The usage of Social Media in Observing how Events may Effect the type of Sentiment between Countries

James Caddock

Abstract

Ever since the birth of social media the idea of using this influx of creative, random and intellectual opinions to observe sentiment has grown in the data science community, leading to the possibility of predicting current and future trends in social media. There are many choices of social media but Twitter, a micro-blogging platform, gives us a lot of extra data that can be valuable in our analysis of how events have effected specific sentiment between countries [1]. There are many challenges owing to the magnitude of data, this includes collecting, filtering and observing carrying large time overheads. Furthermore, the demographic of Twitter [2] must be addressed - users have the potential to skew our findings and impact our results drastically. There is not much we can do about the size of our dataset but we can implement time saving functionality and use a smaller dataset for testing. Our method for Sentiment Analysis (SA) involves using a lexicon, a rule-based tool called Valence Aware Dictionary and sEntiment Reasoner (VADER) [3], which is great for short texts, slang and emojis. In this study we aim to find cases of a change in sentiment that one country has about another or even about itself and then to provide a plausible explanation for this shift through events that have happened either locally or globally in a given time frame.

I certify that all material in this dissertation which is not my own work has been identified.

Contents

1 Introduction **3**

2 Summary of Literature Review and Project Specification **5**
2.1 Literature Review Summary 5
2.2 Project Specification 6
2.3 Evaluation and Testing Criteria 7

3 Design **7**
3.1 Twitter API 7
3.2 General Data Protection Regulation 8
3.3 Demographics and Twitter User Base 8
3.4 Anatomy of a Tweet 8
3.5 Keywords 10
3.6 Vader Sentiment 10

4 Implementation **10**
4.1 Data Preprocessing 10
4.2 Data Sorting 11
4.3 Sentiment Analysis 12
4.4 Visualising the Analysis 12

5 Testing and Exploratory Data Analysis **13**
5.1 Testing 13
5.2 Exploratory 14

6 Validating Results **15**

7 Evaluation **16**
7.1 Quality of the data 16
7.2 Increasing the Dataset 16
7.3 Incorporating the Twitter API into the Design 17

8 Critical Assessment **17**

9 Conclusion **18**

1 Introduction

Relationships between countries are usually implied, and viewed as semi stagnant due to there being a lot of history between them and this is what can define that relationship and the sentiment their people feel towards those of other countries. We want to use our analysis of one countries sentiment towards another to see if there has been any change to general sentiment towards one country due to an event and to see if the change is drastic or whether this national identity and shift in sentiment is only revolved around the actual event and afterwards when the trend tapers off whether the sentiment reverts back to it's original value. A country we're particularly interested in is the United Kingdom (UK) as being based in the UK we have a better idea of the events that have happened and of course in our dataset we have the period of time where the UK officially left the European Union (EU) so there is already an event that we can link towards any change in sentiment other countries have about the UK during this time period. We are going to analyse communications between users and to see if we can find any patterns towards the changes or if there's seemingly no reason for the change, whether we can discover it [4].

In addition to looking at sentiment countries have towards the UK, we will be looking at the EU and seeing how different countries feel about the EU, especially after the UK leaving, it may be interesting to discover whether that has had an impact on different countries sentiment towards the EU. It will also be interesting to note whether the UK sentiment towards the EU matches the UK EU referendum vote, although this is unexpected due to the user base of Twitter, it would be a fair assumption to make that the sentiment towards leaving the EU will be negative and the UK's sentiment towards itself may be also negative. However, we must account for the potential of patriotism that arose from the campaign to leave the EU [5].

The aim of Sentiment Analysis (SA) is to evaluate a person's sentiment or opinion towards particular topics or events and to provide a classification of said sentiment which falls into a general positive, negative or neutral category [6,7], some researchers may refer to SA as Opinion mining although others claim that they are not interchangeable, stating that SA is about detecting the sentiment of a piece of text and assigning it a positive or negative tag, whereas Opinion mining focuses on determining whether the text holds an opinion [8]. Therefore we will only refer to it as Sentiment Analysis to avoid any confusion. Sentiment analysis has a lot of applications to social media, where this constant sharing of opinions, it is easy to find lots of different sources and the accessibility of data found on social media is one of it's main boons [9]. As we are looking towards Twitter as our chosen social media, we must mention how we can use SA to interact with Twitter data. Twitter data consists of tweets which are short texts that can include a large variety of metadata which is something that puts it above the other options. A lot of user's on Twitter will be directing their messages towards others which gives SA the ability to determine not only the sentiment of the text but also the target [1].

Social media is a great source of data for many academics, especially those interested in the scale that we find data online. Twitter is an example of one of these that can provide us with connections between people via user mentions, the geo-location of users and the raw thoughts of it's user base confined to 280 characters per post [10]. People use social media as a way to express their emotions and opinions and data scientists use this knowledge to research into topics in order to better understand them and the way these topics are viewed by user's found on social media. As with Twitter, we are sometimes given the location data of a user, we can use this to group people together by country and we can observe specific countries and discover popular talking points within them.

There are many topics in which data scientists are researching through the usage of social media and big data, and with the assistance of metadata like the geographic bounding box of where a user lives, scientists are able to utilise the data helpfully as there have been studies about using Twitter to look for natural disasters like earthquakes [11], floods [12] and heatwaves [13] as people tend to tweet obvious indicators towards one of those three disasters, they also look into political movements like elections [14] and protest studies [15].

In the case of the natural disasters, usage of these digital communications to analyse and detect real world events is known as social sensing. Work can be done to predict any potential disaster usually before the official government has even managed to get their warnings out. In the article by Sakaki et. al. [11] we see that they were able to predict 96% of significant earthquakes in Japan and that their warnings were faster than the Japan Meteorological Agency. Their research was based on classifying the keywords, length and context of tweets allowing them to pinpoint the centre and trajectory of the quakes.

We may find that messages insinuating heatwaves are not as specific as people will most likely only be commenting on how, "hot it is today", although, this in itself is an indicator that the weather is unusual, especially if it is in a country like the United Kingdom where we are not as used to seeing high temperatures daily. This can be used to create word clouds of common keywords used within different countries on hot days and we can measure this to see on which days that are even hotter than normal what words the population may use instead and therefore we can determine a sense of intensity for the heatwave [13].

Researchers use twitter to study political movements [16], which in terms of national politics it usually includes just that country, however, if the topic in question is like the United States (US) presidential elections we may find people from all over the world talking about it as it could have a big effect on neighbouring and friendly nations [14]. In this case if we want to know only what people from the US think about the elections we would use the location data we have to ignore any user who is not from there.

Researchers have also studied Protest studies, for example an article by Neogi et. al. [15] made an analysis of a protest by Indian farmers in which over 40,000 protesters are committed to three acts, the Promotion and Facilitation act, the Empowerment act and the Protect and Amendment act, which are all designed for the betterment of the farmers.

There have also been studies of the COVID-19 virus both during it's initial and concurrent outbreak [17] and during the lockdown periods [18]. Within the former article they looked at specific countries and classified each country into eight different emotions, which allowed them to gather interesting statistics on each countries broad emotions felt towards COVID-19 [17]. However, it is limited due to it's refined selection of countries that do not account for the global issue that COVID-19 was and continued to be. The second paper looks specifically at the first major lockdown with a time period including the dates of the 16th April 2020 and 30th April 2020 due to them being the second and fourth week of the lockdown respectively. They wanted to see the differences between those dates and to see that sentiment is directly related to the pandemic using Twitter to understand and study the psychology of twitter users during this time frame [18].

With Twitter in mind, it is an important note that the user's that are communicating with each other are highly likely to be geographically close as due to the distance, it's very possible that these user's have something similar to bring them together due to how Tweets cluster [19]. It is quite rare for two people who live very far away from each other and have limited contacts to start messaging as people are used to communicating with people who are close to them and they tend to limit themselves to a small group of people with whom that are constantly keeping in touch. This idea of only communicating with people who are similar and that it's the similarity that induces this

connection is called the homophily principle and it's what makes up a person's network of contacts, in which they are limited to their very small and finite social world [20].

Throughout this paper we will make multiple suggestions towards the sentiment that countries exhibit through how they communicate on Twitter, and we will analyse how each country has responded to Brexit and how their sentiment of the UK may have changed during the moment where the UK officially left the EU.

2 Summary of Literature Review and Project Specification

2.1 Literature Review Summary

As we have previously mentioned Twitter is an attractive option for sentiment analysis, much due to it's 500 million tweets that are posted daily [21], where one tweet makes up a single 280 character post on Twitter. There are many topics discussed on twitter and this allows for an extensive pool of data ranging from interesting to irrelevant opinions that we can collect and measure using SA to determine the intensity of the opinion and it's magnitude - referring to the number of people who may share the opinion [22].

Society is in a constant state of change which is influenced by social, political and economic events, we can quantitatively study the communication and sentiment between one countries population to another, noticing whether sentiment fluctuates or if there's a generally accepted static sentiment. By using geo-tagged tweets we can easily gather sentiment from targeted areas all over Europe and we can use this data to build a network that quantifies the sentiment between country to country. We can even look at individual countries and classify various regional identities [23], communities [24], politics [25] and more [26–28].

Natural language is something which has naturally changed in humans over time. Universal grammar highlights that natural languages obtain underlying standards which determine and restrict the framework of grammar. Natural Language Processing (NLP) is a subset of artificial intelligence, that aims to give computers the ability to dissect, understand, interpret and manipulate human language. This is so that computers can derive the meanings from language exactly the same as humans. We can use the sentiment of human language to uncover any underlying emotions that a person may be feeling. This provides a greater detail of context. Researchers have been challenged by judging natural language sentiment for decades. The complexity involved ranges massively dependant on the size of the text. Text with a large amount of content such as product reviews are known to be simply classified into positive or negative reviews [29]. However, Twitter sentiment analysis is more complex due to the consistent use of abbreviations and slang.

Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text [30]. With this approach a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Different approaches to creating dictionaries have been proposed, including manual [31] and automatic [32] approaches. Generally speaking, in lexicon-based approaches a piece of text message is represented as a bag of words. Following this representation of the message, sentiment values from the dictionary are assigned to all positive and negative words or phrases within the message. A combining function, such as sum or average, is applied in order to make the final prediction regarding the overall sentiment for the message. Apart from a sentiment value, the aspect of the local context of a word is usually taken into consideration, such as negation or intensification [33].

2.2 Project Specification

A more clear and revised aim of this project is to uncover the sentiment that countries have about their neighbours and whether this is and has been effected by events in the world. We want to know if there are positive or negatives trends in sentiment and we will evaluate each countries baseline sentiment in order to justify that their feelings towards another country is unusual compared to their overall sentiment observed. For example we may see that the majority of Finnish tweets are positive, an average of 0.6 where the scale is from -1 to 1, from negative to positive respectively. Therefore their baseline sentiment is quite positive and if we looked at their sentiment towards Sweden and we observe that the average sentiment is higher than the baseline, then we can assume that the Finnish like the Swedes more. A clarification of this project is that the aim is not to create a sentiment analyser but to use VADER, the lexicon-based analyser tool, to analyse, compare and evaluate the sentiment of tweets in order to discern patterns and cause towards pronounced shifts in feelings from one country to another.

The dataset that we use has been collected by SEDA Lab [34], and we have been provided access to this data thanks to Rudy Arthur, one of their members, SEDA Lab is a group of researchers based at the University of Exeter with research applied towards social media. Rudy granted me access to their server which contains hundreds of millions of tweets collected from Twitter, and is actively collecting tweets using the Twitter API to date. Within their collection there are tweets that are geo-tagged from all over Europe which we use for this project. In addition the specific data that we're looking for are tweets that talk about either the UK or the EU and it is important that the tweets used have user mentions as this implies that there is an ongoing conversation and can help to direct sentiments from one country to another and find patterns. User mentions is an attribute within the metadata of a tweet that refers to any user's id and screen name if they appear within the text of the tweet, which is denoted by the @ sign. We go into further detail about user mentions later but it is vital to note that user mentions is a list that can contain multiple users and it is how we create links between user's within our network of profiles.

There are a three main processes that we need and they will each require different resources to create them, the processes are all in some way related to the data that we need and they are; data acquisition, cleaning and analysis. Due to much easier access to the data, acquisition is going to be much simpler, we don't need to communicate with Twitter in any way or request an API key or anything, all the data we need is stored on a server which we can easily ssh into. Now we still need to sift through and relocate all of the relevant data that we actually want, and that is what my data acquisition process will do. we will need to unzip every file that we would like to have the data of and as each file is conveniently named after the date in which it was collected that will be an easy task. The tweets are in the json format and we know that the orjson library is a great library which is pretty fast so we will use that. Using orjson we will create a python program that unzips every file that is between a desired date and then read through the jsons compiling every tweet that includes any keyword that could be referencing the UK or the EU and any tweets replying to tweets that include those keywords. we will make sure to extract the relevant data into new jsons with similar but different names. The relevant data includes; the user screen name, the tweet id, the tweet text, the country, tweet date, the in replies to user's tweet id and any screen name's of mentions. As we don't want to create a folder full of two times the amount of files that we unzipped with a much bigger size, we will unzip, read the data, copy the relevant data then delete each json until we have a folder with just the data that we want. Although most likely unnecessary we will make the program fairly configurable, adding parameters for; a maximum number of files (int), date start (str), date end (str), use exact location (bool), keywords (list). In addition as it's safe to assume that many if not most of the tweets will be in varying languages we am also going to include additional keywords that may be slightly different in other languages, we will do this by using the python library googletrans when first starting the program to convert every keyword into every possible language that may be used in the tweets.

2.3 Evaluation and Testing Criteria

According to our revision of this project our testing criteria has changed. We want to use our dataset that we acquired from SEDA Lab to observe each countries sentiment over time and to note any changes that may occur. With those changes in mind we want to provide reasoning and explain why there has been a change in the sentiment. The events that we're going to focus on has been reliant on what data we can extract, so we decided to look at Brexit and the affects that's had on other countries sentiment towards the UK, as well as looking at the word itself to see what kind of sentiment is used when countries are conversing about Brexit. We also want to keep in mind the baseline sentiment of each country to help prove that in the case of our dataset that the chosen country has sentiment towards the UK or Brexit that is vastly different to their sentiment directed elsewhere.

In order to have success in this project we need to be able to create graphs that can illustrate what the sentiment towards the UK is from countries like Germany, France and Spain. Although, we expect results from most countries in Europe we can not know if we will have enough tweets from every country to veritably confirm their sentiment which is why we only require these graphs from a few notable countries who will have a lot more communication with the UK. We also expect to have a graph of the baseline sentiment for each country, particularly those that have a lot of connections to the UK and finally we would like to have data and display the frequency and sentiment of each country towards the keyword Brexit.

3 Design

3.1 Twitter API

Twitter is a social media micro-blogging platform, where users produce character restricted messages called tweets and disclose their thoughts and activities in real time. Tweets are stored in JSON format and can be collected for free in real-time or purchased retrospectively from Twitter [35]. Access to historic data can also be granted through the use of Twitters developer program, in which a user can hook onto Twitter's API via an endpoint [10].

The Twitter API has a lot of functions that can be easily used to collect very specific data, once you have been provided permission by them and if you would like to collect data as it's being produced on Twitter, then you can set up a system that'll allow you to receive whatever content and metadata you want from Twitter. However, there are certain terms and conditions that you must comply with whether you're collecting the data yourself or acquiring it elsewhere, you must follow them and even though we didn't collect the data from Twitter ourselves, it was still sourced from Twitter so we made sure that our analysis and investigation was not in breach of Twitter's developer agreement and policy [36].

There are other restrictions to the API that include a finite Tweet cap which is different depending on what level of access you have, ranging from 500K Tweets per month to 10 million Tweets per month, there are limits to the amount of data you can request from the tweet and you are even restricted to a character limit for your queries. You may only make a certain amount of requests per 15 minutes and that is also dependant on your level of access, as well as the ability to access the full-archive search of historic tweets which is reserved for only academic research as is most high intensity usage of Twitter's API.

3.2 General Data Protection Regulation

Big data has become a great asset for many organisations, promising improved operations and new business opportunities. However, big data has increased access to sensitive information that when processed can directly jeopardise the privacy of individuals and violate data protection laws. As a consequence, data controllers and data processors may be imposed tough penalties for non-compliance that can result even to bankruptcy. In this paper, we discuss the current state of the legal regulations and analyse different data protection and privacy-preserving techniques in the context of big data analysis. In addition, we present and analyse two real-life research projects as case studies dealing with sensitive data and actions for complying with the data regulation laws. We show which types of information might become a privacy risk, the employed privacy-preserving techniques in accordance with the legal requirements, and the influence of these techniques on the data processing phase and the research results [37].

3.3 Demographics and Twitter User Base

Ideally, when comparing the Twitter population to society as a whole, we would like to compare properties including socio-economic status, education level, and type of employment. However, we are restricted to only using the data that is (optionally) self-reported and made visible by the Twitter users, including their name, location, and the text of their tweets. Overall, we found that Twitter users significantly over represent the densely population regions of the U.S., are predominantly male, and represent a highly non-random sample of the overall race/ethnicity distribution [2].

Along with Twitter being mostly male dominated it is also dominated by the age group of 25-34, with 38.5% of users in that group. Then the second largest group is 35-49 year old's at 20.7%, ages 18-24 and 50+ are tied with 17.1% and the last 6.6% of users are between the ages of 13-17 [38]. What these statistics tell us is that not every person is represented in Twitter and we have a severe lack of people both young and old, although, it is common for people to hold a lower value towards children's opinion as it is believed that they have not lived long enough to truly understand what they're talking about. We don't even have anyone under the age of 13 and that is not a problem because as we mentioned above, we do not hold much value towards their sentiment and feeling and also their opinions will have a lack of impact as their beliefs and opinions at such a young age are likely to be constantly changing as they're finding themselves, which would make for unreliable data [39]. Furthermore, although, we do have a decent percentage of user's over the age of 50, we can probably assume that most of these people are in their 50s, and as the group is much bigger than the 18-24 group we know that technically for the worldwide demographics to be met we would need many more users over the age of 50.

As you can see from Figure 1, we have a lot more tweets from user's living in X than those from Y and Z, which aligns with statistics of Twitter's user base [40]. This is one of the biases of our data but it means that we just have more conclusive data from countries like X where we have a lot more tweets available and can more accurately measure it as a whole. We do not have a solid solution to this bias but we can keep in mind that not everyone from each country is represented and even though the United Kingdom has a lot more users than France we still have the same issue that it does not make up every single person in those countries.

3.4 Anatomy of a Tweet

As mentioned above Tweets are stored in JSON format and have many attributes defining specific things about the contents of the tweet, metadata about the user who sent the tweet and any information about user's who may have been mentioned in the text of the tweet. There is a lot of available data to be gleaned from a tweet but we only need a few attributes in order to successfully dissect them. The 'place' attribute is used by Twitter to designate the location in which each tweet is sent and this is something that has been defined by the user. We use this to extract where a user is from and it is

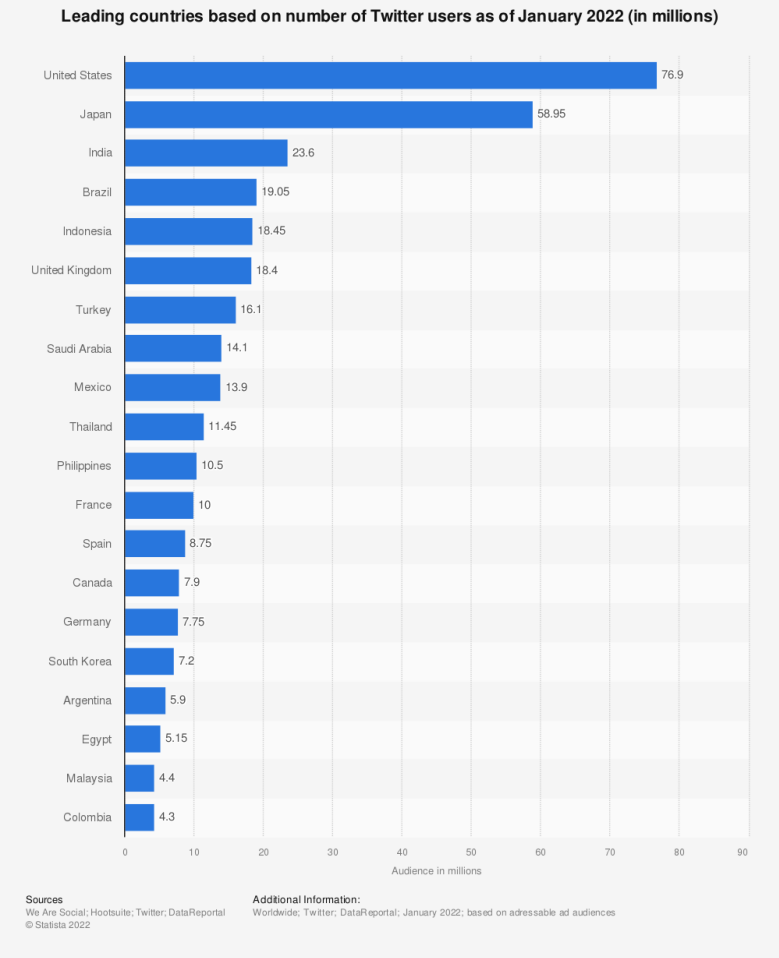

Figure 1: Graph of most common countries who use Twitter [40]

required in order to place a person in the world. As we also want to see interactions between users another important attribute is the 'entities' attribute which holds a list of 'user_mentions' that gives us information about each user mentioned with an '@' in each tweet.

We chose mentions over replies and follower count as that implies a direct connection to each user, whereas a follower may never actually interact with the user and there are a few different possibilities that can invalidate the meaning of a follower. One of these possibilities is that users can purchase followers that are just bot accounts and are not real people, this connection is meaningless as we want real conversations with real people. Another possibility is that a follower may not even interact with the user which would limit their connection to the user to a number which does not provide any useful data except the implication that they like the user, although there's a chance a follower may detest the user and they follow only to fuel their hatred. This is what makes user mentions an amicable choice due to it being a user directing their tweet towards another user which we can draw sentiment from and easily link the tweet as a connection going from the user who sent the tweet A, to the recipient user B.

User's on twitter can decide to display a geo tag on each of their tweets, which will be added to the metadata in the tweet under the "place" attribute. This can include specific data such as the exact longitude and latitude of the location the user sent the tweet to general data like the country code of the country sent. We are only interested in the country code as we want to know what country a user is from and we do not need the exact coordinates of the user as we trust that the country code found is correct and this allows us to sort users in a simpler manner as we can just compare country codes and compile users into group by matching their codes. Unfortunately not all tweets have geo tags, which is unavoidable since there is no obligation on the user to display their location and it would be a

security breach if this was a forced or hidden component. Tweets without country codes are discarded as we can not use them in any way due to the nature of this project. Our approach does have some biases, as by concluding that the place where a user tweets is where they live assumes that the user's geo-tagged tweets were sent from their home which could not always be the case. We attempt to overcome this problem by making a further assumption that the majority of a user's tweets are sent from home as we look at each tweets made by a user and decide that the location they tweet the most from is their home country. Our limitations for this correction lies in how many tweets we have for a user and how many different locations from where they tweet.

3.5 Keywords

We use keywords to sort through the contents of each tweet in the dataset and only continue processing the tweet if one of the keywords are found within the text. This makes defining keywords a good way to focus onto specific topics and divide tweets into sections that allow for further analysis into a chosen topic. For example adding keywords like, 'sport', 'game' and 'player' will get tweets which are likely to be focused on sport. This lets us sort through the dataset quickly as we no longer need to waste time processing a tweet once we discover that it isn't relevant to our desired topic.

The keywords we need are any words referring to the UK, the EU and Brexit. Using these will streamline the dataset, making sure that every tweet received is relevant to the topic we're measuring. With tweets gathered in this filtered dataset we can know for certain that each tweet is talking about these topics that we want and it enables us to make statements about these topics whilst analysing the sentiment discovered in them. Over the years of social media we have developed a way to inform people what the content of our message is, this method is a simply by expressing the topic or opinion after a hashtag, for example #Brexit. this hashtag lets everyone know that this message was about brexit, so in more obscure tweets where a user is replying to another user or something similar and they refer to the topic as 'it', if they also place a "#Brexit" at the end of their message then we'll instantly know that 'it' is brexit and that whatever sentiment we can glean from their text is about brexit.

3.6 Vader Sentiment

We decide to use Vader Sentiment as the sentiment analyser, Vader is a rule-based and lexicon tool used for sentiment analysis where it's strengths lie in sentiments expressed on social media where most messages are short [3]. Vader uses a combination of qualitative and quantitative methods to produce, and then empirically validate, a gold-standard sentiment lexicon which makes it great for the short form tweets found on Twitter. Vader has many features, such is it's ability to process emojis allowing it to detect sentiment through a vast number of them like the laughing emoji which may indicate that an otherwise negative message is a joke. Another useful feature is that Vader can analyse non-English texts with it's usage of My Memory Net [41] which translates the message for Vader, allowing it to conduct it's analysis on an otherwise foreign language.

4 Implementation

4.1 Data Preprocessing

We need to be sure that each tweet from our dataset includes all the attributes that are required so that we can create a profile of each user and successfully place them into a country. The attributes that we are looking for are the "country_code" (the Alpha-2 country code), "user_mentions", "text" (the content of the tweet), "user", "id" (the tweet's unique identifier), "lang" (the language of the text) and "created_at" (exact time of creation). The country code is found in the "place" attribute and user mentions in the "entities" attribute. User mentions is a list of each user that is mentioned in the tweet and we want all of the mentioned user's and their "id" (the user's unique identifier) and "screen_name" (the user's non-static display name) attributes. Finally user is similar to user mentions

except it is not a list and it holds the metadata of the owner of the tweet, similar to user mentions we desire the user's "id" and "screen_name" attributes.

Once we decide what attributes we need we filter through the data and create profiles of each user that has tweets containing these desired attributes. Each tweet is located in a zip file of a JSON file that contains a collection of tweets that were posted on a given hour of a given day, for example tweets from 3pm on the 19th January 2020 will be located in a JSON named geoEurope_2020011915 inside a zip file of the same name, where the format is geoEurope_YYYYMMDDHH. Each of these zip files were located in a single folder which contained our entire dataset before it was processed. As we process each file we add all the valid tweets to a new JSON of the same in a different directory, this allows us to see which json we have completed and if we run into any issues the program can easily pick up from where it stopped as we have a directory full of each processed json.

As we're creating the new processed JSON files, we also create profiles of each user where we add their tweets to a JSON file named after their "id" attribute in a folder called, "profiles". These profiles are structured as follows. Firstly we place the "user" attribute with it's respective "id" and "screen_name" attributes inside. Then we have a "tweets" attribute which is a list that holds every tweet that we have from the user. Finally, we have the "mentions" attribute which is like "tweets" except it holds the "id" and "screen_name" of every user mentioned in the "tweets" attribute.

We can also use keywords to search for tweets that only contain those words, this helps to speed up the process as then we are writing to less files which means we're spending less time on tweets that are not relevant to our chosen topic. The way we implemented keywords was by having a dictionary of keywords where the keys are named after the language that the keywords are written in and the values are lists of the keywords in the keys language. We do this so that whenever we have a tweet that isn't in English, we don't need to translate it as we instead, whenever a new language appear, add a new entry to the dictionary which are the keywords translated into the new language. This saves time as instead of translated for N times, where N is the number of tweets, we're just translating anywhere between 24 to 200, which is unquestionably less than the number of tweets as a single JSON file of an hour's worth of tweets easily reaches the tens of thousands.

4.2 Data Sorting

After the data preprocessing we are left with a profiles folder that is full of unsorted profiles that are hard to distinguish and difficult to manage, which is why we begin by distinguishing each profile more throughout this sorting process. We start by looping through each profile in the folder, we first do a little sanitation by verifying that each tweet in the profile is unique and we remove any duplicates we find. We know whether two tweets are the same as they will have the same "id" attribute due to each "id" being completely unique to each tweet. We also add the "country_code" of each tweet to a dictionary and set it's value to one or if it already exists within the dictionary we increase the value by one. After we have looped through each tweet we determine what country the user is from by which country their tweets are most frequently located. We add a new "place" attribute to the profile containing the country code of their designated country and if it does not exist we create a new folder which is named after the user's country code and we move the profile into this folder. The reason we decided to set a user's country to the place that they've tweeted from most is because it is highly likely that a user will be tweeting mostly from their own country.

4.3 Sentiment Analysis

Now that we have our profiles from our dataset sorted into different folders based on location we can start our analysis of them. Before using VADER to calculate the sentiment of each tweet we want to add every profile to a dictionary, checked_profile with the value of their country code so that we can keep track of all the profiles and start to make connections between users. This allows us to then when we go through each user again, to check their user mentions and if they mention somebody in checked_profile then we know we have a match and we can see that this interaction is between two users that we have in our dataset allowing us to then look at their country of origin and put the sentiment towards the specific average of that country. For example if user A is from the France (FR) and user B is from United Kingdom (GB) we add this to the collection of sentiment from FR to GB. These connections are created in the next part of our analysis, where we loop through every tweet a profile has made and check to see if they mention any user's profiles that we have collected. Once we have calculated the user that the tweet is directed to, we can then use the checked_profile dictionary we created to know what two countries in this instance are communicating with each other. If they are different countries then, after calculating the sentiment with VADER, on a scale of -1 to 1 going from negative to positive where 0 is neutral, we compare the two country codes and if they're different then we add them to a new dictionary, sentiment_over, that is for our analysis of one countries sentiments towards another. In sentiment_over, the keys are formatted in the style of country A sentiment towards country B like "FR over GB", this key is created by concatenating the country code of country A, "over" and country B, where in this case country A is France and country B is Great Britain. We can then access this easily by looking for keys with country A + " over " + country B, allowing us to access each relationship we want to view.

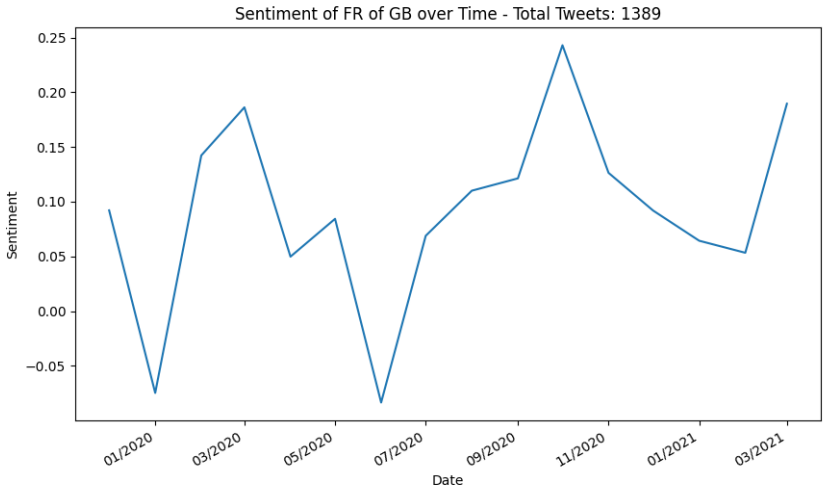

Figure 2: Sentiment of France towards the United Kingdom

4.4 Visualising the Analysis

Figure 2 above is an example of a graph that we make within this section, the graph is showing the number of tweets involved, the change in sentiment over time starting from December 2019 and ending in December 2020, and the averages of the monthly sentiment. We create and output varying graphs and plots that show data in a clean and pleasing format. We first look to make a plot of the change in sentiment over the dataset between each country over country by enumerating over the dictionary and using the key as the title of the graph.

We also use the dictionary that just uses every countries respective tweets to create plots of the countries with the most tweets to looks at their baseline sentiment. To calculate the baseline sentiment for a country, we would take the dictionary holding all of the sentiment for all of their tweets and then we work out the mean of that sentiment and we use the standard deviation to also create some error data that we also use in the graph for our baseline sentiments to show the margin of error and to further explore our data.

Another load of graphs that we create are the keyword graphs which take every tweet by every country that mentions the keyword, one of our keywords was Brexit, and then we individually calculate each countries average sentiment toward that word. This process is much like the baseline sentiment process except it does not use an error bar. We had further plans for graphs but we decided that it did not fit the scope of this project but these ideas are something that we would keep in mind if we were to continue with this project.

5 Testing and Exploratory Data Analysis

5.1 Testing

The dates we chose for the dataset was from December 2019 to December 2020, this time period includes the start and effects that the COVID-19 pandemic had on the world [42]. COVID-19 had a disastrous effect and we have been suffering from it even today, through Twitter we can see how much of the discussion in this time was about covid and if the virus may have brought people together or further apart by using sentiment analysis and comparing how it changes throughout the dataset. In the dataset we also have the British Exit (Brexit), which signified the United Kingdom's (UK) withdrawal from the European Union on January 2020 from the outcome of the 2016 UK EU membership referendum [43].

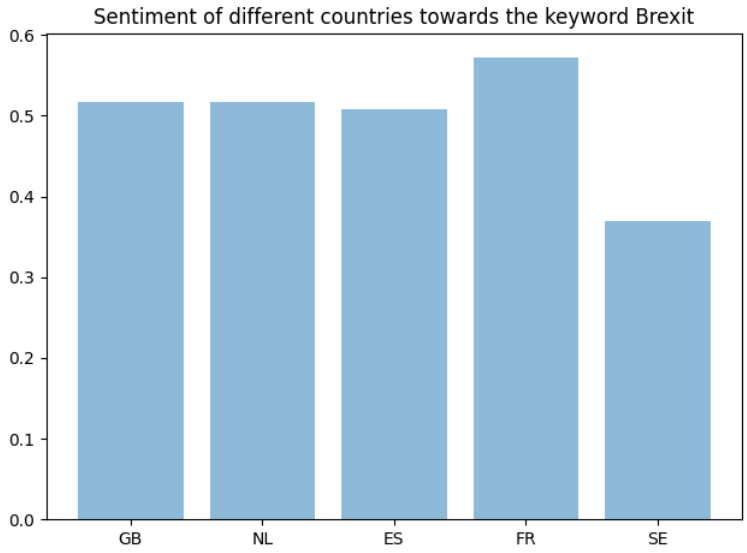

Figure 3: The average sentiment of the word Brexit for the UK, Netherlands, Spain, France and Sweden

There was a potential that even though the data collected was supposed to include only tweets with geo-tags that some tweets may not have these tags and therefore would not provide any use in our research, therefore we made certain that each tweet contains all the pieces of data that we require. This is achieved through a format JSON file that mirrors the same format to tweets except we only include attributes that we want from the tweet and we set the attribute equal to True or False, if an attribute is true then that means that the tweet must include this attribute and should be discarded otherwise, if it is false then that means we would like this attribute but if the tweet doesn't include

it we still want the tweet and if the attribute is not in the JSON we simply do not care about it. For example if we want the text, id and user's id from each tweet but only require the text from the tweet, we would create a format JSON that contains the "id" attribute and "text" attribute set to False and True respectively, as the user id attribute is found in the "user" attribute under "id", we would need to say that the "user" attribute has the "id" attribute that is equal to False.

We noticed that during the period of global lockdown that there were vastly more tweets on those days and that the sentiment found during this time was much more volatile. We believe this is a cause of the extreme effect that the pandemic had on the world and that psychologically it broke a lot of people, so we can imagine why the extremities came to fruition during this time period. However, we did see many positive sentiments come from this period and we believe that may be those wishing others well and hoping for the best in a tough time.

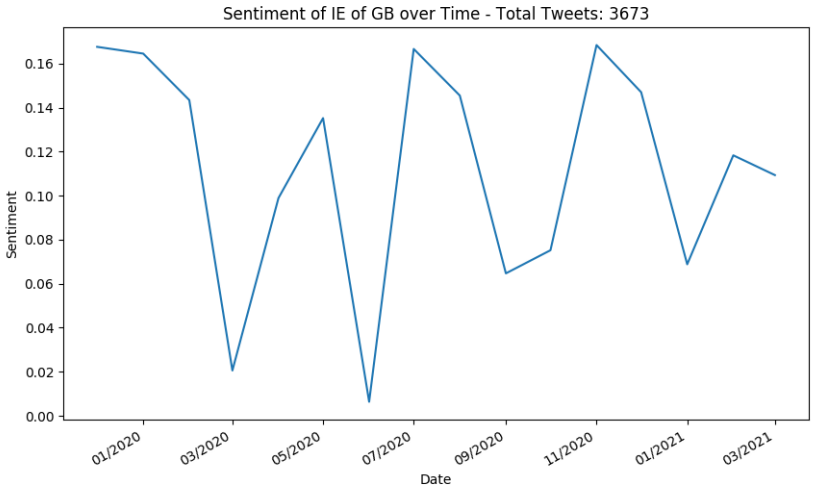

Figure 4: Ireland's Sentiment towards the United Kingdom

5.2 Exploratory

Standard deviation is a widely used measurement of variability or diversity used in statistics and probability theory. It shows how much variation or "dispersion" there is from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values. Error bars are used on graphs to indicate the error, or uncertainty in a reported measurement. They give a general idea of how accurate a measurement is, or conversely, how far from the reported value the true (error free) value might be. Error bars often indicate one standard deviation of uncertainty, but may also indicate the standard error. These quantities are not the same and so the measure selected should be stated explicitly in the graph or supporting text. Error bars can be used to compare visually two quantities if various other conditions hold. For example we can compare the baseline sentiment between country A and country B and see that with the error bar included there is a potential that although it seems that country A is more positive it is also less accurate than country B as country B has a smaller error bar potentially leading to the realisation that country B is in fact more positive. The error bar determines whether these differences are statistically significant. Error bars can also show how good a statistical fit the data has to a given function meaning that if the error bar is much smaller than the average is more accurate. We implemented error bars in our plots for the average sentiment of each country in order to indicate any error, or uncertainty and allow for transparency in the accuracy of our results [44].

As you can see in figure 5, which is a bar graph showing the average sentiment of tweets from the UK, Spain, France, Italy and Germany, the sentiment seems to gravitate towards 0.5 which in this case is neutral, however the UK is slightly higher than the other countries and is sitting closer to 0.6. We created a graph like this for each country and the way we chose which countries to put on the same graph was relative to how much data we have on that country. This is with the goal and idea in mind that countries that we have more data on will give us a more accurate representation of that countries overall sentiment, such that we decided it would make the most sense to rank countries in this way. Another worthy thing to mention about these graphs are the error margins that we included, these done using standard deviation where a positive error is shown as +1 standard deviation above the mean and the negative is conversely -1 standard deviation below the mean.

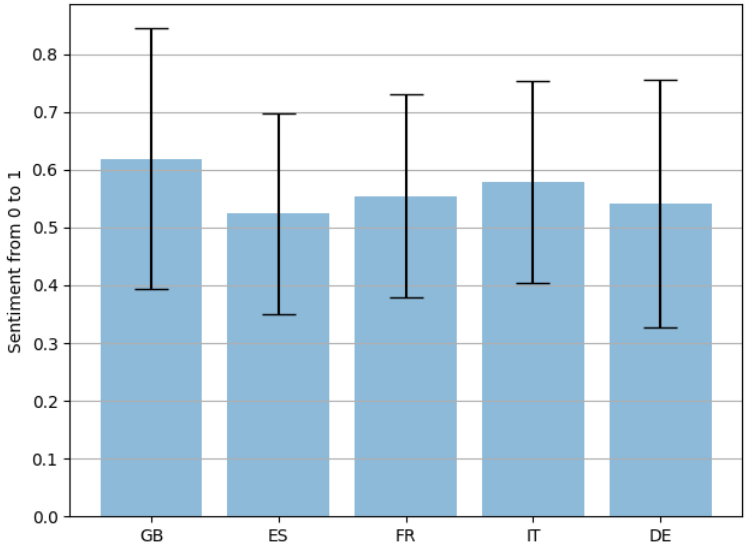

Figure 5: Baseline Sentiment of the United Kingdom, Spain, France, Italy and Germany

6 Validating Results

Throughout our project the aim has been to analyse the sentiment between other countries towards the UK after Brexit, and to discover the baseline sentiment that is found within each countries average tweets to see whether they trend towards negative or positive sentiment in the way that they communicate. This is what allows us to make an accurate judgement as to whether one country truly dislikes another as if a country is mostly negative it's difficult to prove that they dislike another country if their tweets are not more negative than their baseline sentiment. We saw that most countries would trend around a baseline of 0.5, which in the baseline graph translates to neutral, we suspect that this may either be because we did not have enough data or as the error bars on most of our results were fairly large, it seems safe to assume that there were mainly two types of sentiments found within tweets. The first would be extreme positive tweets and the latter would be extreme negative.

We estimate that countries like France, and those who are seen as big players within the EU, were not in favour of Brexit but as we may have formerly believed they are not as interested as the UK are when it comes to talking about Brexit the UK is completely overwhelming when we look at the magnitude of tweets that are discussing brexit from the UK. We hypothesise that Brexit was not a big source of change in sentiment between countries in Europe as there were not many user's talking about it except from those who are from the UK, where the subject was of great important and, where in Figure 3, it may seem that every country has a similar view of Brexit, that is not quite the case if you were to compare the number of tweets that the UK outputted compared to any of the other four countries displayed in the same graph. Brexit had a lot of contention within the UK but it seems

that despite it being about the EU, which a number of these countries are members, other countries in Europe were not very interested in engaging conversation on this topic.

7 Evaluation

7.1 Quality of the data

The quality of our data is influenced by a few factors, one of which is the frequency in which user's decide to communicate and another are bots. We had many instances of user's communicating with other's who lived in the same country as them, as such this meant we had a lot of extra data that we couldn't really use in our analysis as we needed to see conversations between different countries. The solution to this would be to create a sophisticated program that can search through a user's profile and determine whether they are communicating with somebody who lives elsewhere. That way we could remove more profiles from our dataset that do not give us the data we want and this would help to increase the speed at which our program runs. These user's who do not converse with foreigners are responsible for muddling the dataset and making it harder to explore, as well as inflating the size of the dataset without actually adding anything much of worth.

There are many bots on Twitter's platform and as a result they can create a lot of noise surrounding sentiment which can be quite disruptive in research as a data scientist as the data gathered from bots is completely useless and this issue is exasperated due to the fact that bots are constantly tweeting resulting in a lot of redundant data. We were not able to completely scour our dataset from all potential bots due to a lack of time. However, we could have used the sheer quantity of posts made by bots to our advantage as they are usually incredibly similar if not exactly the same. This allows us to spot whether a user is a bot and a solution to removing a bot once detected is as follows. Firstly as we are checking a profile we will look at this text of each of the user's posts, then if their posts seem to be suspicious due to similarities in virtually every tweet we will delete their profile and add their user id to a text file containing blacklisted accounts. This would mean that whenever we are creating new profiles because of new tweets we will first check to see if the user is flagged as a bot and if so we will not make the profile and then we will discard the tweet. Another solution to finding bots is to look at the screen name that they have as some bots are not as malicious as you'd think and they usually announce themselves as bots in their name. This would also allow us to remove accounts of corporations.

7.2 Increasing the Dataset

A larger set of data would allow us to look at trends and patterns in an alternative view as having a much larger period of time to work with lets us do things such as seeing how sentiment changes as trends come and go. Twitter has been operational for around 16 years so if we were able to collect data from the past decade we could more effectively see if change in sentiment between countries is frequent as a disadvantage of our dataset of one and half years is a much shorter lens to be looking for tangible change. Another expansion to the dataset that would result in an interesting yet potentially obvious subject is to continue the dataset from where it ends to the current day. This would most definitely include a influx of negative sentiment towards Russia and positive towards Ukraine due to the recent Russian invasion of Ukraine. Although, this is not quite definitive, we can assume that most countries sentiment of Russia has decreased since then, however we can not know for certain the historic sentiment towards Russia and the sentiment coming from Russia may not be as prevalent as it is elsewhere in the world.

Keyword	Tweets
putin	328186
zelensky	86122
russia|russian	536464
ukraine|ukrainian	687321
keiv|Kyiv	91142
kharkiv	27089
zaporizhzhia	8644

Figure 6: The Number of Tweets containing the keywords relevant to the Russo-Ukrainian Crisis [45]

7.3 Incorporating the Twitter API into the Design

An improvement to the design of the project would be if we had created our own program to collect data using the Twitter API and then we could directly request whatever attributes we wanted from Twitter and this would allow us to not only monitor change in sentiment live but also collect data that we could not acquire from SEDA Labs, like historic data and more recent data. This would be an improvement for the project and would enhance our findings except it would have required more dedicated work that was not in the scope of this project as with our dataset we spent an incredibly long time filtering, sorting and analysing. Adding the collection to that process would mean a lot more time spent on creating the program that collects the tweets from Twitter's API. If we had our own stream of tweets daily we could instantly sort these them as they arrive on our system and as we'd already have a stream set up to collect tweets it would be no challenge to parse the data as we have done in this project. This means we can create a live graph of each countries daily baseline sentiment and research like this to observe daily changes as they're happening and shifts in sentiment that we see can be potentially explained by any ongoing trends or themes from that countries tweets.

8 Critical Assessment

Our project can fall into the trap of being too general, as we want to look at so many different things that the lines can become blurred and it can be difficult to really define what the project is trying to achieve. In the future we would want to define a set focal point which we can have as a constant in the project in order to base our findings around and to work it into the entire project. An example of something we could've done is to use an ongoing conflict, like the war in Ukraine, or an upcoming trend like cancel culture. In using the war we have defined keywords and a specific target country, we'd have Ukraine and Russia, as our two targets and we can see everyday as new advancements are happening how each country responds to it and this can introduce some questions about the war, especially in terms of misinformation and propaganda that may be observed.

A critical problem that we ran into whilst conducting this investigation was that our machine that we ran every process on was not powerful enough to speedily finish each iteration, this meant that we could not make complete use of our dataset as we did not have the time nor money in order to effectively cycle through absolutely everything. Fortunately, we were able to gather enough data to make reflections and statements on our findings and how we believe it could be improved on and maybe even continued.

One such improvement would be to get the power required to compute much more data, as the more data we are able to collect the more accurate, our results will be, ending in a more fruitful conclusion. Whilst we were able to collect lots of data relevant to brexit, as we couldn't gather as much general data as we wanted, it does mean that we do not have as accurate a baseline sentiment model as we'd otherwise like to have.

9 Conclusion

Our revised project had us following the conversation on Twitter about Brexit and it investigated each European countries sentiment towards the United Kingdom in an effort to discern the effect that leaving the EU had on the world. We found that country country country, thought very negatively of the UK around our dataset and in fact, from the official date of the UK leaving the EU there was a spike in negatively all over where even positive country went positive over that month. We can conclude from this that other countries do not have a positive connotation with Brexit and that after leaving the EU countries behaved even more negatively than usual, this means that brexit did have an effect on how the rest of Europe views the UK.

Sentiment analysis is generally quite a crude model compared to it's counter parts, it does not delve too deeply into the content of a person's message and relies solely on the sentiment detected within. If we were to continue this project in the future we would want to use more advanced language models like topic modelling, which is a model that is used for discovering abstract topics that can appear in a dataset, like ours, that is full of sparse text data in various forms of slang [46].

The birth of social media has brought many people together through it's easy access and exploratory nature is the reason that we have so many people from different countries who may normally never have interacted but with the internet they were given this ability to converse. Every countries individual DNA can be seen throughout their footprints on the internet and the otherwise implied relationship that we see between country to country, we can see in real time on places like Twitter where we see new events and trends that spark conversation and arguments that better define and show these relationships. Something that could be expanded in the future is to look much deeper into these relationships and to observe key events that may have defined, or changed a countries sentiment toward another. We would want to steer away from the positive and negative sentiments and move closer towards classifiers that allow us to find what emotions are felt from one country to another, as we know how countries view others in terms of positive or negative, but we don't have a very complex result of these sentiment.

References

[1] A. Balahur, "Sentiment analysis in social media texts," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Jun. 2013, pp. 120–128.

[2] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of twitter users," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 554–557, 2011.

[3] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, pp. 216–225, 2014.

[4] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–41, 2016.

[5] L. Alonso-Muñoz and A. Casero-Ripollés, "Populism against europe in social media: The eurosceptic discourse on twitter in spain, italy, france, and united kingdom during the campaign of the 2019 european parliament election," *Frontiers in communication*, vol. 5, p. 54, 2020.

[6] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[7] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," *Mining Text Data*, pp. 415–463, 2013.

[8] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1014–1023.

[9] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019.

[10] Twitter, "Twitter api documentation," https://developer.twitter.com/en/docs/twitter-api.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW '10*, 2010.

[12] R. Arthur, C. A. Boulton, H. Shotton, and H. T. Williams, "Social sensing of floods in the uk," *PloS one*, vol. 13, no. 1, p. e0189327, 2018.

[13] J. C. Young, R. Arthur, M. Spruce, and H. T. Williams, "Social sensing of heatwaves," *Sensors*, vol. 21, no. 11, p. 3717, 2021.

[14] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, "Analysis of political discourse on twitter in the context of the 2016 us presidential elections," *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626, 2017.

[15] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of indian farmers' protest using twitter data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100019, 2021.

[16] J. Pal and A. Gonawela, "Studying political communication on twitter: the case for small data," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 97–102, 2017.

[17] A. D. Dubey, "Twitter sentiment analysis during covid-19 outbreak," *Available at SSRN 3572023*, 2020.

[18] I. Priyadarshini, P. Mohanty, R. Kumar, R. Sharma, V. Puri, and P. K. Singh, "A study on the sentiments and psychology of twitter users during covid-19 lockdown period," *Multimedia Tools and Applications*, pp. 1–23, 2021.

[19] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," *Proceedings of the ACM SIGIR: SWSM*, vol. 63, 2011.

[20] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[21] G. Stricker, "The 2014 #YearOnTwitter," 2014. [Online]. Available: https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

[22] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation*, 01 2017, pp. 502–518.

[23] R. Arthur and H. T. P. Williams, "The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales," *PLOS ONE*, vol. 14, pp. 1–14, 2019.

[24] P. Parau, A. Stef, C. Lemnaru, M. Dinsoreanu, and R. Potolea, "Using community detection for sentiment analysis," in *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2013, pp. 51–54.

[25] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "Politwi: Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 24–33, 2014.

[26] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," in *Data and Applications Security and Privacy XXIX*. Springer International Publishing, 2015, pp. 170–187.

[27] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in *The Semantic Web – ISWC*. Springer Berlin Heidelberg, 2012, pp. 508–524.

[28] S. H. W. Ilyas, Z. T. Soomro, A. Anwar, H. Shahzad, and U. Yaqub, "Analyzing Brexit's impact using sentiment analysis and topic modeling on Twitter discussion," in *The 21st Annual International Conference on Digital Government Research*. Association for Computing Machinery, 2020, p. 1–6.

[29] P. Baid, A. Gupta, and N. Chaplot, "Sentiment analysis of movie reviews using machine learning techniques," *International Journal of Computer Applications*, vol. 179, pp. 45–49, 2017.

[30] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[31] R. Tong, "An operational system for detecting and tracking opinions in on-line discussions," in *Working Notes of the SIGIR Workshop on Operational Text Classification*, 2001, pp. 1–6.

[32] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions of Information Systems*, vol. 21, no. 4, p. 315–346, 2003.

[33] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Security Informatics*, vol. 4, 2015.

[34] S. Lab, "Seda lab's blog page." [Online]. Available: https://blogs.exeter.ac.uk/seda-lab/

[35] T. Butcher, "Objects of intense feeling: The case of the twitter api," *Computational Culture*, 2013.

[36] Twitter Developers, "Developer agreement and policy," *Twitter Developers: San Francisco, CA, USA*, 2020.

[37] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, "Privacy issues and data protection in big data: A case study analysis under gdpr," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 5027–5033.

[38] J. Shepherd, "22 essential twitter statistics you need to know in 2022." [Online]. Available: https://thesocialshepherd.com/blog/twitter-statistics

[39] R. J. Osborne, B. F. Bell, and J. K. Gilbert, "Science teaching and children's views of the world," *European Journal of Science Education*, vol. 5, no. 1, pp. 1–14, 1983.

[40] S. S. Inc.., "Leading countries based on number of twitter users as of january 2022." [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[41] Translated, "Mymemory documentation," https://mymemory.translated.net/doc/.

[42] H. Li, S. Liu, X. Yu, S. Tang, and C. Tang, "Coronavirus disease 2019 (covid-19): current status and future perspectives," *International journal of antimicrobial agents*, vol. 55, 2020.

[43] BBC, "Brexit: What you need to know about the uk leaving the eu." [Online]. Available: https://www.bbc.co.uk/news/uk-politics-32810887

[44] B. F. Life, "Interpreting error bars." [Online]. Available: https://www.biologyforlife.com/interpreting-error-bars.html

[45] E.-U. Haq, G. Tyson, L.-H. Lee, T. Braud, and P. Hui, "Twitter dataset for 2022 russo-ukrainian crisis," 2022. [Online]. Available: https://arxiv.org/abs/2203.02955

[46] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.