library

	A	B	C	D	E	F	G	H	positive	ne	J	K	L	M	N
1	Tweet Text	Location	Time and Date	fear	anger	anticip	trust	surprise	positive		negative	sadness	disgust	joy	anticipation
2	@SkyNews If you don't want	Aberaman, \	2021-02-24 13:35:55	0	0	0	0.142857	0	0.285714		0.1428571	0	0.142857	0.142857	0.142857143
3	@Sheppard250 @wildmount	South East, I	2021-02-24 13:37:53	0	0	0	0	0	1		0	0	0	0	
4	Just had my Covid vaccine toc	Ealing, Lond	2021-02-24 13:38:30	0	0	0	0.166667	0	0.5		0	0	0	0.166667	0.166666667
5	Today's lockdown reading is..	Castlereagh,	2021-02-24 13:38:39	0	0	0	0	0	1		0	0	0	0	
6	@liz_lizanderson Can I join yc	Camberwell	2021-02-24 13:40:28	0	0	0	0	0	0.5		0	0	0	0.25	0.25
7	Doesn't bode well for my trip	Bangor, Nor	2021-02-24 13:41:08	0	0	0	0.333333	0.166667	0.333333		0	0	0	0	0.166666667
8	Good to hear at least they are	Lambeth, Lo	2021-02-24 13:41:22	0	0	0	0	0	0.333333		0.3333333	0.333333	0	0	
9	Spring Budget 2021: what to e	Houghton-le	2021-02-24 13:41:37	0	0	0	0.25	0.25	0.25		0	0	0	0	0.25
10	End of April feels #lockdown I	Scotland, Ur	2021-02-24 13:41:41	0	0	0	0	0	0		0	0	0	0	
11	@Sprinter0712 Don't be so dr	Hull, Englan	2021-02-24 13:41:52	0.166667	0.166667	0	0	0	0.166667		0.1666667	0.166667	0.166667	0	
12	@taxitalkmike @saferoadsno	Camberwell	2021-02-24 13:42:16	0.125	0	0	0.125	0	0.375		0.125	0.125	0	0	0.125
13	First vaccine done https://t.c	Dundonald,	2021-02-24 13:46:40	0	0	0	0	0	1		0	0	0	0	
14	@11DICKO @mikesaltsman19	Stockport, E	2021-02-24 13:46:47	0.333333	0	0	0	0	0		0.3333333	0	0	0	0.333333333
15	@BallouxFrancois I thought 't	Wetherby, E	2021-02-24 13:48:22	0.142857	0	0	0.142857	0	0.285714		0.2857143	0	0	0	0.142857143
16	Post-lockdown academic fant	Sheffield. En	2021-02-24 13:48:34	0.066667	0.066667	0	0.2	0	0.266667		0.0666667	0.066667	0.066667	0.133333	0.066666667

Figure 3: a snippet of the excel file created after running the NRCLex sentiment tool.

analysed text to 2 sentiment and 8 emotions and gave 10 separate scores and the Textblob library presented a single polarity score. Both provided interesting and relevant results, and both could help identify different themes to the tweet. I analysed sentiment using NRCLex in the same program that I streamed the tweets. This method allowed for the Tweet Text, Location, Time and Date and all of sentiments and emotions scores related to be saved in the same excel file at the same time:

To append all the data to the excel file I used a list data structure called 'excelinput' which would get created if the tweet passed all the required checks. Once the sentiment had been analysed for that tweet the results would be appended to the list and 'excelinput' would be added a new list to the Excel file.

```
# get first element the tweet ie (tweet text) and run a sentiment anlaysis on it
text_object = NRCLex(tweet)
emotion = text_object.affect_frequencies

# for each element in the dictionary append that to the list
for key in emotion.values():
    excelinput.append(key)
```

Figure 4: a code snippet of the code to analyse sentiment and appending it to a dictionary..

The first line in the code above, takes the tweet text and runs it in the sentiment analysis classifier and the second line saves the values of each of the scores assigned to a dictionary called emotion. There is a key for each sentiment and emotion and a corresponding key which contains the score returned from the NRCLex classifier. Only the score is required and stored in the excel file so using a for loop each score is appended to the 'excelinput' list.

For the Textblob sentiment analysis classifier, I ran the tweets through the sentiment tool once all the tweets had been collected. I duplicated the tweets contents of the Excel file from the NRCLex analysis and created a separate python program, that would iterate through each line in the new excel file and classify the sentiment polarity of each of the tweets and save and append the score to as a new column.

```
for row in sheet.iter_rows(min_row=1, min_col=1, max_row=sheet.max_row, max_col=2):
    print ( "on row: "+ str(y))
    tweet = []
    # get each row from file and append to list
    for i in row:
        tweet.append(i.value)

    # get first element the tweet and run a sentiment anlaysis on it
    text= str(tweet[0])
    emotion = TextBlob(text).sentiment.polarity
    emotion = str(emotion)

    #for key in emotion.values():
    tweet.append(emotion)

    #append the whole list to new excel file
    newsheet.append(tweet)
```

Figure 1: a code snippet of the method to analyse the tweets with the Textblob tool.

The code snippet above shows the for loop created to iterate through each row of the Excel file. The variable 'emotion' in this program stores the polarity score and is converted to an string before appending it the list 'tweet' that contains the content for each row.

Data Analysis

Once all the tweets had been collected and run through both sentiment analysis classifiers it was time to pre-processes the dataset for analysis. This involved creating a word dictionary that would be used to create word clouds and a write a python program that would calculate the PMI scores.

Creating a word dictionary

I created a new python program to create the word dictionary. I needed to create two different word dictionaries for the word clouds: one for the number of word occurrences in positive tweets and one for the number word occurrences in negative tweets. To make the positive word dictionary, I iterated through the Excel file generated after the Textblob analysis and first checked if the polarity score was greater than 0, which would indicate that the tweet is positive in sentiment.

The program then tokenizes the tweet into words and then each of the words is passed through a lemmatizer, this process groups all the inflected words together and would help produce a better understanding of what topics and word concepts were being tweeted about the most. After lemmatizing the words, I would create word dictionary by adding every new word in the dictionary with a corresponding counter value and increase it every time the word re appeared. The word dictionary was then sorted from most frequently used word to least frequently and the first 500 words that were not a stop word were saved in a text file. This same process was repeated for the negative dictionary by modifying the python script to save and process tweets that had a sentiment polarity score of smaller than 0.

```
if sentiment > 0:

    sentimentcounter = sentimentcounter + 1

    #clean tweets

    text = word_tokenize(text)


    text = [lemmatizer.lemmatize(w) for w in text]


    for words in text:
        if words in word_dict:
            word_dict[words] = word_dict[words] + 1
        else:
            word_dict[words] = 1
```

Figure 2: a code snippet of how the word dictionary was made.

Calculating Pointwise Mutual Information

To write the program that would calculate the pointwise mutual information score between frequently occurring words and positive and negative tweets, a similar approach to the word dictionary program was taken. The first part of the program involves creating a word dictionary using the exact method in the word dictionary program. However, instead of only saving words from specific sentiment, a counter for the number of times a tweet uses that sentiment is created. This is so the probability of an independent occurrence of positive tweet can be calculated and each word will have the number of its occurrences stored in the dictionary which can be used to calculate each word's independent occurrence probability.

```
for key , v in word_dict:
    if key not in stopwords.words() and key.isalpha() and v > 100:
        #go through tweets and store the number of times both the sentiment and keyword have appeared in a tweet
        print("starting loop")
        for row in sheet.iter_rows(min_row=2, min_col=1, max_row=sheet.max_row, max_col=13):
            tweet = []
             # get each row from file and append to list
            for i in row:
                tweet.append(i.value)

            text= str(tweet[0])
            text = text.lower()
            sentiment = tweet[8]

            if key in text and sentiment > 0:
                #print("in for loop")
                both = both + 1
        #calculate PMI
        probability_a = k/y
        probability_b = sentimentcounter/y
        probability_a_b = both / y
        try:
            PMI = math.log(probability_a_b/(probability_a * probability_b))
        except:
            print("PMI erorr")

        datainput = str(key)+","+str(PMI)+","+str(k)+","+str(sentimentcounter)+","+str(both)
        f.write(datainput + "\n")

    both = 0
```

Figure 3: A code snippet showing the method of calculating PMI scores.

To find the number of times both events occur together the Excel file is iterated through and a counter called 'both' increases each time each event occurs. This process is inside multiple for loops to ensure each combination is calculated. To limit the number of words processed and increase speed of the program only words that occurred more than 100 times were passed through. The for loop was run twice, once for the positive tweets and once for the negative tweets. Both times the results were stored in a text file so that the words with the highest PMI scores could be analysed.

Results

After collecting and processing the dataset the final step was to analyse the tweets. The tweets dataset was analysed using PowerBI. The data was classified using two sentiment tools, the NRCLex sentiment analysis tool which assigned a value between 0 and 1 for each the emotions and sentiments, the closer the value is to 1 the more associated the tweet is to that sentiment or emotion. The second classifier was the Textblob Sentiment analysis tool which assigned a score between -1 and 1 is assigned, depicting a sentiment polarity with a score being between -1 and -0.9 as negative 0 as neutral and a value between 0.1 and 1 as positive. Depending on the type of analysis being done one of the results from the tools were chosen as the data.

In figures 9, 11, 19 – 22 the data used was analysed using Textblob and in figures 10, 12, 13-16 the data used was analysed using NRCLex. The data for the word clouds (figures 17 and 18) used the results from the word dictionary python program. The data is Table 1 was from the results of the pointwise mutual information python program. The visualizations and analysis chosen were to try and answer some of the research questions mentioned in the introduction and to uncover and find meaningful themes about covid-19.

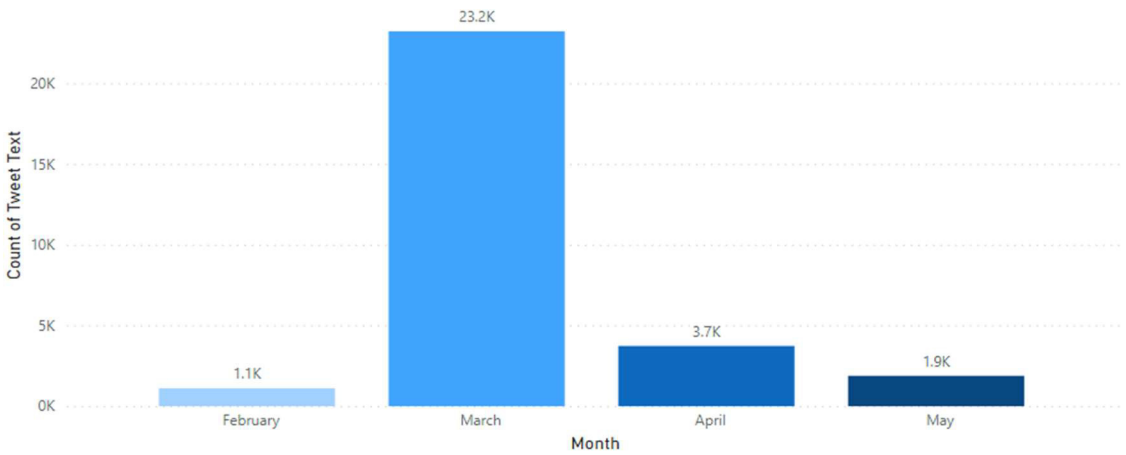

Figure 4: A column chart to show the number of tweets collected per month.

The total number of tweets in the dataset was 30,008 the breakdown by month is shown in figure 8. However, in figures 12-15 the total number of tweets compared is 20,446 due to cleaning of the data by location.

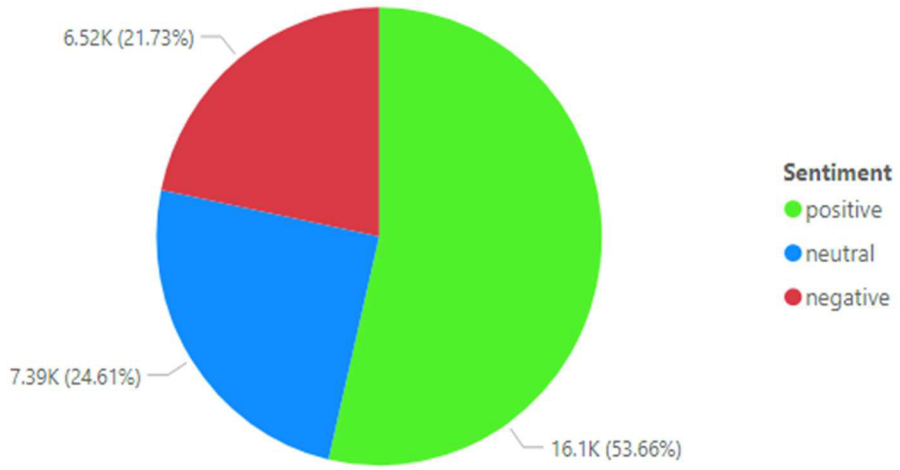

Figure 5: A pie chart to show the distribution of positive, negative, and neutral sentiment.

In the pie chart, figure 9 This pie chart shows that 16,100 tweets in the dataset collected between 26 February 2021 – 11 May 2021 had a positive sentiment polarity, accounting for over 50% of the tweets. 7,390 tweets had a negative sentiment polarity and 6,520 had a negative polarity, showing that there were more tweets with a neutral sentiment than a negative sentiment. People in the United Kingdom conveyed a more positive sentiment compared to a negative or neutral sentiment during the time period the tweets were collected in.

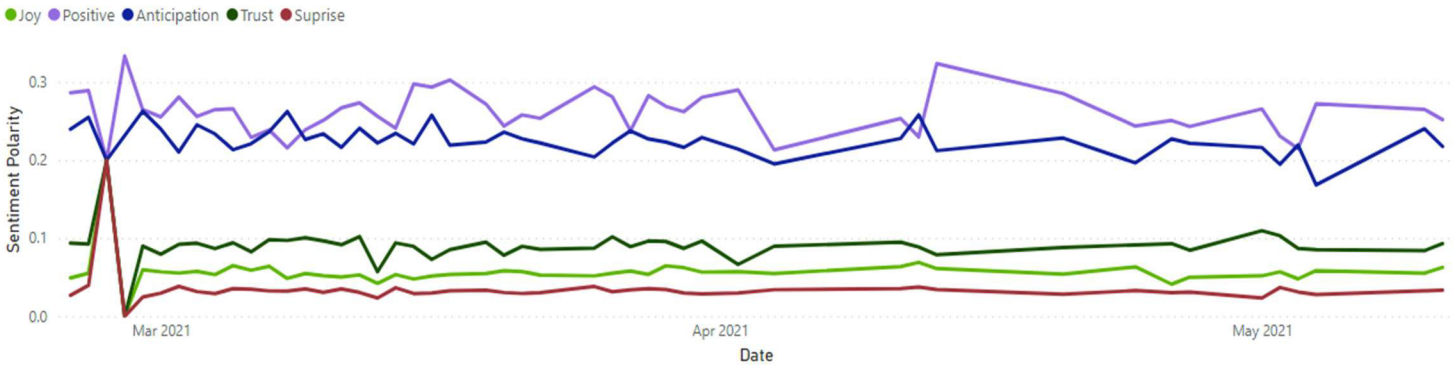

Figure 6: A Line Chart to Show Average Polarity of Joy, Positive, Anticipation, Trust and Surprise Sentiment Over Time

In figure 10, a line chart of the Joy, Positive, Anticipation, Trust and Surprise sentiments over the collection period is shown. On Average, Positive sentiment and Anticipation are the highest in the tweets almost consistently having a score between 0.2 and 0.3. The Joy, Trust and Surprise sentiment lines remain similar throughout the collection period. There is a sudden drop in Joy, Trust and Surprise emotions on 13th March 2021 this could be due to the fact it had been a year since covid 19 had hugely impacted the world. There is a sudden increase of Positive sentiment on 13th April, this is the day after the restriction easing began in England which may indicate that with easing of restrictions and re-opening of services such as

hairdressers people react more positively. Anticipation and Positive sentiment remain similar up until 12th April when Positive sentiment remains high, and the Anticipation sentiment line drops slightly. This finding could indicate that there was a high anticipation level with the easing of restrictions.

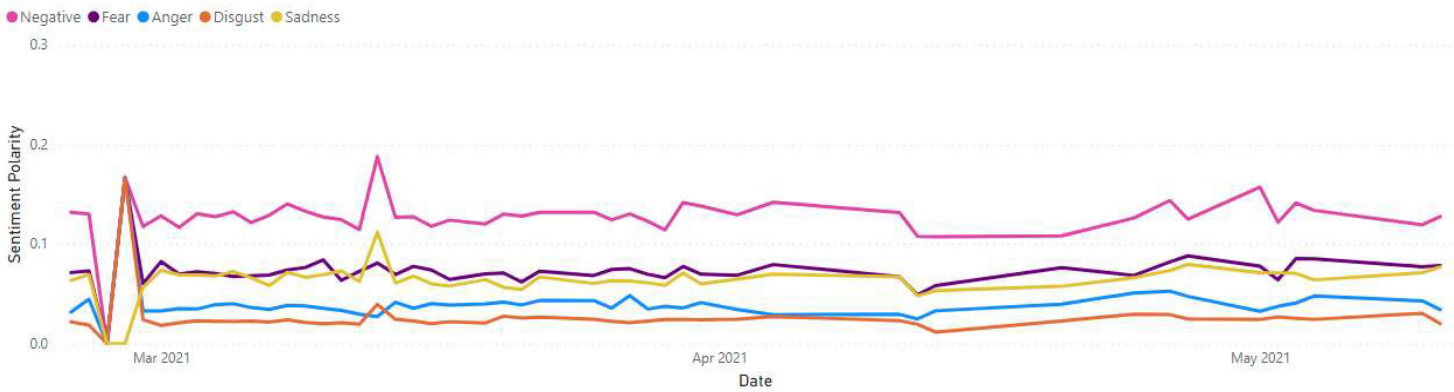

Figure 8: A Line Chart to Show Average Polarity of Negative, Fear, Anger, Disgust and Sadness Sentiment Over Time

In figure 12, the line graph for Negative, Fear, Anger, Disgust and Sadness is shown over the collection period. The 5 sentiment lines remain almost in sync with each other the whole time, the Negative sentiment remains within the 0.1 and 0.2 range and the other 4 sentiments stay between 0 and 0.1. there Is a sharp increase in Negative, Disgust and Sadness sentiment on the 13 March 2021, this finding correlates with the finding in figure 10 where there was a decrease in Positive sentiment around the same time, showing that people become more negative towards the pandemic due to the fact it had been a year since covid-19 had impacted the country. From April 12 there was decrease in Negative sentiment again this finding correlates with the finding figure 10, since this was the day restrictions eased in England.

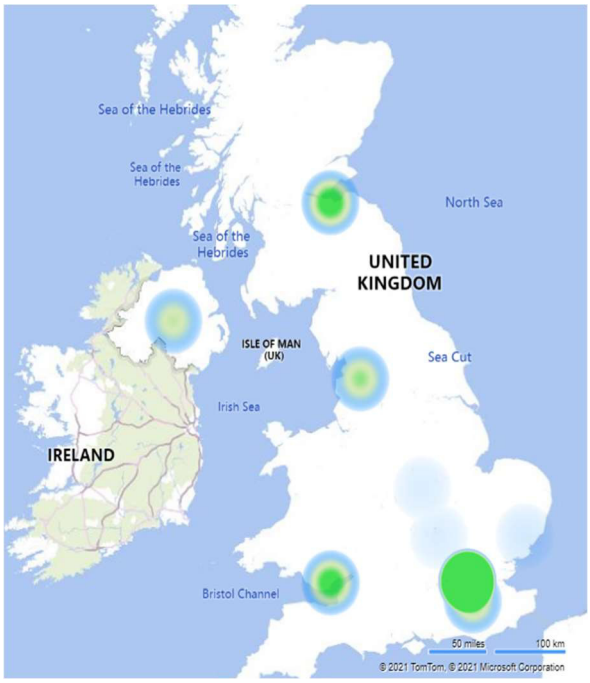

Figure 7: A heatmap showing average sentiment across the country.

In Figure 11, the heatmap shows the average sentiment polarity in different regions of the United Kingdom. The colour scale ranged from red representing a negative Sentiment to green to represent a Positive sentiment. The sentiment polarity used was from the Textblob sentiment analysis where the sentiment polarity was between -1 and 1. As the map shows the average of all the tweets between 26 February 2021 – 11 May 2021 and as seen in figure 9 that the majority of tweets were positive, mostly the map shows green sections which indicated positive sentiment in that area. There was an extremely high Positive sentiment around South East of England. Around Wales and Scotland there was a higher Positive sentiment than Neutral and around the North West of England and Northern Ireland a more Neutral sentiment than Positive sentiment. A primarily neutral sentiment was observed around the East Midlands.

The next 4 figures were created to analyse if there was significant change in sentiment during the different months in the collection period.

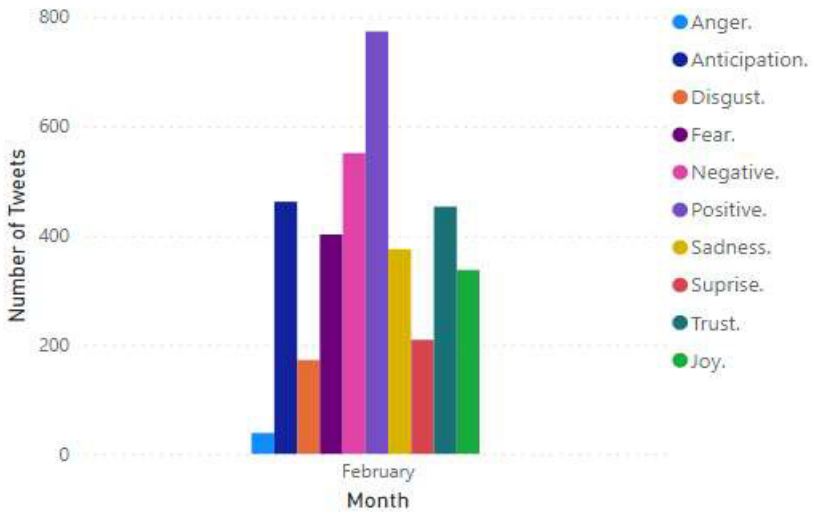

Figure 9: A bar chart to show the number of tweets by sentiment in February.

Figure 13 is a bar chart to show the number of tweets by sentiment collected in February and as shown in Figure 1, a total of 1,108 tweets were collected in February. During February 2021, the United Kingdom was still in its 3rd official lockdown but during the days the tweets were collected in this month the roadmap to restrictions easing had been published (gov.uk, 2021). The graph shows most of the tweets were Positive in sentiment in February with 774 tweets, followed by Negative Sentiment with a total of 551. Anger was the least prevalent sentiment with only 38 tweets. Over 400 tweets with Anticipation and Joy sentiment were found. Over 35% of the tweets this month showed anticipation this could be due to the fact the roadmap had been released and people were waiting for the restrictions to ease.

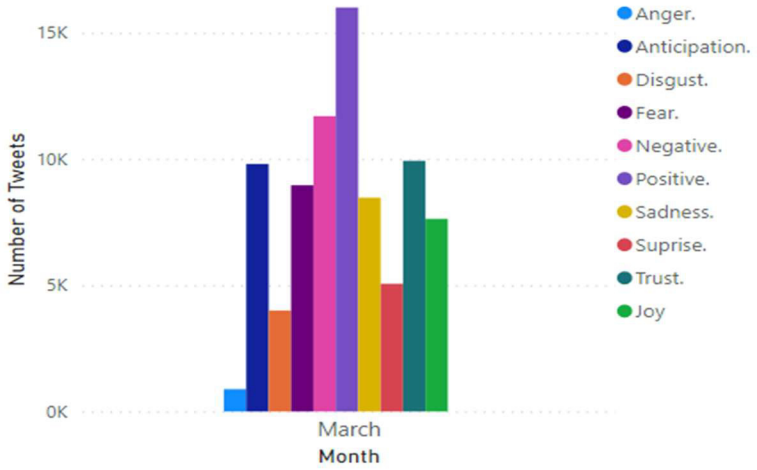

Figure 10: A bar chart to show the number of tweets by sentiment in March.

Figure 14 is a bar chart to show the number of tweets by sentiment in March. As shown in Figure 1, 23,220 tweets were collected in March. There is great similarity between the bar chart for February and March with Positive sentiment tweets prevailing with 16,000 tweets and almost 12,000 tweets with Negative sentiment. However, Surprise was one of the sentiments to have grown between the months. The sentiment Surprise accounted for more than 20% of the tweets in March. March 2021 marked a year since the pandemic had been announced and a high number of tweets with surprise emotions could indicate people were not expecting to still be dealing with covid-19 for a whole year (Gov.uk, 2020).

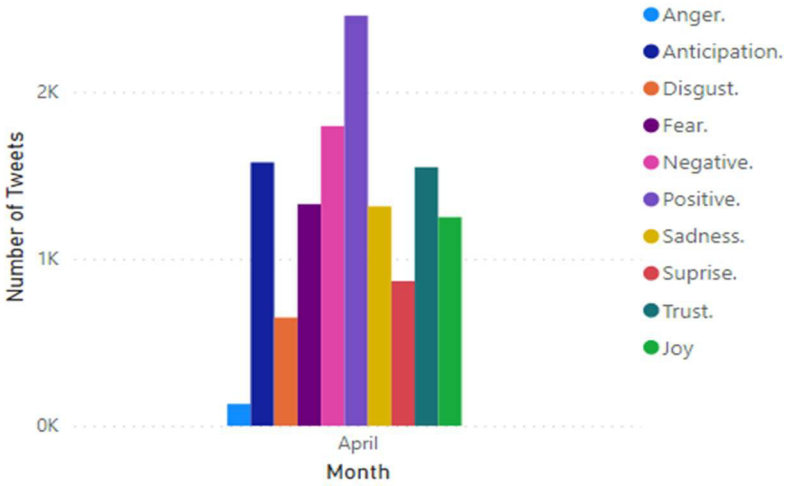

Figure 11: A bar chart to show the number of tweets by sentiment in April.

Figure 15 is a bar chart to show the number of tweets by sentiment in April. As shown in Figure 1, 3,734 tweets were collected in April with 2456 of those having Positive sentiment making it the most common sentiment that month just as the previous 2 months. One of the changes in the month of April includes an increase of Sadness sentiment in tweets, apart from that the

distribution of sentiment is similar with the previous months. The increase in the number of tweets with sad sentiment is surprising since in this month a lot of the restrictions began to ease (gov.uk, 2021) and cases in the United Kingdom had dropped. (JHU CSSE COVID-19 Data, 2019)

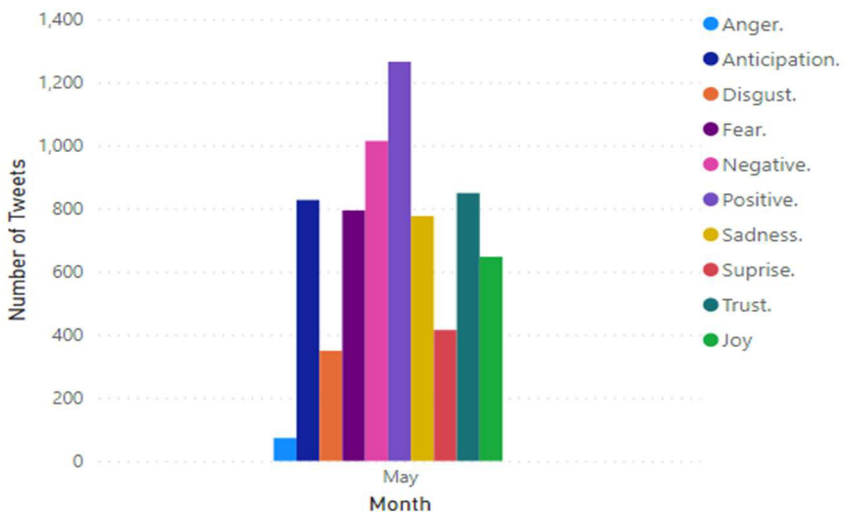

Figure 12: A bar chart to show the number of tweets by sentiment in May.

Figure 16 is a bar chart to show the number of tweets by sentiment in May. As shown in Figure 1, 1,884 tweets were collected in May. This was the last month of collection and displays a very similar distribution between the sentiments as the previous 3 months, over the whole collection period Positive sentiment has been the most common and Anger has been the least used. An increased number of tweets with a Fear sentiment is seen in May with over 40% of tweets showing that sentiment compared to the previous month where Fear sentiment was seen in 35% of the tweets. The increase in fear may be due to the fact a new 'variant of concern' had been discovered and cases of that variant had been found in the United Kingdom people could potentially have been fearing another lockdown or set of tighter restrictions.

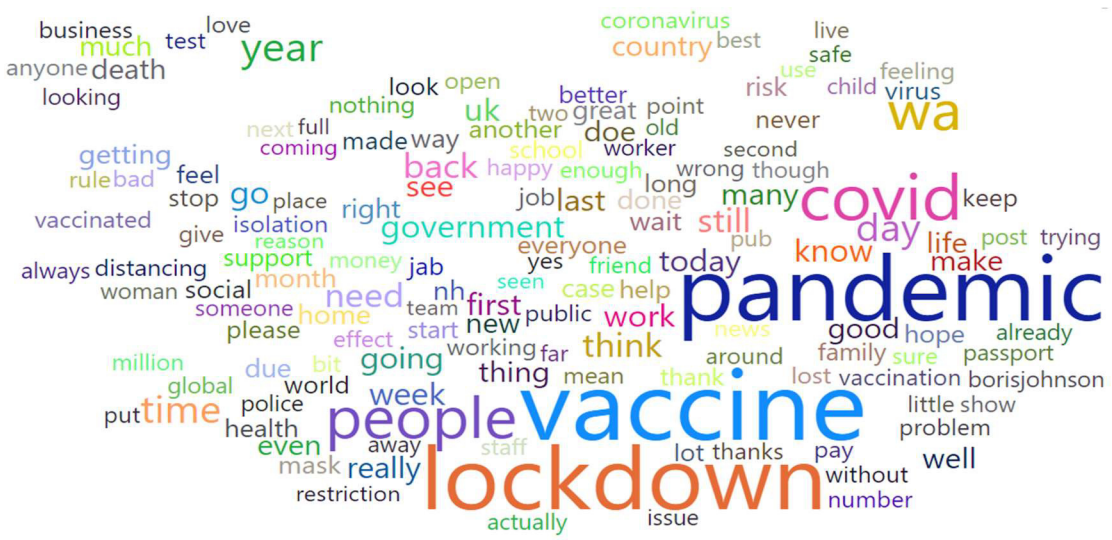

Figure 13: Word Cloud to show the most common words used in Negative Tweets.

Figure 17, shows a word cloud of the top 150 most common words found in Negative tweets. The 5 most common words include vaccine,pandemic, lockdown, covid, people 4 of these words were part of the keywords. Other noticably interesting words not in the keywords list include government, passport, pub, death, distancing.

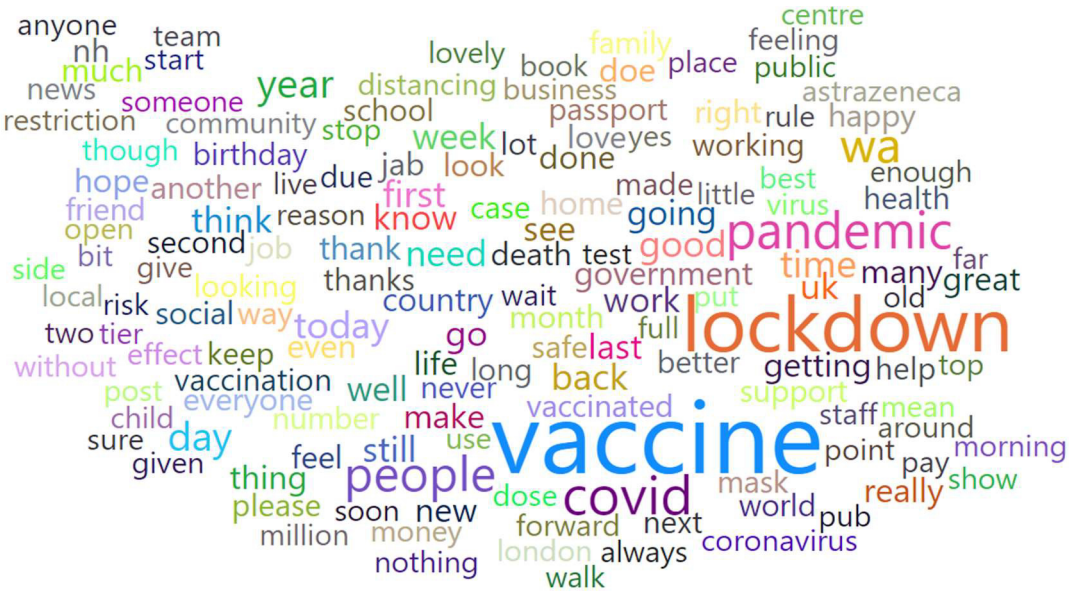

Figure 14: Word Cloud to show the most common words used in Positive Tweets:

Figure 18 shows a word cloud of the top 150 most common words found in Positive tweets. Interesting vaccine was the most popular for positive and negative tweets followed by lockdown, covid, pandemic, people again very similar to the negative tweets. However, other common words not in the keyword list include day, astrazenca, hope, health.

The next four figures were created to analyse and compare the sentiment between the four constituent countries in the United Kingdom.

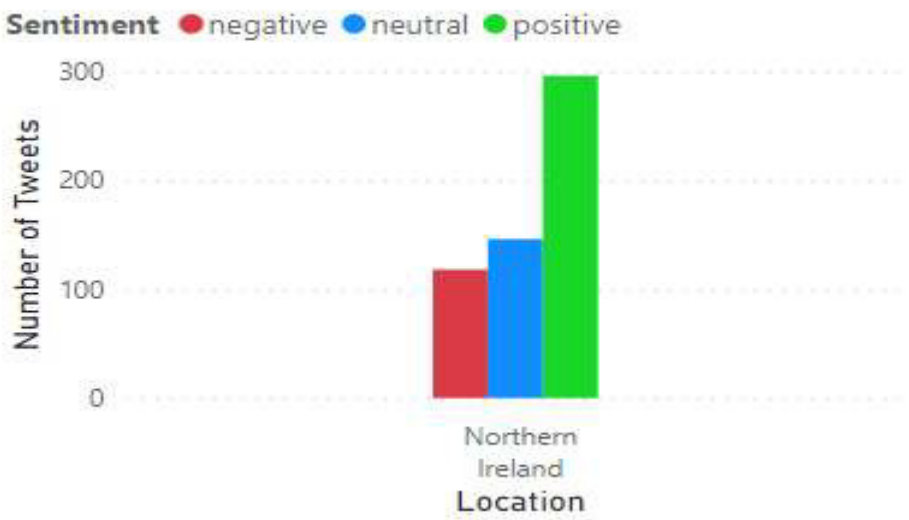

Figure 15:A bar chart to show the number of tweets tweeted by sentiment in Northern Ireland.

Figure 19 , shows a bar chart of the number of tweets by Sentiment that were geo tagged on Twitter with Northern Ireland. The majority of tweets were of Positive sentiment indicating that overall people in Northen Ireland had a Positive sentiment about the pandemic. There were more tweets with a Neutral sentiment than with a Negative sentiment which corresponds with the overall sentiment of the country as seen in figure 8.

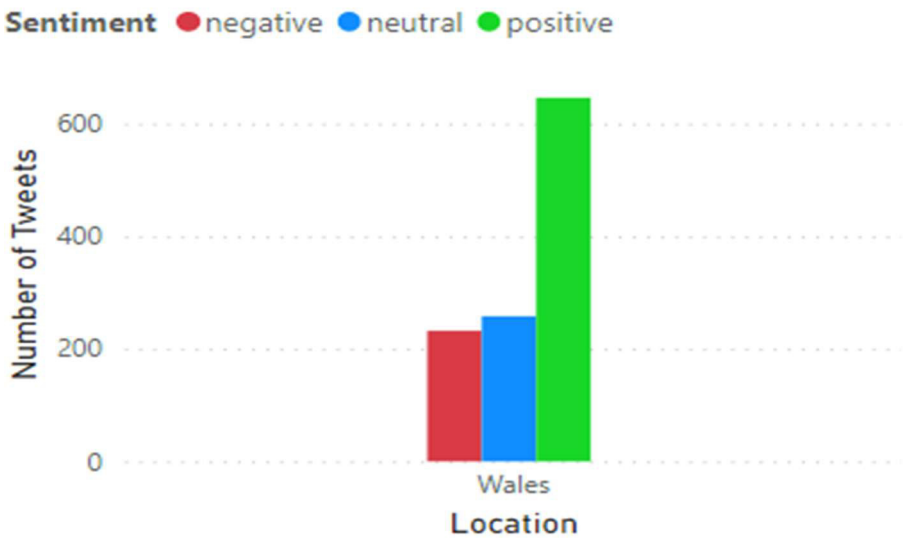

Figure 16 : A bar chart to show the number of tweets tweeted by sentiment in Wales.

Figure 20, shows a bar chart of the number of tweets by Sentiment that were geo tagged on Twitter with Wales. The majority of tweets were of Positive sentiment indicating that overall people in Wales had a Positive sentiment about the pandemic. There were more tweets with a Neutral sentiment than with a Negative sentiment which is the same as what Northen Ireland has as shown in figure 19 and corresponds with the overall sentiment of the country as seen in figure 8.

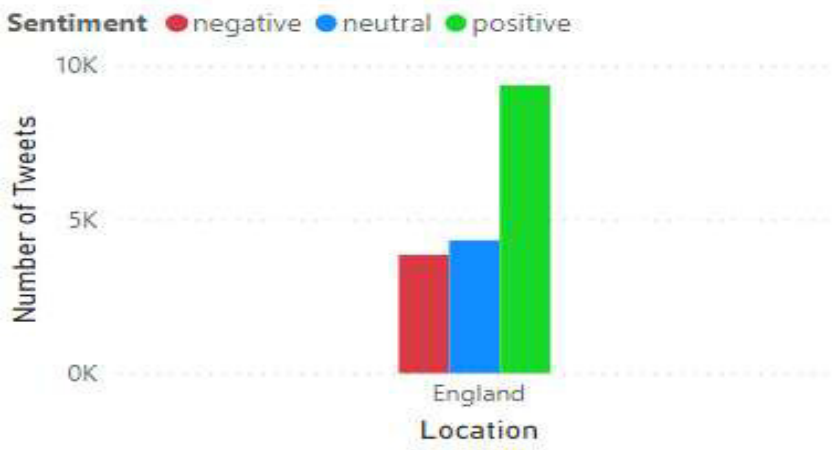

Figure 17: A bar chart to show the number of tweets tweeted by sentiment in England.

Figure 21, shows a bar chart of the number of tweets by Sentiment that were geo tagged on Twitter with England. The majority of tweets were of Positive sentiment indicating that overall people in England had a Positive sentiment about the pandemic. There were more tweets with a Neutral Sentiment than with a negative sentiment which corresponds with Wales and Northen Ireland as seen in Figure 19, 20 and the overall sentiment of the country as seen in figure 8.

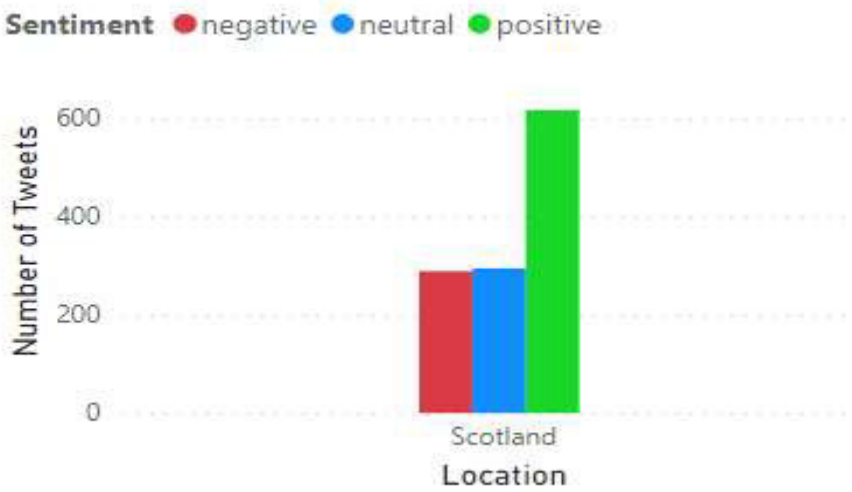

Figure 18 : A bar chart to show the number of tweets tweeted by sentiment in Scotland.

Figure 15 shows a bar chart of the number of tweets by Sentiment that were geo tagged on Twitter with Scotland. The majority of tweets were of Positive sentiment indicating that overall people in Scotland had a Positive sentiment about the pandemic. There were almost and equal number of tweets with a Neutral sentiment and a Negative sentiment which doesn't match the Northern Ireland, Wales,and England in figure 19, 20 , 21 and doesn't corresponds with the overall sentiment of the country as seen in figure 8. The Neutral sentiment and Negative sentiment each account for 24% of the total tweets collected in Scotland.

No.	Positive	Negative
1	excited	catch
2	ahead	office
3	scotland	power
4	science	current
5	return	long
6	giving	fear
7	medical	feel
8	glad	human
9	covidvaccine	trust
10	keep	tier
11	massive	lost
12	become	card
13	completely	global
14	bit	pandemic
15	travel	see
16	based	contract
17	including	law
18	trip	bring
19	daughter	may
20	wonderful	tell

Figure 23: A table to show the top 20 words with the highest PMI value for Positive and Negative sentiment.

Table 1 shows the top 20 words with the highest Pointwise Mutual Information (PMI) value for both Positive and Negative Tweets. The word with highest PMI value in Positive sentiment tweets was 'excited' which means the word occurred more frequently than expected in Positive tweets, a word not maybe directly related to the pandemic. Some of the interesting words in the positive sentiment column include Scotland, science, medical, covidvaccine, travel words that can be associated with the pandemic and showing a more positive outlook towards those topics.

The word with the highest PMI value in Negative sentiment tweets was 'catch' a word that could be associated with the pandemic and more particularly with 'catching Covid-19'. Other interesting words include office, fear, trust, tier, lost, global, pandemic, contract that can be associated with the pandemic showing the topics and themes that were associated with a more negative outlook due to the pandemic.

Summary of Results

After analysing the graphs and findings above it can be said that overall, throughout the collection period the United Kingdom had a positive outlook towards the covid-19 on Twitter. With over 50% of tweets showing a positive sentiment and it being the most common sentiment found over the 4 months and in each of the 4 constituent countries in the United Kingdom. This finding opposes with the findings found when researching similar work, the study on 'Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study' concluded a negative outlook on the pandemic due to its high number of negative tweets. This finding from my project indicated that people were much more positive about covid-19 during my collection period. This is something I anticipated since people were more used to the idea of being in a lockdown, practicing social distancing and there was a 'light at the end of the tunnel' due to the steady vaccination roll out. Vaccinations were also the most common theme found in the data, it was the most common word despite tweet sentiment and was also prevalent in the Positive PMI table. Showing that mostly people had a positive outlook towards the vaccine.

I also wanted to see if there was a different in sentiment between the 4 constituent countries but looking at the graphs there was not any significant all 4 countries had mostly a positive outlook. Another finding in the results showed that during significant restrictions easings the number of positive tweets increased which shows that people were happy that the country was coming out of the lockdown.

Evaluation of Results

Evaluation of collecting tweets.

Using the python library, Tweepy to connect to the Twitter API worked very well, due to the level of access to Twitter API multiple filters on one Twitter stream listener was not possible and a workaround was done but apart from that Tweepy was a very useful library. The method to stream and save the tweets also worked successfully and the tweets were saved in an easily readable Excel file. However, one improvement I would make in the method of collecting the tweets would include pre-processing part of the tweet before it is saved into the Excel file. Even though I pre-processed the tweets using Unicode, removing any links and @mentions from the Twitter may improve the sentiment analysis scores. Another improvement I would make to method would involve improving the manual aspect to collecting the tweets. Due to the fact I had to manually start the program every day and complete regular checks on the running system, improving the error handling of the python program could help alleviate part of the issue.

Evaluation of sentiment analysis tools

NRClex is an emotion-based sentiment analysis tool that uses lexicons, a list of words and the emotions that they convey to predict sentiment. It is a useful tool as it aims to detect different emotions as well as including positive and negative sentiment. This tool provided the tweets dataset with a more detailed and thorough analysis. The Textblob sentiment analysis returned both a sentiment polarity score that indicated whether the tweet was positive, negative, or neutral and a subjectivity score between 0 and 1 where 0 is very objective and 1 is very

subjective and also used a lexicon-based approach. One of the drawbacks of using a tool that associates the text with lexicons is that often the emotion wanting to be expressed is different to the one that is associated with the lexicon used. This could mean that some of tweets were incorrectly labelled.

Both tools were useful as they provided results using different methods of sentiment analysis. However, due to time constraints a comparison between the results was not done, a comparison could have provided more insight around the true sentiments of tweets and if both the tools provided similar sentiment assignments to the tweets.

Evaluation of analysis of results

Overall a large majority of the tweets collected and analysed were exactly what was required, some of the tweets were not usable due to the geo tag used when tweeted by the user. Tweets can be tagged using a certain geo tag that may not be necessarily accurate i.e., a full location was not provided or may not even be a real location. When processing the tweets for analysis based on locations only tweets that had tweeted using one of the 4 constituent countries that make up the United Kingdom were used England, Scotland, Wales and Northern Ireland. This meant that locations that had been tagged in part of the United Kingdom but not mentioned the related country was not included. An improvement in either the collection method of the location or processing post collection would be beneficial in the future to allow for more accurate analyses to be performed. All the analysis and visualisations created provided useful and interesting results but to further improve the analysis on the tweets certain factors could be more specifically analysed. For example, word clouds for specific dates could be created to see if key events are spoken about different than a regular quarantine day. More precise location analysis of counties could also have been performed than just analysing the differences between the constituent countries.

Conclusions

The aim of this project was to analyse the sentiment that people in the United Kingdom have about Covid 19 using Twitter. To achieve this aim a set of objectives had to be accomplished which included:

1. Creating a tweet dataset
2. Defining the sentiment of each of the tweets
3. Pre-process the dataset for analysis.
4. Analysing the sentiment of the dataset.

The first aim to create a tweet dataset was achieved. The tweets were collected by connecting and streaming the Twitter API using the Tweepy python library, the streamed tweets, the location tagged, the date and time of the tweet were then stored in an Excel file. I was able to collect 30,008 tweets over 26 February 2021 – 11 May 2021. I believe this was sufficient amount to be able to analyse people's sentiment about the pandemic, since a lot of the questions and findings I anticipated to find at the start of the project were achieved and observed. However, the method of collecting the tweets meant that I had to manually run and end the program every time to save tweets, I attempted to save tweets every day for the same amount of time but as the project progressed it become more difficult to maintain that. March had the greatest number of tweets collected compared to the other 3 months and collecting a larger dataset would have been more beneficial in getting a truer estimate of sentiment. The advantage to a larger dataset would include the possibility to explore more in detail the sentiment changes and pinpoint more significant dates and event for the changes, it would also make the findings more accurate.

The second aim to define the sentiment of each of tweets was achieved as well. I was able to classify the sentiment of the tweets using two different sentiment analysis tools, NRCLex and Textblob. Using NRCLex 2 sentiments and 8 emotions were analysed and using the Textblob tool a polarity score was given to each of the tweets. Having two different types of sentiment scores was useful when trying to analyse the data as different types of analysis were able to be conducted. However, due to time constraints I was unable to conduct comparison between the tools which could have helped in understanding how accurate the sentiment tools are.

The third aim to pre-process the tweet dataset and prepare it for analysis was successfully achieved. Once the data was collected it was cleaned using PowerBI. To be able to calculate Pointwise Mutual Information scores and to create the word clouds using the dataset a word dictionary was made and the use of the tokenizer and lemmatize tools from the NLTK library were utilised to create the dictionary.

The final aim to analyse the dataset was also achieved. Using the visualisations and tools available on PowerBI I was able to analyse the average sentiment score throughout the collection period, the number of tweets per sentiment every month, word clouds for Positive and Negative sentiment and more. Although, I was able to create multiple visualisations a more in-depth analysis could have been conducted if the project timescale was not as short.

Another benefit with a longer timescale would mean larger dataset could have been collected and more accurate sentiment could be gathered.

Overall, the results showed that the sentiment of people in the United Kingdom about Covid -19 was positive despite what conversations people were having in person around the pandemic, online on Twitter the majority of the tweets were positive in sentiment. There were peaks in sentiment during key dates, positive sentiment increased around the days when restrictions were being eased and increase in anticipation was also seen during the days before such key events. These results coincided with the results found in other similar studies reviewed. However, a result not found in the published literature that was found during my project was the topic of most tweets was around vaccines most likely due to the fact that there was a large rollout of vaccination during the collection period. Vaccine was the most common word found in tweets despite the sentiment of the tweet. This project has been able to successfully accomplish an analysis on the sentiment of the coronavirus pandemic in the United Kingdom using Twitter.

Future work

Within the timeframe provided to complete the project I believe I was able to complete and conduct a good analysis on Covid 19 sentiment in the United Kingdom on Twitter. However, if given more time I would have liked to conduct a comparison between the sentiment analysis tools used as well as testing other sentiment analysis tools available on the dataset collected. I would have also preferred to collect more data, more often over a longer period of time than just 4 months which could improve the accuracy of the findings and reveal more predictive factors for change in sentiment.

The project's results found some similarities with previously published literature and it's results such as the majority of tweets having a positive sentiment. In the future it could be interesting to see how sentiment changes 'post-pandemic' and if the keywords chosen are still topics of conversation.

Reflections on learning

When deciding to take on this project, I was aware that it would be a challenge. I chose a project in which I had very little knowledge and would require me to spend a lot of time researching and learning. I was always interested in sentiment analysis and was aware of the concept however, the process on how sentiment analysis is performed was a skill I had to learn and find tools that analysed sentiment. Connecting to Twitter and streaming the needed tweets was a completely unknown topic for me as well. However, on top of learning these new skills I was able also to further develop my knowledge in Python.

Having regular weekly meeting with my supervisor allowed me to get continual and useful feedback on the progress of the project and allowed me to stay on track. Reflecting on my time management, the first half of the project that was completed before easter was well executed and I was able to keep on track with the work plan I had written. Dividing the project into sub-tasks helped me keep on track and work efficiently. Due to personal reasons related to the pandemic, I was unable to work on the project properly for several weeks after easter.

Which meant that the second half of the project which involved analysing the results and writing up the report was completed on a much tighter time constraint than anticipated. Nonetheless, I was able to persevere and complete the report and produce a final project that I can be proud of.

References

Alto, V., 2020. *Understanding Pointwise Mutual Information in NLP.* [Online]
Available at: https://medium.com/dataseries/understanding-pointwise-mutual-information-in-nlp-e4ef75ecb57a
[Accessed 24 May 2021].

analysis, I. f. G., 2021. *Timeline of UK coronavirus lockdowns, March 2020 to March 2021.* [Online]
Available at: https://www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf
[Accessed 25 May 2021].

Anon., 2020. *NRCLex 3.0.0.* [Online]
Available at: https://pypi.org/project/NRCLex/#description
[Accessed 24 May 2021].

bbc.co.uk, 2020. *Coronavirus: Greatest test since World War Two, says UN chief.* [Online]
Available at: https://www.bbc.co.uk/news/world-52114829
[Accessed 22 May 2021].

Boon-Itt, S. &. S. Y., 2020. *Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study..* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7661106/
[Accessed 16 May 2021].

Das, S. K. A., 2021. *Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network..* [Online]
Available at: https://link.springer.com/article/10.1007%2Fs12065-021-00598-7
[Accessed 16 May 2021].

datacamp.com, 2018. *Stemming and Lemmatization in Python.* [Online]
Available at: https://www.datacamp.com/community/tutorials/stemming-lemmatization-python
[Accessed 24 May 2021].

Devika M D, S. C. A. G., 2016. *Sentiment Analysis:A Comparative Study On Different Approaches.* [Online]
Available at: https://core.ac.uk/download/pdf/82425196.pdf
[Accessed 27 May 2021].

docs.microsoft.com, 2021. *What is Power BI?.* [Online]
Available at: https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
[Accessed 24 May 2021].

GeeksforGeeks, 2020. *Removing stop words with NLTK in Python.* [Online]
Available at: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[Accessed 24 May 2021].

GOV.UK, 2020. *Full list of local restriction tiers by area.* [Online]
Available at: https://www.gov.uk/guidance/full-list-of-local-restriction-tiers-by-area
[Accessed 17 May 2021].

Gov.uk, 2020. *Prime Minister's statement on coronavirus (COVID-19): 23 March 2020.* [Online]
Available at: https://www.gov.uk/government/speeches/pm-address-to-the-nation-on-coronavirus-

23-march-2020
[Accessed 14 May 2021].

gov.uk, 2021. *COVID-19 Response - Spring 2021 (Summary).* [Online]
Available at: https://www.gov.uk/government/publications/covid-19-response-spring-2021/covid-19-response-spring-2021-summary
[Accessed 26 May 2021].

GOV.UK, 2021. *Face coverings: when to wear one, exemptions, and how to make your own.* [Online]
Available at: https://www.gov.uk/government/publications/face-coverings-when-to-wear-one-and-how-to-make-your-own/face-coverings-when-to-wear-one-and-how-to-make-your-own#:~:text=Face%20coverings%20must%20be%20worn,a%20member%20of%20the%20public.
[Accessed 25 May 2021].

JHU CSSE COVID-19 Data, 2019. *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.* [Online]
Available at: https://github.com/CSSEGISandData/COVID-19
[Accessed 14 May 2020].

lexalytics.com, n.d. *Sentiment Analysis Explained.* [Online]
Available at: https://www.lexalytics.com/technology/sentiment-analysis
[Accessed 24 May 2021].

Lin, Y., 2021. *10 TWITTER STATISTICS EVERY MARKETER SHOULD KNOW IN 2021 [INFOGRAPHIC].* [Online]
Available at: https://www.oberlo.co.uk/blog/twitter-statistics
[Accessed 22 May 2021].

Mohammad Abu Kausar, A. S. M. N., 2021. Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak. *(IJACSA) International Journal of Advanced Computer Science and Applications,* 12(2), pp. 415-422.

MonkeyLearn, 2020. *Sentiment Analysis Applications.* [Online]
Available at: https://monkeylearn.com/blog/sentiment-analysis-applications/
[Accessed 24 May 2021].

NLTK.org, n.d. *NLTK Documentation.* [Online]
Available at: https://www.nltk.org/
[Accessed 24 May 2021].

planspace.org, 2015. *TextBlob Sentiment: Calculating Polarity and Subjectivity.* [Online]
Available at: https://planspace.org/20150607-textblob_sentiment/
[Accessed 24 May 2021].

Pokharel, B. P., 2020. *Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal.* [Online]
Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624719
[Accessed 16 May 2021].

Roul, A., 2021. *Sentiment Analysis- Lexicon Models vs Machine Learning.* [Online]
Available at: https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746
[Accessed 24 May 2021].

Schmidt, C. W., 2012. *Trending now: using social media to predict and track disease outbreaks.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261963/
[Accessed 25 May 2021].

T. Vijay, A. C. B. D. a. P. K., 2020. *Sentiment Analysis on COVID-19 Twitter Data.* [Online]
Available at: https://ieeexplore.ieee.org/document/9358301
[Accessed 16 May 2021].

textblob.readthedocs.io, n.d. *TextBlob: Simplified Text Processing.* [Online]
Available at:
https://textblob.readthedocs.io/en/dev/#:~:text=TextBlob%20is%20a%20Python%20(2,classification%2C%20translation%2C%20and%20more.
[Accessed 24 May 2021].

Tweepy, n.d. *Streaming With Tweepy.* [Online]
Available at: https://docs.tweepy.org/en/v3.10.0/streaming_how_to.html
[Accessed 17 May 2021].

twitter.com, n.d. *About Twitter.* [Online]
Available at: https://about.twitter.com/
[Accessed 22 May 2021].

WHO, 2019. *Timeline: WHO's COVID-19 response.* [Online]
Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#event-0
[Accessed 14 May 2021].