09
11
19'

BDM- DISC 325

PROJECT REPORT

GROUP 2
FARRUKH MASOOD
AREESHA AMIR
KASH KUMAR
SAFIULLAH KHAN

TABLE OF CONTENTS

Dataset description:

The dataset that was presented to us was in an extremely raw format, with 185944 rows and 12854348 columns. The dataset itself contained a total of 61934 number of unique and original tweets on the date Friday 11th October 2019. In the initial period, when our dataset was still in its raw form, each tweet in our dataset contained almost 60 attributes which after filtering down, was reduced to 41 unique attributes allotted to each tweet. These attributes included the time and date of each tweet, the tweet and user ID, the written text of the tweet, the username of the user who tweeted it, the screen name of the user, if the particular tweet is truncated, the tweet ID, number of retweets and number of favorites, the number of followers and friends of the user, date of creation of the account etc. From the filtered 41 attributes from the dataset, we extracted 27 of the most relevant as well as meaningful attributes in order to carry out a further and in-depth analysis on. Upon further investigation, we identified various trends and patterns within the dataset, this led us to conclude that the tweets in the dataset were primarily centered on the upheaval caused in the UK due to Brexit. These will be highlighted and explained later on the report.

Context and Background:

The primary theme of the dataset revolved around Brexit and the turmoil that was caused everything happening in the United Kingdom. The tweets represented the point of views either in favor or against the UK's decision to end their long-withheld ties with the European Union. The data of the tweets was presented between a limited time frame of 2 hours and 30 minutes on 9th October where most of the tweets were being generated from personal accounts as well by many esteemed news agencies from around the world including Reuters top news, The Bloomberg, The Hindu, the guardian and Sky news which clearly reflects the importance of the topic at hand and how it took twitter by storm. The data is directly reflective of the emotions, point of views as well as their sentiments towards Brexit and the daily updates that surfaced regarding the topic in 2019, when the UK was the closest to executing their action of separation.

The update regarding Brexit on the 11th October 2019, included that President Boris Johnson was preparing for the parliament to sit on 19th October to pass the Brexit deal through and to allow the UK to leave the EU on the 31st October (Political Intelligence). Twitter was

responsible for giving birth to the ever so famous “Leave vs stay” argument regarding exiting the Union. This was just one of the events that was happening in Britain which caused the upheaval on twitter. Boris Johnson was the key player in the Brexit movement, along with some important figures whose point of views had an effect on the referendum including the leader of the Labor party Jeremy Corbyn who was anti-Brexit.

Since the UK general elections approached, it can be seen in the dataset that there are quite a few tweets that mention voting. Dominating discussions on twitter were regarding what the outcomes of the elections should be and their effects on the UK if or if not, Brexit occurs. Despite the backlash and efforts of the labor party, the Brexit party was a clear winner of the Elections in December 2019 and on 31st January, after three and a half years of negotiations, the UK became the first country to leave the EU. (SpringerLink)

Methodology:

We performed our entire project solely using Python. Our methodology can simply be divided into three key areas through which we were able to use the raw dataset and through a laborious process, were able to accomplish the standardization of data.

1- Data pre-processing

2- Exploratory Data Analyses

3- Visualizations and Analyses

Data PreProcessing:

The size of the initial, raw and uncleaned dataset was a massive 457 MBs, constituting of 185944 rows and 12854348 columns. The data was incoherent and anomalous in its initial state and contained many inconsistencies and errors. In order to assure that any meaningful analysis could be carried out on our dataset, it was essential to carry out crucial and extensive preprocessing. We made use of the three main stages of data preprocessing for our data cleaning and extraction.

Data transformation:

We started our data transformation process by firstly loading our Dataset into python, moving line by line, reading each line as a string of attributes. The primary rule of action was that we created a list “Remove_from_lines” and stored those characters in this list which were

unnecessary and irrelevant in our string of attributes. The line of code that follows, assists in removing all those characters from our string wherever they might appear and replace it with a blank character.

Secondly, we noticed that in the raw data, there was inconsistency in the number of columns as most of the rows had a different number of columns and the data from one cell was spilled into the next. This irregularity was caused by a property of the CSV file in which the comma (,) is used as a separator and this had led to a creation of a new column with every occurrence of a comma. To rectify this problem, we used a line of code to find and remove all of the commas and substitute them with a blank character. Hence all the attributes were presented as a consistent string in one column per row.

The data was then transformed to categorize the string into unique attributes by using the heading of the attributes and storing all the characters that appear between the headings to be stored into the attribute. For example, the extract shows that the attribute date_time should contain the characters that appear between "created_at" and "id:". The read attributes are then removed from the rest of the string of attributes to avoid any repetitions in the code

```
#Data Transformation
source_file = open("Group_2.csv", "r") #original file
df_to_append = []
#characters to remove from each line
remove_from_lines = ['"','+','0000']
count = 0
for line in source_file:
    line = line.replace(",", " ") #removing commas from csv

    for word in remove_from_lines:
            line = line.replace(word,'') #removing unncessary characters

    date_time = line[line.find("{created_at") + 12 : line.find("id:")]
    index = line.index(date_time)
    line = line [index : :]
```

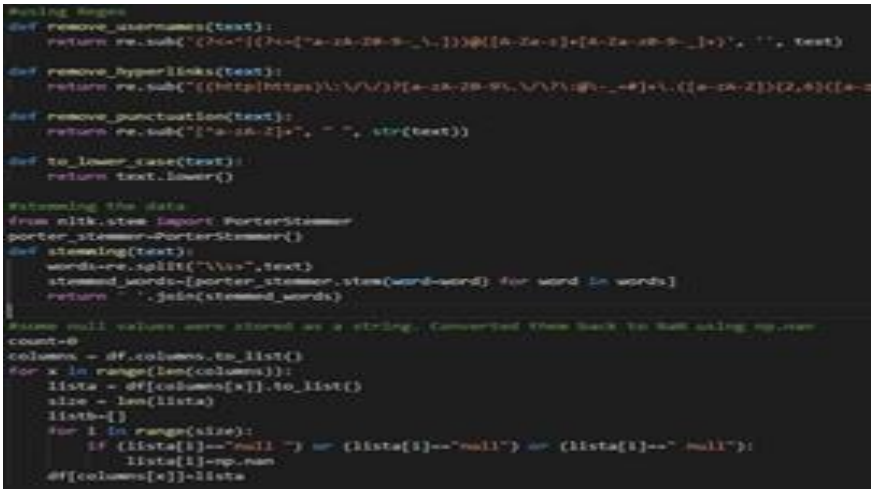
Data cleaning:

Only one in three rows of the dataset actually contained data. Every two rows after a row of data were blank and it was imperative to remove these rows. This was done by writing a python code to eliminate the entire row if it contained only commas or blank characters.

```
# removing 2 lines spacing in the data
for line in cleaned_data:
    if line[0] == "," or line[0] == " ":
        continue
    cleaned_formatted.write(line)
```

We then made use of the Regex function which is a built-in python function to facilitate data cleaning. The 4 codes that can be seen under, namely “remove_username”, “remove_hyperlinks”, “remove_punctutaions” and “to_lower_case” are simply parts of the regex function. The first code “remove_username” is simply used to remove the ‘@’ symbol before the username of the person who tweeted so that we can easily extract just the username of the account. The “remove_hyperlinks” code is used to remove source files and any other irrelevant links which are not necessary for our data or analysis, therefore the code asks for any unnecessary symbols to be removed from the data. The “remove_punctuations” code is used to remove any punctuation marks which does not any value to our analysis, bit instead makes our data over flooded and complicated. The code removes any punctuation mark that is present in the string but keeps those that occur between lowercase or uppercase A-z. Finally, the “to_lower_case” code is used to simply convert all of the uppercase letter into lowercase to make the comparability of data easier. Some users might have used uppercase letters for the same word others have used lowercase letters for. Converting all letters to lowercase will eliminate any discrepancies and will help to get accurate results.

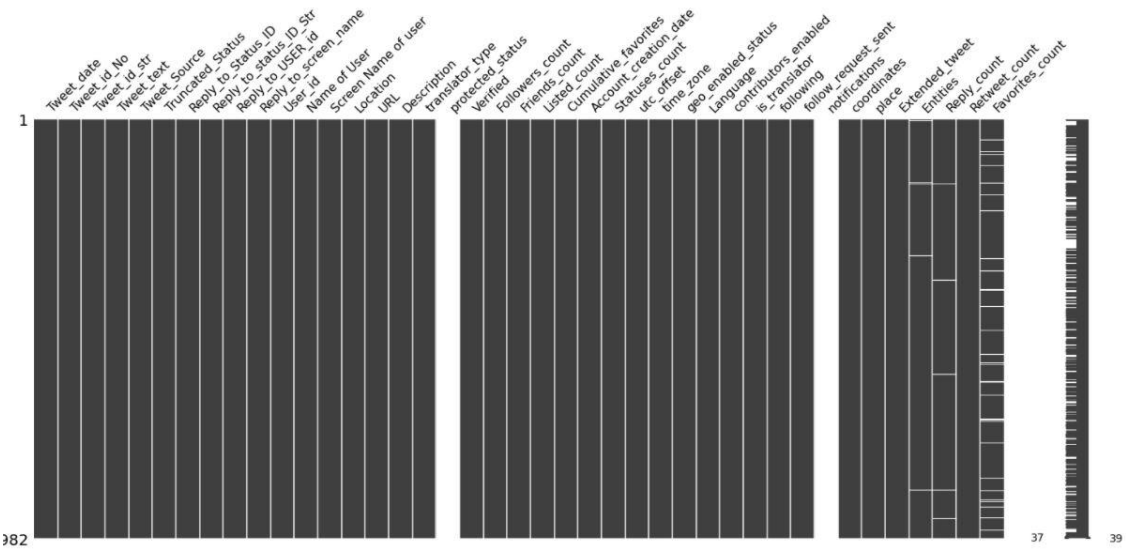
Finally stemming has been carried out to reduce certain words back to their root form, so for example our program does not consider “that” and “that’s” as separate words. This is done primarily to keep to data from being over complicated and so that the final analysis is unaffected.



In the final portion of the data cleaning, it was observed that the null values in some of the attributes were stored as a string and this hinders the final analysis. A python code was then executed to convert those null values back to datatype NaN values. This allows for the exclusion of the null values from our dataset and guarantees a more meaningful and precise analysis.

Data reduction:

The last part of the Data preprocessing is the Data reduction. The following python code includes the names of the 14 attributes which were not retained as the data in these columns contained null values almost 80% of the instances and would have caused discrepancies in the dataset and erroneousness in the analysis. More white space in the following figures show more null values in their respective attribute. There were quite a few attributes which contained mostly null variables which clearly fail to add any values to our data and instead will end up making our end analysis inaccurate and full of misleading values, therefore we decided to drop these attributes and include those that contained relevant information.



Following is an excerpt from the python code we executed to drop the 14 null attributes. The last line of code is used to reduce data redundancy and repetition from the data because the "tweet_id_no" for each tweet is supposed to be unique and this code will assist in verifying that there is no duplication

```
#Dropping columns due to high presence of null_values
#removing Tweet_id_str column due to repitition of same data / reducing data redundancy
# df.isnull().sum()
df = df.drop(columns=['Reply_to_Status_ID', 'Reply_to_status_ID_Str',
                      'Reply_to_USER_id','Reply_to_screen_name','protected_status',
                      'utc_offset','time_zone','Language','following','follow_request_sent',
                      'notifications','coordinates','place','Tweet_id_str'])
df.drop_duplicates(subset ="Tweet_id_No",keep = False, inplace = True)
```

```
[61934 rows x 27 columns]
['Tweet_date', 'Tweet_id_No', 'Tweet_text', 'Tweet_Source', 'Truncated_Status', 'User_id', 'Name of User', 'Screen Name of user', 'Lo
cation', 'URL', 'Description', 'translator_type', 'Verified', 'Followers_count', 'Friends_count', 'Listed_count', 'Cumulative_favorit
es', 'Account_creation_date', 'Statuses_count', 'geo_enabled_status', 'contributors_enabled', 'is_translator', 'Extended_tweet', 'Ent
ities', 'Reply_count', 'Retweet_count', 'Favorites_count']
```

After elimination of 14 attributes from an initial 41 attributes, 27 potentially relevant attributes were extracted as these attributes offered prospective for a comprehensive and accurate breakdown and analysis of the available data. The finalized 27 attributes are:

1. Tweet Date	1. URL	19. Statuses count

2. Tweet ID number	2. Description	20. Geo-enables statuses
3. Tweet text	3. Translator Type	21. Contributors enabled
4. Tweet Source	4. Verified	22. Is Translator
5. Truncated Status	5. Followers count	23. Extended tweet
6. User ID	6. Friends count	24. Entities
7. Name of user	7. Listed count	25. Reply count
8. Screen name of user	8. Cumulative favorites	26. Retweet count
9. Location	9. Account creation date	27. Favorite's count

Finalized dataset:

The dataset that was finalized after a thorough cleaning and preprocessing was reduced -------- and is now a file of ---------MBs, 61934 rows and 27 columns of relevant and value adding data on which we have carried out further analysis. No important data was lost during the preprocessing and hence can be determined that the data preprocessing procedure was successful in eliminating unwanted data and achieved data standardization. The new dataset was stored in "Final Formatted File.xlx." In python, this master dataset was stored in pandas' Data Frame "df" that was unaltered until the end program execution. To conduct any further analysis, we majorly relied on temporary variables to filter out or modify any data.

Exploratory data analysis:

Part 1

After the successful expedition of data cleaning and pre-processing, we finally moved towards the exploratory data analysis of our dataset and for that purpose we instituted some general statistics which can be specified below:

```
Total tweets =  61934
Tweets with URL =  13576
Verified Tweets:  1142  Non-Verified Tweets:  60792
Tweets made by accounts who had greater than or equal to 500 followers:  33500
Tweets made by accounts who had less than 500 followers:  28434
Number of tweets which contain brexit in their text:  54251
Tweets with more than 1000 replies:  3332
```

We used pandas to carry out our data exploration and were able to calculate important statistics for example how many tweets in total were made during the two and a half hours of data provided to us. Moreover, it can be seen that the number of verified tweets is 1142 which are significantly less than the number of non-verified tweets which total to 60792. This means that mostly general public or bot accounts were responsible in generating most of the tweets and giving their own personal viewpoint whether they support Brexit or not. The number of tweets made by people with 500 followers or more is almost 33500 tweets which means that these were users who were significantly active on twitter as you have to be actively posting tweets to gain followers. Out of 61000 total tweets, almost 54251 tweets mentioned the word “Brexit” which indicates that the topic of interest on Twitter was revolving around the Brexit news and activities. Finally, almost 3332 tweets had more than1000 replies which signifies the intensity on which discussions were being carried out on twitter, refuting or supporting opinions of a user in regards to Brexit.

EDA part 2:

For the next part of our exploratory data, we analyzed which accounts that were tweeting during the time frame had the greatest number of followers and therefore are, in correlation, verified accounts. The primary purpose for carrying out this analysis is to judge and decipher the credibility of the tweets as well as to give a rough idea about how authentic the content posted by the account is. As expected, the top accounts with the most followers who are tweeting amongst a political and international topic would be media agencies, news channels and coverage sites.

Name of User	Followers_count
Reuters Top News	20779566
The Guardian	8039751
Bloomberg	5627978
The Hindu	5546778
Sky News	5185423

The account that posted tweets during the timeframe of our dataset included OfCourse London's top news organization Reuters top news, followed by another UK famous, The Guardian. The US's Bloomberg has the third most follower count and The Hindu being the fourth. This analysis helped us to see that not only UK based news organizations are actively participating and are up to date with the activities of Brexit, the reality is that news companies all over the globe are equally as invested to know what the new headlines are and are stand by to broadcast any news on their platform. As we know that during the time frame of our data, news of Brexit being delayed was circulated, news channels clearly took to twitter to be the first to initiate any debates regarding the news and to provide authentic information to the users of twitter,

Part 3:

```
Total Tweets:  61934
Total number of Retweets:  30996674
Total Unique Users:  30537
Number of tweets containing a URL:  13576
Number of Tweets that are replies:  52288
The extracted dataset from twitter comprises of tweets from Fri Oct 11 07:03:12  2019  till Fri Oct 11 09:26:50  2019
```

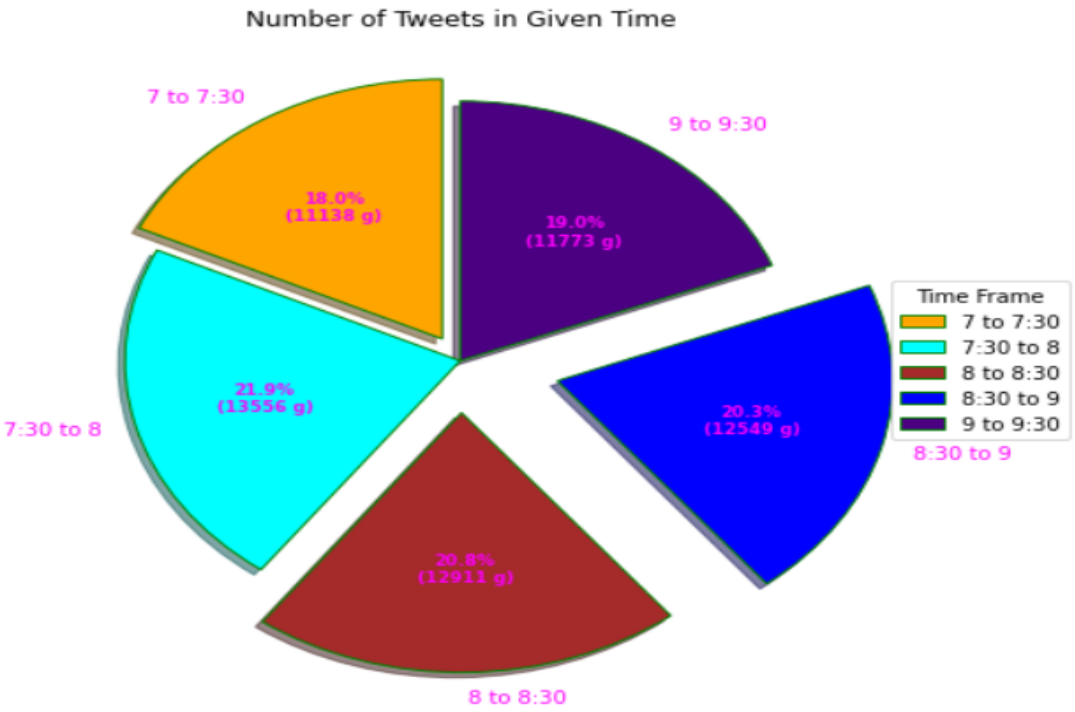
In the final part of our exploratory data analysis, we decided to keep it simple and end it with a few basic yet important statistics. As we can see that the total number of retweets is significantly greater than the number of newly generated tweets. This shows that users were more likely to retweet or share a pint of view or tweet rather than post one themselves. Another explanation for the difference between the tweet and retweet numbers could be due to the bot accounts. This is because Bot accounts are primarily created to boost or promote a tweet or point of view, and

since there are quite a few number of bot accounts in our dataset as we will see later on, it is highly likely that this is what caused the retweet count to be significantly higher.
There were around 30 thousand unique user accounts, this included verified accounts, normal accounts of users as well as bot accounts, all together responsible in creating and retweeting tweets. In the end we simply have calculate the time frame ranging from the first tweet made in our dataset to the time the last tweet was made in our dataset, giving us 2 hours and 30 minutes of data of tweets.

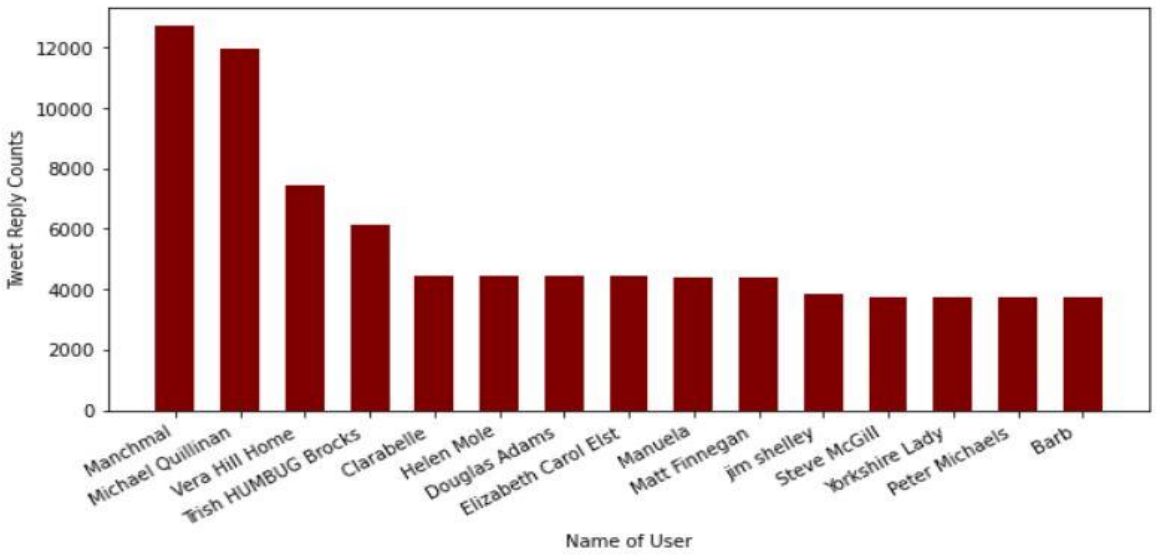
Data Analysis and Visualizations:

Tweet by time analysis:

On 9th October 2019, the court delayed its decision regarding the Benn Act, therefore forcing the Prime Minister Boris Johnson to seek a Brexit extension. Mister Johnson promised the centrist Conservative MPs that we will not cave for a deal against Brexit and that he will not join hands with Nigel Farage, assuring the people in support of Brexit that he will definitely give the deal his all. The delay in the court decision is probably the reason why we see an influx of tweets discussing the future of Brexit.
A parliamentary hearing was carried out in order to discuss the stance and standings of the people on the debate regarding Brexit on 9th October, a few hours before the data recorded in our dataset. Michael Barnier, the chief negotiator for the Brexit spoke his mind by saying that an agreement was not possible at the time. Davis Sassoli however looked at the court's extension to be a way for the British people to express their views (The Guardian), for that reason the public took refuge in twitter to let out their thoughts and blatantly state where their loyalties lie. It is probably due to this state of affairs that we see a divided percentage of tweets within the 2 and a half hour of data that we had. The public was active for the entirety of the 2.5 hours, continuously expressing their opinions and retweeting to support. We can see that there was no one time period during which the number of tweets were extremely high or low, however the number of tweets seem to be consistent in each half an hours' time.



Tweet reply counts:



We plotted maximum number of replies that a tweet got with the names of users as shown in the bar graph below. The user named “Manchmal” got the maximum number of replies **(12,712)** in

this timeframe. We used an SQL query to find the text of that tweet made by Manchmal. The results are highlighted in blue as follows:

tweet_text	Reply_count	Account_Created
RT @PascalLTH: This did not happen. Bercow and Sassoli talked about the importance of both Parlia…	3689	Wed Mar 29 18:00:11 2017
Because everyone needs a break from :;)u00a3@@(/ Brexit!	12712	Wed Mar 29 18:00:11 2017
RT @mpc_1968: 1/ A few clips from this week's @CommonsForeign discussing trade China Brexit an…	2	Wed Mar 29 18:00:11 2017
RT @ottocrat: US regulatory culture. Letu2019s not go down this path Brexit Britain. https://t.co/F…	1	Wed Mar 29 18:00:11 2017

We searched for this tweet on twitter and found the tweet which was made on 11th Oct 2019. The actual tweet is shown below:

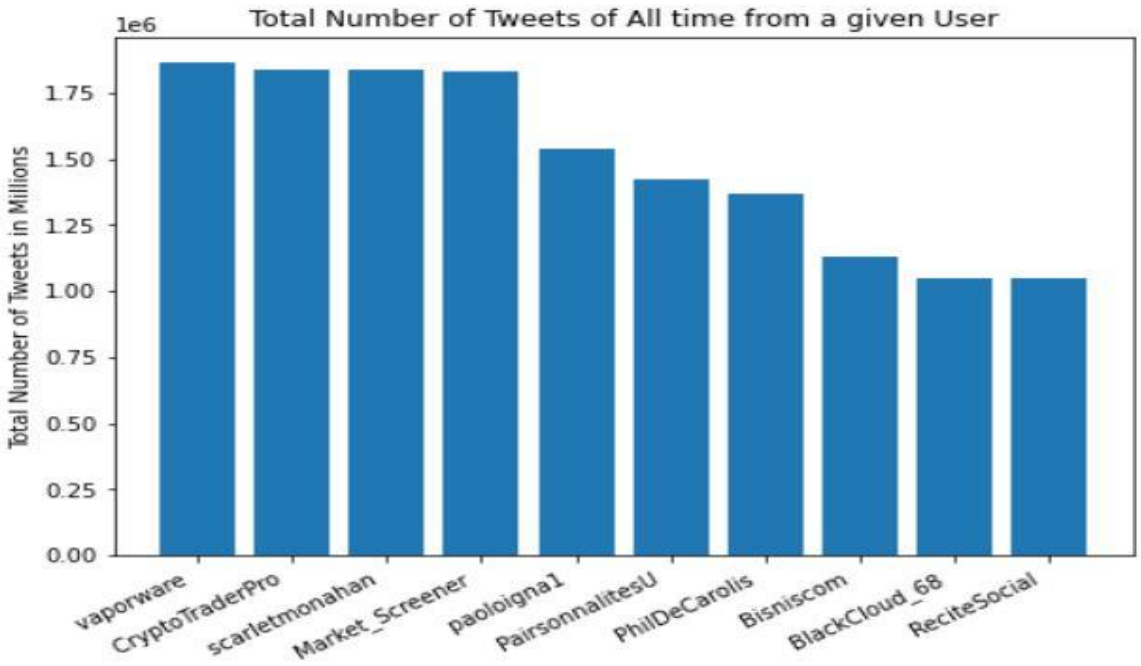
This shows that people at that time were tired of the controversial heate[illegible] and wanted to take a break from it. This is the reason why around 12K p[illegible] hours to this tweet among all other serious tweets made. Michael Quillinan's tweet got the second highest number of replies and its text is shown below:

tweet_text	Reply_count
RT @simoncoveney: Hard to disagree - reflects the frustration across EU and the enormity of whatu2019s at stake for us all. We remain open to fu2026	11988

The tweet was basically a retweet of the Simon Coveney's tweet. Simon Coveney is the Defense Minister of Ireland. He was of the opinion that UK should remain in the European Union as can be seen by his original tweet. People who replied mostly were from UK and replies had a huge difference in the opinions of people whether UK should leave EU or not.

Total Number of tweets all time from a given user:

For a better analysis we also made a graph to see which of the users who tweeted during the specified time frame of our dataset have been the most active been throughout. We carried out this analysis in an attempt to see if the users who are tweeting or retweeting on October 9th are actual authentic accounts or just bot accounts. Hence, we plotted a graph of total tweets of all time for a given user to see which users from our dataset have tweeted the most times. There were a few accounts that had over 1.75 million tweets been posted or retweeted from their account. It was an odd discovery when we saw the number of tweets allotted to one account to be so high, yet there were quite a few accounts whose total tweet count easily goes beyond a million tweets.

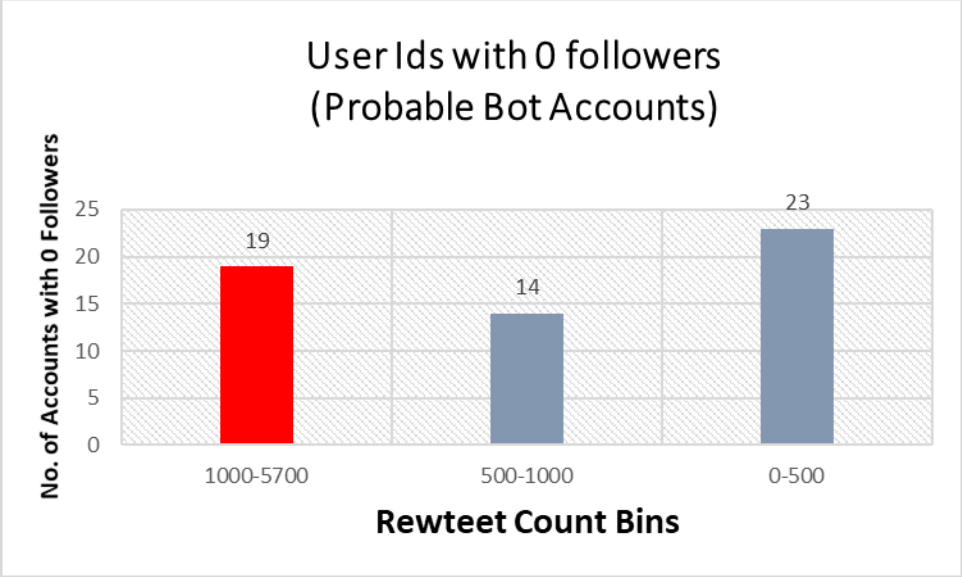


This is when we came to the realization that it might be physically impractical for an account to have made over 1.7 million tweets, unless it was created back in 2010 and has been fairly active ever since. The only other explanation was that these accounts were probably bot accounts who were specifically created for the purpose of re-tweeting millions of times to make a particular

tweet or point of view viral on twitter. We carried out a further analysis on bot accounts to clarify whether these accounts with the extremely high tweet count are regular accounts or bot accounts.

Bot Account Analysis:

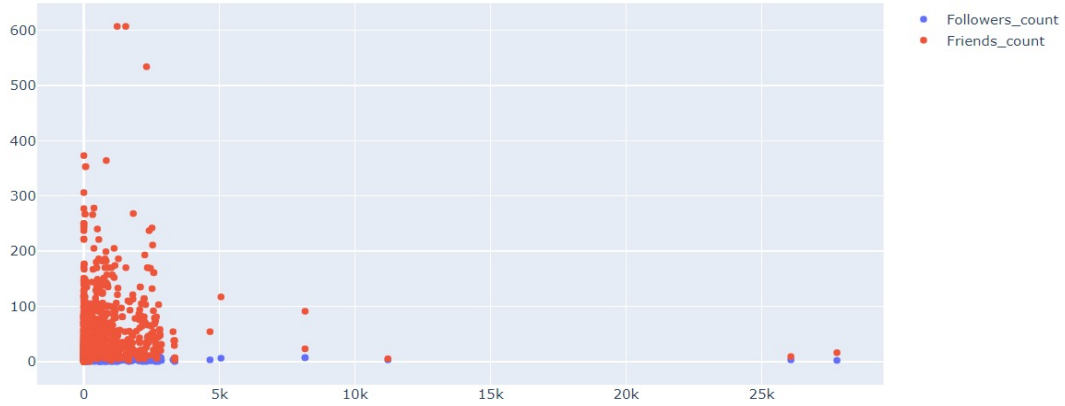
After seeing that there were accounts with more than 1.75 million tweets (as shown by statuses_count), we decided to carry out an analysis to cross-check the existence of bot accounts in our data. The most basic bot analysis we carried out was firstly by finding out the number of accounts with 0 followers and then creating a graph against the number of tweets made by these accounts with zero followers against the three retweet count bins.



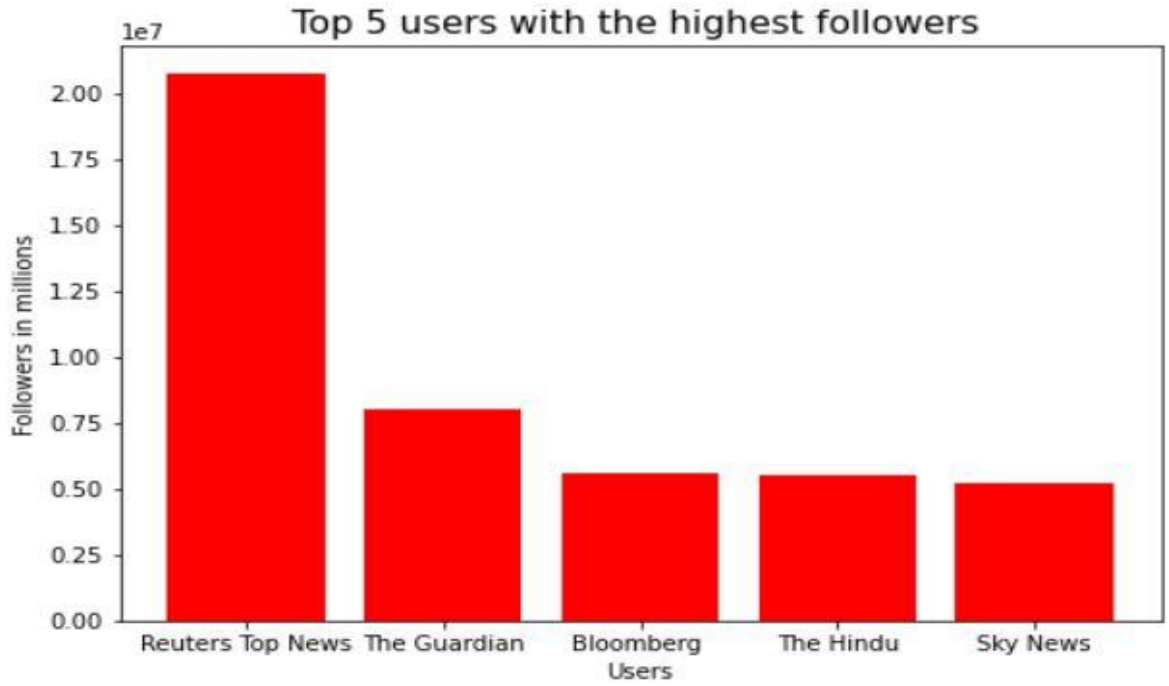
We carried out another visualization and plotted these values on a graph to clearly see the bot accounts and analyze them with respect to their followers_count and friend_count as well. The X axis shows the number of Retweets made by a given user's unique ID and the Y axis plots their respective followers and friends counts to see how many bot accounts there might be in our dataset. The red dots show the friends count and the blue dots show the follower's count. We can see that the blue dots have a few accounts who have 0-5k retweets and only very few who had more than 25k.

From the given scatterplot results, we can observe that there are quite a few accounts, especially on the extreme right side, that have followers and friends count almost equal to 0 or less than 5, but their retweet count was significantly high (>25,000 for 2 accounts, and >5,000 for 6 accounts

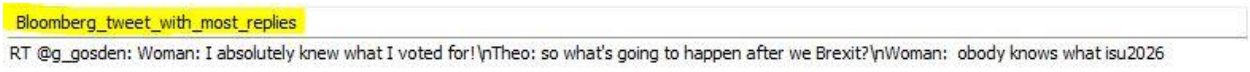
as shown by the graph). Even though there were a lot of accounts with very high followers and friends count who had retweeted between 0-2,500, but a couple of outliers with close to no followers and friends are obvious in the data. Since we cannot be conclusive without proof-verification, we can speculate that these could be potential bot accounts and their purpose could be to promote tweets to support/ go against a certain trend/issue on the twitter.



Top users with highest followers:



To further analyze the users with highest followers, we made a bar graph showing the followers of top five accounts from which the tweets were made. As discussed in the EDA, the accounts with highest followers belonged to the news agencies. Reuter is present on the top of the list with more than 20 million followers. There are American news agencies like CNN, NY Times who have more twitter followers than Reuter. It shows that the big American media and news agencies did not involve much in the Brexit debate. Reuter being one of the biggest European news agency, was on the lead in the followers count as shown in the graph below. Similarly, The Guardian and Sky news being British companies were second and fifth on the list respectively. Bloomberg is the only American Media agency included in top five most followers account and quite surprisingly Bloomberg tweeted **33 times** which shows keen interest of Bloomberg in the Brexit discussion. We dig deeper to find Bloomberg's tweet that got most replies (**256**). The SQL query result is shown:

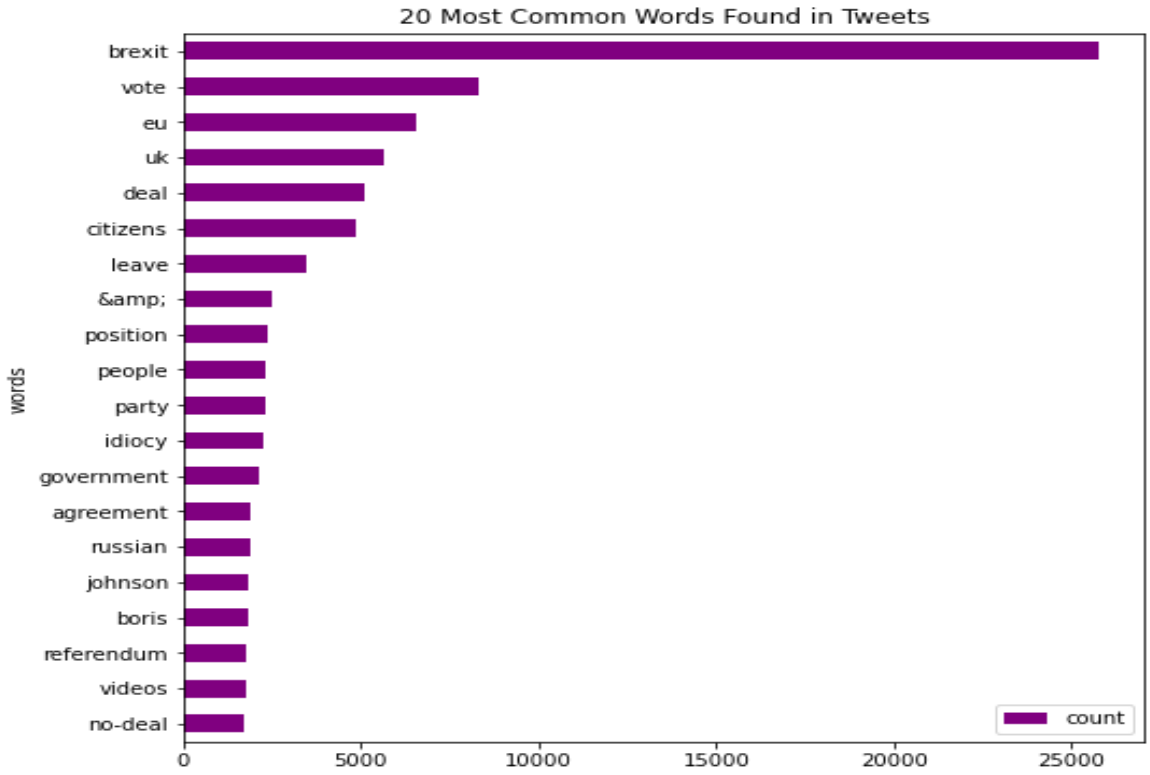


The above tweet was basically a retweet of another tweet that was done by **Simon Gosden. Esq.** on the 11th Oct. The tweet was against the Brexit as can be seen from the picture below:

Most common words found in tweets:

In order to see whether the context of the tweets in our dataset was actually revolving around Brexit, it made sense to extract the most used words in the tweet to get a better understanding of what the tweets are actually about and what is the most common themes amongst them. We made use of Python to create the visualization below to clearly show the most used 20 words throughout the course of the 60000 tweets presented to us. The first graph that we extracted had a

lot of errors as it contained words like ‘the’, ‘it’, ‘https’ to be some of the most commonly used words, however they were of no actual use to us to perform any sort of meaningful analysis, therefore we filtered out those words and created a graph which plots only the relevant words



It was no surprise that Brexit was the most used word across the tweets, with being motioned almost 25000 times which was drastically more than any other word was mentioned. As we can see that almost all of the top used words in one way or another refer to the events of Brexit, involving words like UK, vote, EU etc. The reason behind finding the most common words was to better understand our data and validate that in fact the topic of discussion on the two and a half hours on 9th October was mistakenly about Brexit.

Word Cloud:

To understand the words that were being used throughout the tweets, extracting and analyzing only 20 words seemed less to provide a meaningful conclusion. It is for this purpose, we created a word cloud which is a type o visualization that analyzes the tweets, filters out the most used words and the size of the word is directly corresponding to the number of times that word was

used in the tweets provided in the dataset. Even though we already know that the 20 most common words used are related to Brexit, we wanted to get an even thorough examination of all of the other words that were seen quite a few times so that we can draw any interesting conclusions or findings regarding some other topic or point of view.

As a first impression, we can clear see words like Brexit, UK, Position, deal are the ones with the biggest font, hence the ones that have been used the most. These words yet again are directly related to Brexit in one way or another, however it is unclear that the tweets that mention Brexit

or the UK or vote are done so in a positive context or a negative context. To answer that, we have carried out Sentimental Analysis that has been explained in the later part. We can see other words that are not enlarged but are still essential in the Brexit argument on twitter for example words like “leave”, “October vote”, “Brexit deal” are some of the phrases and words that directly correlate with the Brexit activities on twitter as well.

We can also see the names of some important personalities in the word cloud, this could be that they are either involved in Brexit and are being discussed by the users or are some famous personalities who have made tweets regarding Brexit and therefore have been highlighted or retweeted many times. We can see names like Boris Johnson, Varadkar and Hartley Brewer being mentioned in the cloud, therefore signifying their importance.

We can also see words which on first glance don’t really seem to associate with Brexit but have still been used many times and therefore made an appearance in the word cloud. These include words like “Swedish boyfriend” and “dragon den”. Other words which are used to support the Brexit argument in one way or another like “citizens”, ‘agreement”, “treaty”, “government” and “enforcement officers” are simply just some common words that were used throughout a number of tweets and retweets to convey the emotions of the users regarding Brexit.

Word Count Analysis using My SQL Workbench

We used My SQL Workbench to determine the count of famous words in the tweets' text. Using Word Cloud, we determined the frequently occurring words in the tweets. We used the My SQL queries, including keyword LIKE and the operation %, to locate the specific words in the tweets. A sample query is shown below, which estimates the word "Brexit" in the tweet texts. It is important to note that My SQL is a case insensitive language, so it does not matter if we write Brexit or brexit in the syntax.

```
select count(tweet_id)
as BREXIT_COUNT from d.twt
where
tweet_text like '%Brexit%';
```

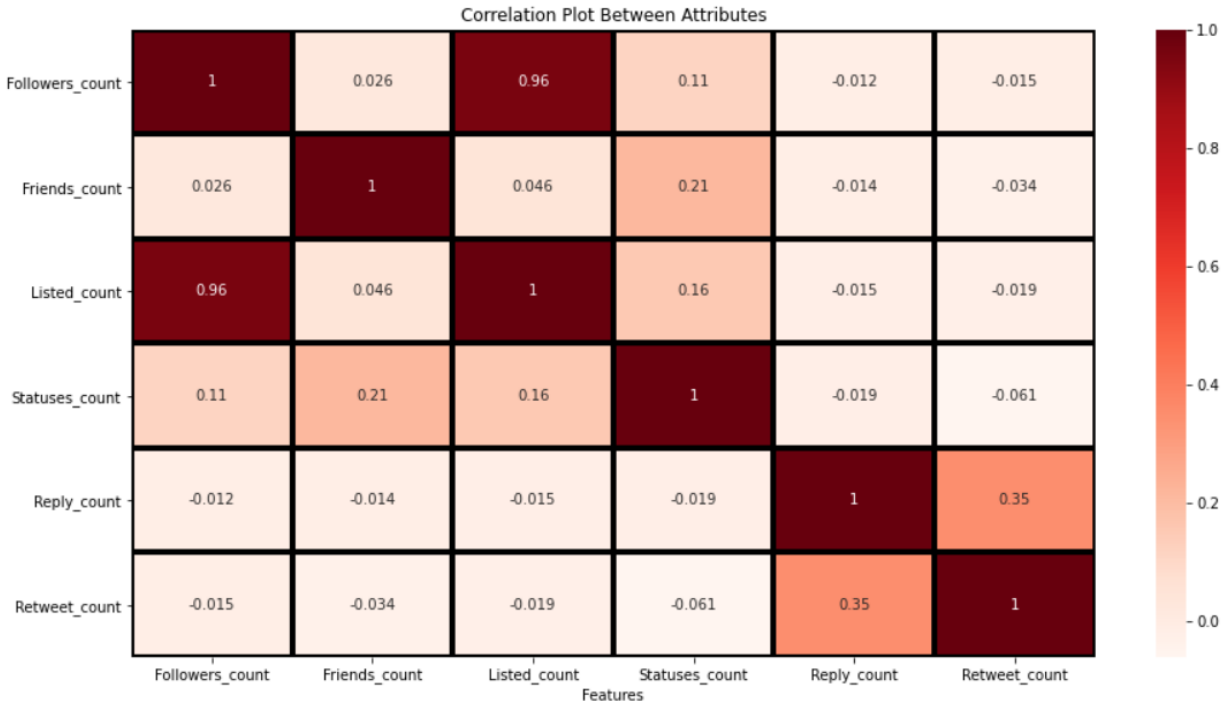
Here, **d** is the database and **twt** is the table where the attributes are stored. The output of queries is shown below:

According to our analysis, the most repeated word in the tweets was "Brexit," and it came in **31400** tweets. It was due to the fact that a heated discussion was taking place globally about this topic. People were using the word Brexit to present their opinion whether UK's decision to leave the European Union would be a good one or not. Moreover, **20143** tweets had the word "European Union or EU" included, and its rationale is self-evident since the UK had to vote either to leave or remain in the European Union. Similarly, the word "UK" appeared in **11324** tweets.

Talking about personalities, **Boris Johnson**, the conservative party leader, was mentioned in **4727** tweets, whereas **Jeremey Corbyn**, the leader of the Labour Party, was discussed in **1133** tweets. The rationale why these politicians were mentioned widely is that the British Elections were quite near, and the winning party would have to decide on the matter of Brexit. Due to the Brexit discussions and the elections going on simultaneously, the words like vote, party, election, win, and lose were frequently used in the tweets.

Correlation Matrix:

Correlation is used to determine the relationship between two attributes of a dataset. It shows the tendency of two variables to move together. If two variables increase together, the correlation is said to be positive and negative otherwise. Correlation of two variables ranges from -1 to 1 where 1 shows the perfect positive correlation. In our model, we have used five variables to make a correlation matrix as shown in the figure below. The relationship strength is shown by the purple-yellow color scale. Red color means perfect correlation whereas white color shows no correlation between the attributes.

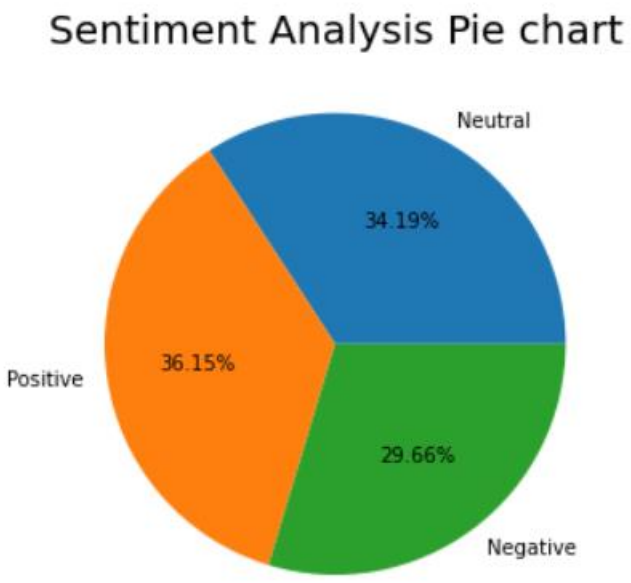


On the diagonal, all the boxes are yellow colored showing perfect correlation. It is because the correlation of an attribute with itself is always perfect (i.e. +1). We can see from the figure that the strongest relationship we have is between Listed_count and Followers_count. It has light yellow color which shows a correlation of around 0.9. The rationale behind is that the people with more followers generally remain more active and have therefore more listed counts. Moreover, a weak relationship is present between the statuses count and the friends count. This shows that on average, the accounts with more followings have more status counts. A plausible explanation is that with time, the followings of an account increases and so does the number of tweets since creation of the account. But the low value of correlation shows that it is not strictly the case. Lastly, there is no correlation of Reply_count with any of the variable as shown by the purple cells. The correlation between friends count and followers count is also zero which potentially shows the existence of bot accounts in data. Since, bot accounts may have thousands of friends but very little followers, the correlation between these two attributes may have reduced significantly by presence of a bot.

Sentiment Analysis:

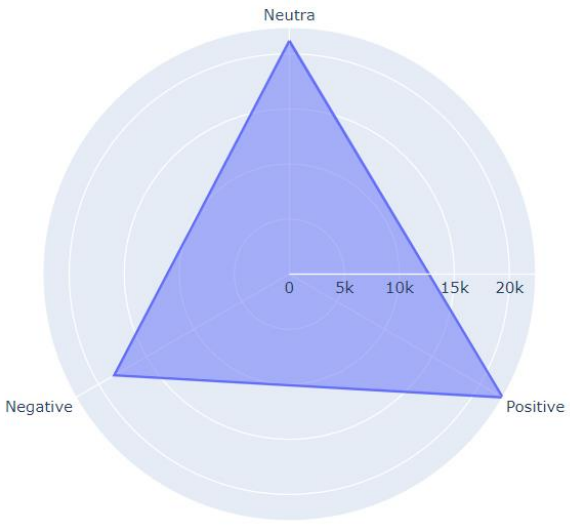
Carrying out a sentiment analysis was extremely essential for our specific data primarily to decipher the sort of emotions that are surrounding the topic at hand. A sentiment analysis is basically "Social Listening" in which social media, twitter in this scenario, is used to dig deep into the conversations of users online to monitor the context of their tweets and whether the emotion that surrounds these tweets ad mentions is positive, negative or neutral. (Sprout social) Since Brexit is a topic regarding which people have varying as well as extreme viewpoints, it was important to carry out a sentiment analysis on the tweets in our dataset to comprehend what the majority sentiment is. With the help of Python Library Text Blob, we set out to find the polarity of the tweets in our dataset. Polarity is measured on a scale of -1 to +1 meaning that a polarity greater than 0 and less than or equal to +1 indicates that the tweet is skewed positively, if the polarity is less than 0 and greater than or equal to -1, it indicates that the tweet is negatively skewed and finally a polarity of 0 simply means that the tweet at hand has neutral sentiments attached to it.

When a sentiment analysis was carried out on our dataset, we found out that 36.15% which totals to 22,388 tweets has a positive sentiment associated with them. This primarily includes all of the users who are tweeting in favor of the UK leaving the EU. 34.19%, which are about 18,370 tweets contained a negative sentiment and included the users who were anti-Brexit and tweeted against the UK separating from the Union. 29.66%, 21176 of the tweets were neutral which could mean that these were users who are either confused or unsure of their stance on Brexit or it can also include users or accounts that are tweeting or retweeting simple facts and information regarding Brexit, without implying any sort of sentiment.



We carried out another form of visualization for the sentiment analysis which is the Radar Chart to offer a more comprehensive and holistic overall assessment through multivariate data. Through the Radar chart, we were able to clearly visualize and compare the current situation that was prevailing during the time frame of our tweets and whether the popular opinion was pro, anti or undecided. The graph below represents the number of tweets associated with each category labelled Positive, Negative and Neutral. Each Bin (small circles inside the circle that label the number of tweets) is made at an increment of 5000 tweets, the further you move from the center to the edge of the circle, the more tweets that category will contain.

We can clearly see that the negative sentiment falls in the 4th bin which means that it has around 15k-20k tweets that represent a negative emotion. Neutral sentiment tweets also fall into the 5th bin but are closer to the bin lower limit of 20k tweets. Positive sentiment tweets fall in the last bin which contains more than 20k tweets and is the closest to the edge of the circle, being the highest number of tweets out of all three. We can therefore safely determine that the majority of the tweets made on 9th October had a positive sentiment and offered a high prospect of being pro-Brexit.



Mapping Analysis:

To better present the number of tweets from each country, we mapped tweets to their creation locations. Most of the users did not disclose their locations, and therefore, tweets from those accounts were not considered in this analysis. To determine the longitude and latitude of each country, we downloaded a file called "LatLong.txt" which contained the Latitude and Longitude

details of all countries in the world. We grouped all the tweets according to their locations and created a list of countries from where the tweets were made. Finally, we found the relevant details of all the countries present in the list from the file "LatLong.txt." Only evident locations having a considerable tweet count were used in this analysis. Due to this, mostly European countries were included in the list of countries. A table representation of this information is shown below:

	Country	Code	Latitude	Longitude	yes/no	Magnitude
8	Republic of Austria	AT	47.33333	13.33333	1	9623
14	Kingdom of Belgium	BE	50.75000	4.50000	1	27
16	Republic of Bulgaria	BG	42.66667	25.25000	1	105
41	Republic of Cyprus	CY	35.00000	33.00000	1	610
42	Czechia	CZ	49.75000	15.00000	1	118
43	Federal Republic of Germany	DE	51.50000	10.50000	1	163
45	Kingdom of Denmark	DK	56.00000	10.00000	1	16
50	Republic of Estonia	EE	59.00000	26.00000	1	13
53	Kingdom of Spain	ES	40.00000	-4.00000	1	172
55	Republic of Finland	FI	64.00000	26.00000	1	63

To map the information, we used the library **plotly** to create a map that shows the number of tweets made from each country. The frequency of tweets is represented by a green color scale, as shown in the map.

EU civil war breaks out as Austria spearheads rebellion against post-Brexit budget hike

AUSTRIA is on a collision course with incoming president of the European Commission Ursula von der Leyen after reiterating its vehement opposition to the EU's insistence that each of the EU27 should pay a minimum annual contribution of 1.1 percent of GDP.

By CIARAN MCGRATH
PUBLISHED: 00:42, Fri, Oct 11, 2019 | UPDATED: 08:14, Fri, Oct 11, 2019

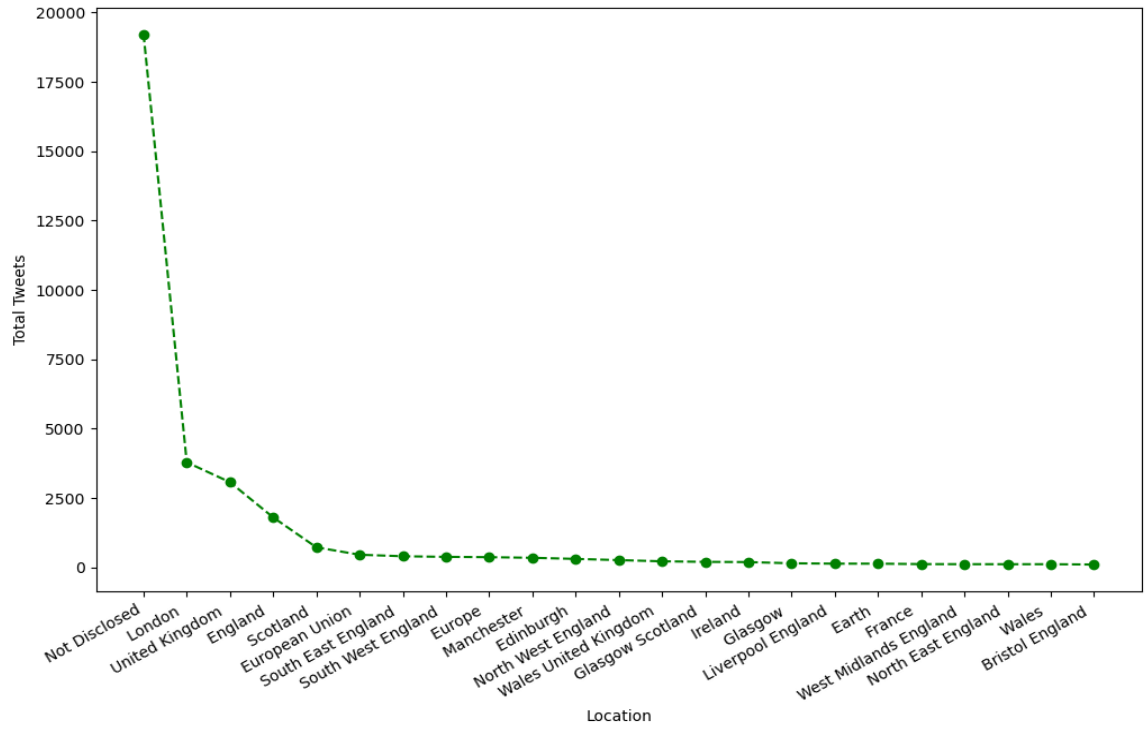
The area highlighted in the dark green color in the United Kingdom. It shows that the most significant number of tweets (**12,241)** were from the UK. It is self-explanatory as Brexit is all about the UK deciding to leave or remain in the European Union. After the UK, Austria is the country with the most number of tweets. It is because of the hike in budget after the Brexit. So, the civil war broke out in the European Union. The finance minister of Austria, Eduard Muller, said in a speech, "EU would need to tighten its belt after Britain, which makes an annual contribution of more than £13billon, leaves the bloc, rather than simply expecting everyone else to pay more." This news was published in the article of Express-News UK. Moreover, **129 tweets** from Austria mentioned this article in their text. The article published on 9th Oct, 2019 and a sample tweet text is shown in the figure below.

RT @Hillchaser: EU civil war breaks out as Austria spearheads rebellion against post-Brexit budget hike https://t.co/zotcC1bQ27

Furthermore, the map in fig 2 shows that most of the tweets were made in the European Countries like Austria, Cyprus, Spain, Germany, etc. We did not include the countries from where the tweets were less than 100. And therefore, we can see that a very negligible amount of tweets were made from other parts of the world like Asia, Australia, Africa, etc. No representation of these areas shows that Non-European people had very little interest in Brexit.

Digging deeper into location Analysis:

Since location is set by the users and Twitter allows multiple ways to set location. One can set location in terms of city, country, continent, or some other place. This was hard to track using a map analysis. So to further analyze the location of the user ids, we made a line graph to count the tweets from each location. Here, the location could be any place set by the user.
Another challenge in doing this location analysis was the **null values** of location. Since some users do not feel comfortable or necessary to disclose their location, so **Nan** was showing in their locations. We converted all the nan values into "**Not Disclosed**" and plotted the following graph:



As can be seen from the graph, many users did not disclose their location. Most of the people belonged to London or UK, which validates our research and the map analysis. While from all other areas, there was roughly the same number of accounts. Still, almost all the disclosed locations were in Europe. This shows the keen interest of Europeans in the Brexit discussion. Moreover, some users set very vague locations like European Union and Earth. Most accounts did not have a location, showing that many of the accounts could be fake accounts, and/or probably people did not want to disclose their identities while discussing a highly controversial topic like Brexit.

Conclusion:

From our tweet data analysis, we found that most of the tweets, mainly from Europe, wanted the United Kingdom to leave the European Union and supported Brexit. It can be indicated by the word count analysis where the word "leave" occurred in **4602** tweets, whereas the word "remain" was used in **3567** tweets. Our research supports the facts in reality since the UK did leave the European Union on 31st January 2020. Furthermore, our analysis shows that non-Europeans had a mixed opinion about Brexit, as shown by their tweets. Also, there was very little response from continents Asia, Africa, Australia, and Antarctica. It shows that people in these areas either had little interest in Brexit or chose to remain silent about it.

Having said all this, there is one significant limitation in our dataset – the small timeframe of the dataset. The Brexit discussion has been spread over the years while our data consisted of only 2.5 hours of tweets, making it hard to believe in the results generated by our analysis. Time is an essential factor, and therefore, our analysis could have been more robust if we used data of multiple years and examine the opinion and discussions over the years. Moreover, to obtain more accurate results about Brexit, we should analyze the data from various sources like news agencies, google searches, articles, and interviews, etc.

Work Cited:

Boris Johnson tells Tory MPs if Brexit delayed he would not fight election on no-deal platform - as it happened | Politics | The Guardian [WWW Document], n.d. URL https://www.theguardian.com/politics/live/2019/oct/09/brexit-latest-news-boris-johnson-plans-emergency-saturday-sitting-of-parliament-after-eu-summit-live-news (accessed 5.3.21).

del Gobbo, E., Fontanella, S., Sarra, A., Fontanella, L., 2020. Emerging Topics in Brexit Debate on Twitter Around the Deadlines. Soc Indic Res. https://doi.org/10.1007/s11205-020-02442-4

London, 2019. Brexit Update 11th October 2019 - Public Affairs, Consultancy and Lobbying. Political Intelligence. URL https://www.political-intelligence.com/brexit-update-11th-october-2019/ (accessed 5.3.21).

The Importance of Social Media Sentiment Analysis | Sprout Social [WWW Document], n.d. URL https://sproutsocial.com/insights/social-media-sentiment-analysis/ (accessed 5.3.21).

The top ten most-followed news accounts on Twitter - Press Gazette [WWW Document], n.d. URL https://www.pressgazette.co.uk/the-top-ten-most-followed-news-accounts-on-twitter/ (accessed 5.3.21).