Geographic Sentiment Distribution of COVID-19 in the UK

COMSM0017: Applied Data Science

Andrew Corrigan
Engineering Mathematics (of UoB.)
ac16534@bristol.ac.uk

Faizaan Sakib
Computer Science (of UoB.)
ss16161@bristol.ac.uk

Eleanor Begbie
Engineering Mathematics (of UoB.)
eb16727@bristol.ac.uk

Arabella Peake
Engineering Mathematics (of UoB.)
ap16260@bristol.ac.uk

Mark Nicholl
Computer Science (of UoB.)
mn16660@bristol.ac.uk

Leechay Moran-Allen
Computer Science (of UoB.)
lm16154@bristol.ac.uk

***Abstract*—Infecting over 4.44 million people, the coronavirus is a great threat to the global economy and the well-being of individuals. This investigation collects and compares various data sources focused around the ongoing pandemic of COVID-19. A large database of UK based COVID-19 tweets is collected before a developed model is applied to estimate the sentiment of the data-set. The overriding sentiment is compared to various other metrics such as the FTSE-100 value and the number of UK COVID-19 cases. The data is displayed on an online dashboard via easily interpreted and interactive graphs. Overall, correlations are found between Twitter sentiment and various COVID-19 metrics giving us a deeper understanding of the social effects and public sentiment towards pandemic over time.**

I. Introduction

To date, the coronavirus has taken the lives of over 300,000 individuals; the COVID-19 pandemic is undoubtedly the biggest challenge today's world has had to face. Mathematical modelling is an essential tool for understanding the spread of the disease, the infection rate and for making future predictions. This report is concerned with the geographic sentiment distribution of COVID-19 in the UK.

Social media is the largest, most accessible platform on which information is shared- enabling individuals to express their opinion to a large number of others. Millions of people every day participate in the writing, sharing and reading of tweets, allowing content to spread viciously. Twitter is one platform containing an incredible library regarding the public's attitude towards the COVID-19 pandemic at any time, making it a valuable platform for analysis.

Data is collected from four sources: tweets (Twitter), COVID-19 dataset (the Johns Hopkins University Centre for Systems Science and Engineering), FTSE-100 (obtained from ShareCast) and news articles (The Guardian). All of the collected data is processed such that is is viable to use in analysis, for example the removal of all upper case letters and the removal of all numbers. Deeper details of this are given in Section III-B. Having processed the data, is it stored as a `CSV` file or as an SQLite database where appropriate.

Appropriate methods are then explored to visualise, analyse and compare the data. Supervised and unsupervised learning is explored and discussed in order to determine the sentiment of the tweets before we investigate the relationships present between the collected data sets.

This report will begin by analysing the overall sentiment of tweets, i.e. whether the general attitude is positive or negative in a specified area. An investigation into the correlation between sentiment and the number of cases will take place. We will also investigate how the number of cases correlates to the FTSE-100 market price.

II. Ethics and Safety

Before preceding, it is important to note the ethical and privacy issues that can arise when conducting analysis using Twitter data. With more households and individuals having access to social media platforms, there are currently 330 million Twitter users. As the use of Twitter increases, its value as a data source increases also. Despite being an insightful data source, (containing a library of society's feelings, opinions and news) there are many key issues within social media research, such as whether social media spaces are private or public spaces.

Where some social media platforms are clearly private, such as your Facebook profile, others are seen as public spaces for online communication between individuals, such as Twitter. In the later case, the distinction of privacy becomes unclear. Content (tweets, photos, news articles) shared and published on Twitter is publicly accessible via the Twitter API and/or via data re-sellers. Additionally, the default setting of an individuals Twitter account is public. These factors mean we do not know the extent to which individual users of Twitter are aware of the public accessibility of their data/information. Working with a data-set containing a large number of tweets poses the ethical concern that one is unable to gather individual, specified consent from every tweet. It is assumed that by using Twitter, users are accepting that their data may be used in such an investigation, however this assumption can be contested as the extent to which users understand this assumption is blurred.

Twitter is not the only social media platform on which these ethical concerns are present. When using this sort of data in any investigation, these questions must be asked and researchers must be wary not to breach any legal privacy terms.

III. Data Preparation

The data, on which sentiment analysis will be performed, forms the basis of this project. The dataset is independently produced, due to the project's specific aims. The social media platform Twitter is chosen as the main source of data for analysis. The following characteristics make Twitter a suitable choice:

- it is made up of publicly available text-based data entries, known as "tweets"
- it is more suited to topical discussion compared to other social media platforms
- it has a large active user base resulting in a high volume of accessible data
- there are a number of well-documented and configurable APIs available for fetching tweets

Once the sentiment analysis of the tweets is carried out, it is essential that there is a reference point which can be used to evaluate the performance of the model. There does not exist a labelled dataset based on sentiment analysis, in relation to COVID-19. Thus, in addition to the Twitter data, a number of datasets are gathered to be used as sources of ground truth that can help to verify the predictions of a sentiment analysis model. They include:

1) COVID-19 data
2) FTSE-100 data
3) UK-based news articles relating to COVID-19

The data for all sources mentioned above is collected within a 3 month time-frame between 25/12/2019 to 25/03/2020. This takes into account a preliminary period where there is no expected influence on the data, as the specified time-frame begins a week before the first known cases of COVID-19 were reported [1]. Furthermore, the length of the period considered should be sufficient for performing temporal analysis of sentiment.

A. Collection

1) Twitter: The tweets are fetched using TWINT [8]. It is a scraping tool which does not use the Twitter API, bypassing the limitations that comes with it. This means there is no limit to the number of tweets that can be fetched. More importantly, tweets can be fetched in terms of their location by city, which is necessary for the purposes of geographical analysis.

In order to obtain tweets relevant to COVID-19, the search is performed by keywords "coronavirus" and "COVID-19". This filters out tweets that do not contain these keywords. TWINT ensures that all variations of the keywords (e.g. lowercase, uppercase, captialised) and their hashtags are considered. A total of 10 cities are surveyed for COVID-19 related tweets, which include London, Bristol, Manchester, Liverpool, Birmingham, Leeds, Oxford, Glasgow, Belfast and Cardiff. This ensures that the tweets are pooled from an adequate number of cities, representing most regions of the UK.

Unlike the Twitter API, TWINT does not have data streaming capabilities. Furthermore, attempting to filter tweets by their location significantly increases the time TWINT takes to fetch them. It would not be feasible to have a dataset that attempts to update in real-time. Nevertheless, it would also not be possible to filter tweets by location with the Twitter API, leaving TWINT as the only viable option for the purposes of this project. As a result, the tweets are fetched across the time-frame specified in Section III.

A limit of 500 tweets are fetched for each city per day across the time-frame. The list of dates is obtained using `pandas.date_range()` with the required start and end dates passed as arguments. The dates are then iterated through, fetching tweets through TWINT for each city. These tweets are further filtered to only include non-popular and non-verified tweets. This makes up the main set of tweets.

This is repeated with a limit of 100 tweets, to obtain two additional sets of tweets. The first includes only popular tweets, and the second includes tweets that are both popular and from verified accounts. These datasets can be seen as more credible sources of sentiment that can be used to cross-validate the sentiment analysis model against the main set of tweets.

It is worth emphasising that the number of tweets fetched from any day for any city may fall below the limit provided. This is because there may not have been that many tweets, which include the keywords, posted that day. Furthermore, TWINT finds tweets by location only if the tweet has location data encoded in it. This option is set by the user themselves and is turned off by default. This means that a subset of the actual population of tweets for each city is instead being sampled. These factors make it inevitable that there is some variation of the number of tweets obtained for each city. The level of variation can be seen in Figure 1, where the vast majority of the tweets originate from London. There is some concern regarding the cities for which far less number of tweets have been obtained. In particular, only 258 tweets are obtained from Oxford in a three month period, which seems unlikely even with the consideration that a smaller population is being sampled. Repeated attempts of fetching tweets from these cities failed to address this issue. This is accepted to be a result of TWINT's unreliable functionality.

There is a total number of 28876 tweets in the main set. The verified, and verified + popular sets are made up of 16985 and 7930 tweets respectively.

2) COVID-19: The COVID-19 dataset provided by the Johns Hopkins University Center for Systems Science and Engineering is used [5]. This dataset includes the number of COVID-19 infections, recoveries and deaths for the UK, along with every other country, over the course of the pandemic. The number of cases of each category is updated as a cumulative figure. This dataset has been known to be a well-reputable source of data during the pandemic. It is updated every hour.

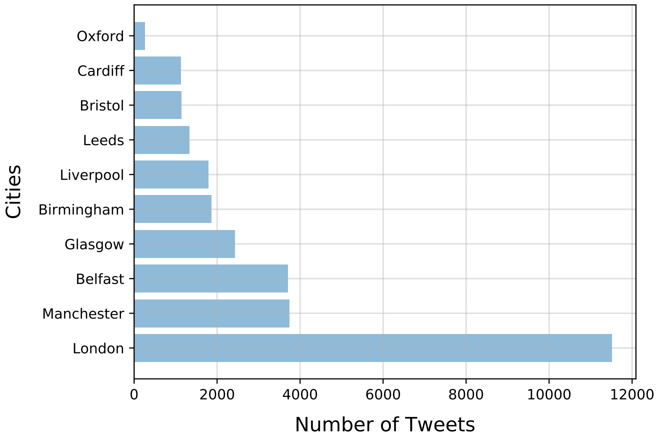


Fig. 1: A bar graph showing the distribution of tweets across every city in the Twitter dataset.

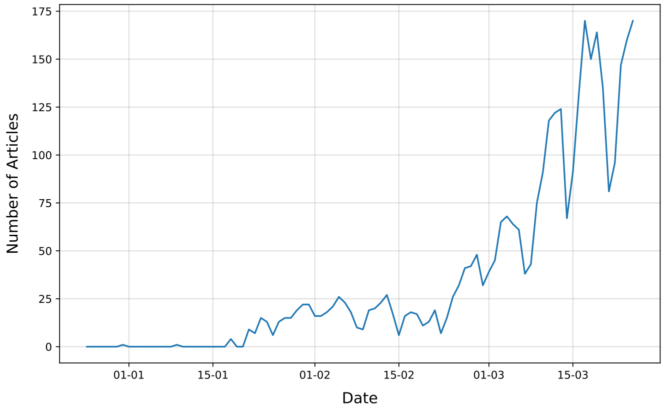


Fig. 2: A line plot showing the increase in the number of COVID-19 related articles produced by The Guardian over time.

3) FTSE-100: The data for the FTSE-100 is obtained from ShareCast [7]. It has the open, close, low and high price of the FTSE-100 recorded for each day that the market has been in operation.

4) News Articles: This dataset consists of news articles related to COVID-19. The articles should primarily be from the UK, similar to the focus of the project. The Guardian is used as the source of this dataset. It is based in the UK and hosts a well-documented API available for obtaining its articles.

The API is used to filter articles by The Guardian, to ensure that they contain either of the terms “coronavirus” and “COVID-19”. They are further filtered by date, to only get articles within the 3 month time-frame. However, there is a complication with the API, where the articles are returned one “page” at a time, with each page including a maximum of 50 articles. Therefore, a single request may not actually be returning all articles that are found by the search query. To circumvent this, the number of pages that exist for the query can be obtained. With this, each page making up the search query’s overall output can be requested. By iterating through the pages, a list of all of the articles can be aggregated. A total of 3284 articles related to COVID-19 are collected.

B. Pre-Processing

All pre-processing is done with the use of `Python`. The `Pandas` library is also used to represent the raw datasets as data structures that can be analysed and manipulated as needed.

1) Twitter: Each tweet obtained from TWINT has a total of 31 fields. Many of these fields are unnecessary and are therefore removed. This is done in place with `pandas.drop()`.

Recall that there are three different Twitter datasets (main tweets, popular and popular + verified) that are collected. Columns named “popular” and “verified” are added to specify whether the tweets are of those nature. Among the three datasets, the main tweets are mutually exclusive as they explicitly do not meet the search conditions of the other two sets of tweets. However, the popular + verified set of tweets may have some number of duplicate tweets that also appear in the popular set. These tweets need to be removed to avoid repetition of data. Each tweet has a unique ID assigned by Twitter. Using this attribute, the popular + verified set is filtered by the tweet IDs found in the popular set. If any tweet is found to have a tweet ID that also exists in the popular set, it is removed from the popular + verified set.

Following this, the actual text that the tweets are comprised of need to be processed in preparation for the sentiment analysis model. It is vital that all text that is not valuable input for the classifier must be removed, it is also vital that the input is normalised so that words with the same sentiment can be represented similarly for the classifier. This will be done through several key steps:

- **Removing all numbers**: Sentiment is hard to derive from numbers, for example dates can only carry sentiment in relation to other dates. The benefits of standardising the input and removing numbers outweigh the potential sentiment lost by removing them.
- **Replacing upper case letters**: The input for vectorisers is case sensitive and as such would read capitalised and uncapitalised words as having different sentiment. This is often not the case and as such all upper case letters will be replaced with lowercase to standardise input.
- **Removing stop-words**: Stop words are words that carry little meaning in the context of the sentence. These are divisions of natural language such as articles, prepositions and pro-nouns. e.g. ”the”,”and”,”it”.
- **Removing Links**: Links often occur in tweets and whilst the link could lead to a website that could be analysed to understand the sentiment of the tweet this is in-practical.

2) News Articles: Among the articles that were retrieved through the Guardian API, some of them did not have the search keywords within the article. This seemed to be an issue of the API. Any articles not including the keywords are consequently removed.

3) FTSE-100: The data for the FTSE-100 is correct and concise for the purposes of the project. No preprocessing is necessary.

4) COVID-19: There are some instances of non-countries recorded with cases (e.g. the Diamond Princess cruise ship) that are not required and therefore dropped. Furthermore, there are negative values which sometime appear. This is known to be an error in the dataset. These values are therefore removed.

C. Storage

The Twitter dataset is populated as a `CSV` file. Due to the high volume of data, querying it becomes inefficient. This issue must be resolved as the dashboard will contain many visualisations that will attempt a large number of queries at the same time. This may lead to a significant increase in load time for the dashboard if a `CSV` were to be used. Therefore, once the required pre-processing is complete, the tweets are stored as an SQLite database. The FTSE-100 and news datasets are also stored in the SQLite database, as separate tables.

The COVID-19 dataset is retrieved from the source as a `CSV` file. As the source regularly updates the data, the dashboard aims to utilise this to display live data relating to COVID-19. The `CSV` is relatively small, and figures are accumulated upon as a time series. Due to the data's simplicity and its evolving nature, there is no need of storing it in a database. If it was to be stored in one, it would unnecessarily require two operations to be put in place, retrieving the data, and then updating the dashboard's database. Instead, the `CSV` is fetched, manipulated, and processed within the source code. The file is not stored on disk.

IV. Data Exploration

Exploration of the data is necessary in order to ensure that sufficient data has been gathered for the purpose of this project.

1) Twitter: Assessing the content of the popular and verified set of tweets, it is found that a large proportion are factual, with many containing COVID-19 related statistics and references to news articles. The sentiment of these tweets is therefore neutral as there is not an underlying opinion in most cases. For this reason, for future sentiment analysis, only tweets which are not popular nor verified are considered. When investigating the geographical location of the tweets, it has already been highlighted that the majority of the tweets pooled are from London, as seen in Figure 1. This leads to an uneven distribution of tweets pooled based on geographical location.

2) News Articles: As previously mentioned, some of the retrieved articles did not have the keywords present and such articles were removed as these are not relevant for the purpose of this project.

3) FTSE-100: Due to the nature in which the FTSE-100 data is updated and recorded, certain dates may not portray the true behaviour of the stock market. Stock exchanges do not operate at the weekend or on national holidays. For these dates, the stock market prices are set to the market close price of the previous day. The other three data sets obtained contain their true value for every date in the time range. Therefore, this is an important aspect which needs to be considered when comparing the time series of the FTSE-100 data with those obtained using the other data sets.

4) COVID-19: After exploring this data set, no significant concern was highlighted which could notably affect further modelling and analysis.

V. Modelling

To extract usable information from the tweet data we apply several techniques for sentiment analysis. The effectiveness of the techniques can be compared with the COVID-19 tweet data, we will therefore be able select the model which gives the most reliable sentiment analysis and use that model to infer the sentiment to be displayed and compared on the website. We will explore two main branches of machine learning: supervised and unsupervised learning.

To make sure each model developed can be compared effectively and validly a series of standardised performance measures will be used. It is also essential that for each method these performance measures are evaluated on the same labelled sub-set of the collected tweet data.

Three main measures of performance will be used:

1) **Precision**

$$\frac{t_p}{(t_p + f_p)} \tag{1}$$

Where t_p is the number of true positives and f_p is the number of false positives. Precision is the model's ability to not label a negative sample as positive, a measure of the models robustness to false positives.

2) **Recall**

$$\frac{t_p}{(t_p + f_n)} \tag{2}$$

Where f_n is the number of false negatives.Recall is the ability of the model to find positive samples, it is the models robustness to false negatives.

3) **Weighted F1 Score**

$$F1 = \frac{(P \cdot R)}{(P + R)} \tag{3}$$

Where P is the precision and R is the recall, previously defined. This is a combination of both the precision and accuracy and gives an intuitive score for the overall performance of the model. The relative combination to the average for each score is equal. It is a measure of the incorrectly classified cases compared to the accuracy which is simply a measure of the correctly classified cases.

A. Supervised Learning

Supervised learning maps an input to an output based on a set of example input-output pairs. This means that to learn the mapping the method must have example input-output pairs or supervision set to learn from. For a reliable classifier the training supervised set must be sufficiently large so that the mapping can be reliable for future extrapolation of the model. The tweet data collected for sentiment analysis is not labelled

Corpus	**Size**	**Features**
Sentiment140	1,600,000	82,221
COVID-19	28,000	9,544

TABLE I: A comparison of the training corpus and the COVID tweet corpus

and will not be able to be used as the supervision set. The set can either be labelled or a similar labelled training set could be used. As the size of the data-set required would be large, manually labelling a data set would be impractical. Therefore a similar data set will be used to train the supervised methods and then the learned model we be applied to the tweet data-set. There are many positive-negative sentiment data-sets available, it is essential for validity that the training set is as close as possible to the tweet data. We therefore will use the sentiment140 data-set [4]. This is a set of 1,600,000 pre-labelled tweets, each tweet is labelled as being either positive, negative or neutral. The model learned from this data set should be transferable to the Twitter data-set as both are tweet data. That means the input should be of similar structure and the same pre-processing steps outlined earlier can also be used. Here we show a comparison of the training data-set and the COVID-19 data set.

The sentiment140 data-set has a much larger number of tweets and features, this means that the classifiers learnt from this corpus can be transferred to the smaller corpus.

Two supervised learning techniques are compared. The first being a Support Vector Machine (SVM) with a Term Frequency - Inverse Document Frequency (TF-IDF) vectoriser and the second being an SVM with a word2vec vectoriser.

1) SVM with TF-IDF: The first model used is the currently accepted state of the art performance for sentiment analysis [3], therefore this will be the starting point and acts as a performance baseline for further development. This model is split into two sections: the vectoriser and the classifier. The vectoriser converts the raw text data into a matrix of dimension equal to the number of features. This matrix is then weighted, TF-IDF is a combination of two measures. Term frequency and Inverse document frequency. This are combined as follows. The score of the word t from the document d in the set of documents D is:

$$TFIDF(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{4}$$

$$tf(t, d) = log(1 + freq(t, d) \tag{5}$$

$$idf(t, D) = log(\frac{N}{count(d \in D : t \in d}) \tag{6}$$

This works by counting how many times a word appears in the entire document, or in our case the entire corpus, and weights the words inversely proportionally to this. This reduces the effect of words that appear throughout the corpus, typically words that carry little meaning. This means that the classifier is able to focus on words that are specific to the token and therefore the label. With the text data vectorised this can then be inputted to a classifier, in this case an SVM is used. An SVM is a binary classifier, it works by selecting a hyperplane that best separates a pre-labelled data set. It does through maximising the margin between the positively labelled training data and the negatively labelled training data.

2) SVM with Word2Vec: TF-IDF is a versatile method however it has some limitations. For large a corpus the vector representation is large, this can make the input vector large and reduce the accuracy of the classifier. The TF-IDF vector is also corpus-specific; when transferring the trained vectoriser from the sentiment140 data-set to the COVID tweet data-set some information could be lost. To potentially increase the accuracy and transferability of the classifier a word2vec vectoriser is used and compared. Word2vec is a deep learning technique using a two layered neural network to learn vector representations of individual words. It works by generating a vector of a set dimension, and distributes a representation of a word across these dimensions. Each element in the vector contributes a different aspect to the meaning of a specific word. This enables more complex representations of words. The weightings are learnt via the the two layered neural network. The word vector is typically trained from an extremely large corpus, in this case 1.6 billion words. This vector allows various mathematical computations on the words.

B. Unsupervised Learning

Unsupervised learning draws inferences from datasets consisting of unlabelled input data. As mentioned previously, the tweet data collected for sentiment analysis is not labelled, hence unsupervised learning is an appropriate method to explore. Here, a combination of a TF-IDF vectoriser and K-means is used in order to cluster the data into three classes, each representing either a positive, negative or neutral sentiment.

K-means clustering is a centroid-based approach to grouping data, with each data point belonging to the cluster with its nearest centroid. Here, the algorithm is initialised with $k = 3$ random centroids. K-means then employs a squared error criterion, whereby each data point is assigned to its closest centroid, forming clusters. The new centroid point is computed from the mean of the data points in each cluster. The process is then repeated as the data points are reassigned to their new closest centroid and the new centroid point is computed based on the data points in the new cluster. K-means is an iterative process which haults when the centroid points have stabilised; when there is no reassignment of points from one cluster to another. Here, K-means is therefore used to group partition the tweets into three groups based on their sentiment similarity.

C. Model Performance

To select the best model the performances must be compared. To do this accurately and to give a reliable measure of how well each method performs on tweet data-set a labelled subset of the collected Twitter data will be used. This has been manually labelled and as such can be used as the ”Golden Standard” set assumed to be 100% accurate. This also means

Model	Precision	Recall	Weighted F1 Score
SVM with TF-IDF	0.64	0.62	0.60
SVM with Word2Vec	0.76	0.73	0.72
K-means with TF-IDF	0.38	0.36	0.33

TABLE II: The precision, recall and weighted f1 score for each model. These are evaluated as the average value over all three of the classes.

that each model is tested on an identical data-set so that results are directly comparable.

In order to analyse which model has been able to classify the tweets the most accurately, we consider three evaluation metrics: the precision, the recall and the weighted F1 score, which are all averaged over the three classes. The values of these can be seen for each model in Table II.

Predictably the SVM with Word2vec achieved the highest precision, recall and F1-score, marginally higher than the TF-IDF SVM however notably higher than the unsupervised K-means. Therefore the SVM with word2vec will be used to generate sentiment scores for the collected Twitter data. These scores will consist of a label of either Neutral, Negative or Positive and a confidence score for Negative and Neutral that can be used as a measure of how negative or positive a tweet is.

VI. Evaluation

A. Sentiment Analysis Results

Figure 3 shows a time series of the number of negative, positive and neutral tweets from the collected Twitter data. the first trend this graph shows is that over time the number of tweets found to mention COVID-19 increase dramatically, With a large increase immediately after the first UK death. The figure also shows that the number of negative and neutral tweets increased at a much greater rate than positive. This shows that as COVID-19 has developed the majority of the tweets are negative rather than positive. The graph also shows that after the 15th of March the number of tweets for negative, neutral and positive plateau showing that for the tweets in the database, the interest peaks. It's worth noting that this around the time when the UK-wide “Lock-down” was implemented. After the lock-down the negative sentiment does not increase at the same rate. Interestingly the graph does not show a large increase in tweets after the first UK case, there is a brief spike around 2020-02-03 however this settles until the first UK death. This shows that until the first UK death COVID-19 was not mentioned frequently suggesting that it was not something people were aware of or worried about until the first UK death.

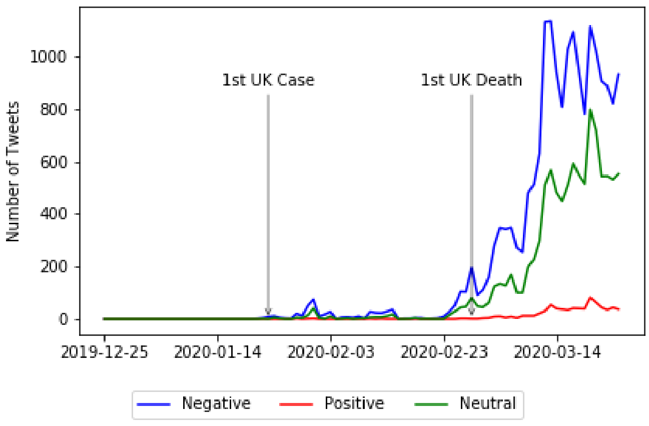


Fig. 3: A time series of the number of positive, negative and neutral tweets mentioning the COVID-19 keywords. The time series also shows certain key events; the 1st UK case and 1st UK death.

Figure 5 shows the sentiment of the tweets containing certain key words. This highlights specific words of interest and shows the general sentiment of the tweets towards the certain aspect of COVID-19. Unsurprisingly words such as ”coronavirus” and ”cancelled” are found to have mainly negative sentiment associated with them. There are however some surprising words associated with negative sentiment, the second most negative word was ”BBC” with roughly 80% negative tweets. This shows that of the tweets collected, the majority mentioning the BBC were negative suggesting that they are unhappy with the BBC's coverage of the pandemic. Similarly NHS has one of the higher proportions of negative tweets, showing that the NHS is associated with negative sentiment. This could be as the tweeters feel negatively towards the NHS and it coping with the pandemic or alternatively expressing concern about the NHS. Classifying sentiment as strictly negative or positive limits the analysis, the exact sentiment cannot be identified however it does help highlight general trends.

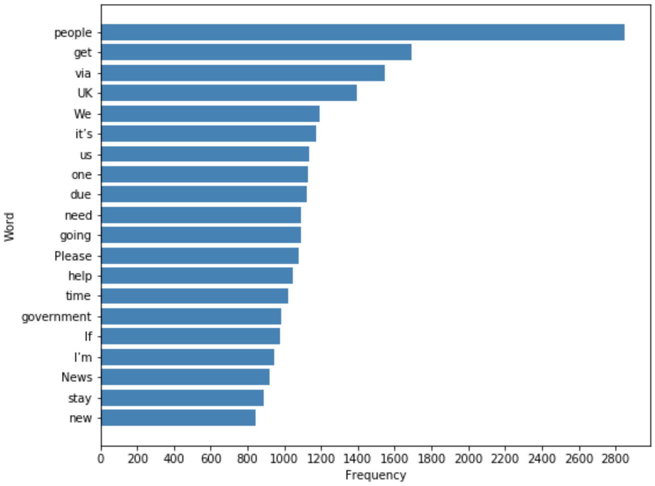


Fig. 4: A bar graph showing the frequency of the top 20 words. It excludes search terms used to gather tweets. This highlights trends in tweets about COVID-19

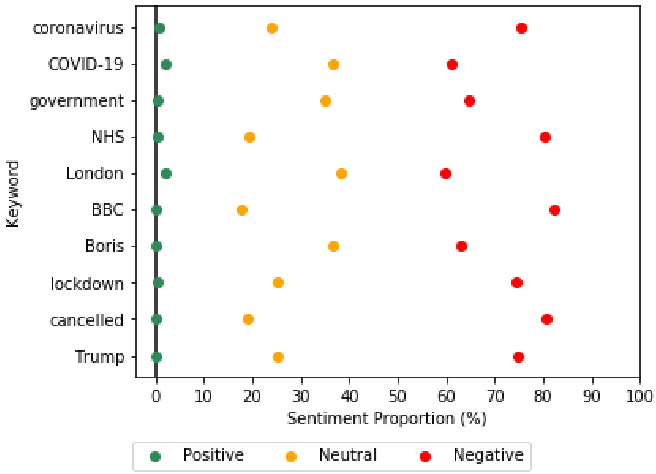


Fig. 5: A plot of selected key words. For each word shows the percentage of negative, neutral and positive tweets containing the word is shown. This is a measure of the sentiment towards each of the selected key words in the context of the pandemic.

B. Model Validation

In order to assess the performance of the model, its results have to be validated against trusted secondary sources of sentiment. This is achieved by examining the correlation between the results of the sentiment analysis, and the other sources. This can give an insight into how well the model is able to make predictions that reflect true sentiment relating to COVID-19.

As mentioned in Section III, three additional datasets were gathered along with the Twitter data. From these datasets, the following variables will be used as sentiment indicators:

- Number of confirmed cases of COVID-19 in the UK
- Change in FTSE-100 relative to the market open price of 25/12/2019
- Number of news articles published that are related to COVID-19

These variables are either directly, or indirectly, influenced by the COVID-19 pandemic. They are also independent of the sentiment analysis performed on the tweets. Similar to the analysis of the tweets, they can be represented as time-series data. As such, these sources help to build a reference point for sentiment towards COVID-19.

The number of tweets labelled as "negative" sentiment per day is used as the primary variable representing the model. This will be compared to the secondary sources of sentiment.

All variables must be normalised to allow for their comparison. Each variable is constrained to the same 3 month period that the tweets were collected within. This is done during the pre-processing of the data, as explained in Section III-B. Next, the values of different units across the variables are transformed to ensure that they fit a uniform scale. This is accomplished with the use of the following equation:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{7}$$

where z_i is the adjusted value for the i^{th} value of variable x. This normalises values to the range $[0, 1]$. In this particular case, 0 represents no negativity, while 1 represents strong negativity. Note that the concept of negativity depends on the nature of the variable itself. For the purposes of this task, it is accepted that high levels of negativity is synonymous with high numbers of: tweets labelled as "negative" sentiment, confirmed COVID-19 cases, news articles relating to COVID-19. In contrast, for the FTSE-100 data, a high market price, or its increase, indicates confidence and hence positivity. Therefore, the FTSE-100 data would produce adjusted values near to 1 where there is no negativity. This is the reverse of how the other variables are defined in the same scale. To ensure that the normalised FTSE-100 prices are in line with the other variables, the complement of Equation 7 is taken to be z_i. With this, the lower the market price is, the higher the adjusted value that will be produced.

When the normalised values are overlapped, as shown in Figure 2, there is a visual indication of correlation between the variables. The lines follow a clear trend throughout the three month period. There is minimal activity at the start of the year, near to when the pandemic had only just begun. There is some fluctuation in the FTSE-100 market price at this point, but this is expected as it reflects the constant change in market prices. Following this, there is an increase in magnitude of all variables as the effects of the pandemic start to be noticed. This is further followed by a sharp rise, coinciding with the date of the first COVID-19 related death in the UK on 05/03/20. Levels of "negativity" remain at high levels upto and beyond the day the nationwide lockdown was announced on 23/03/20. Meanwhile, the growth in the number of COVID-19 cases in the UK shows an exponential rise.

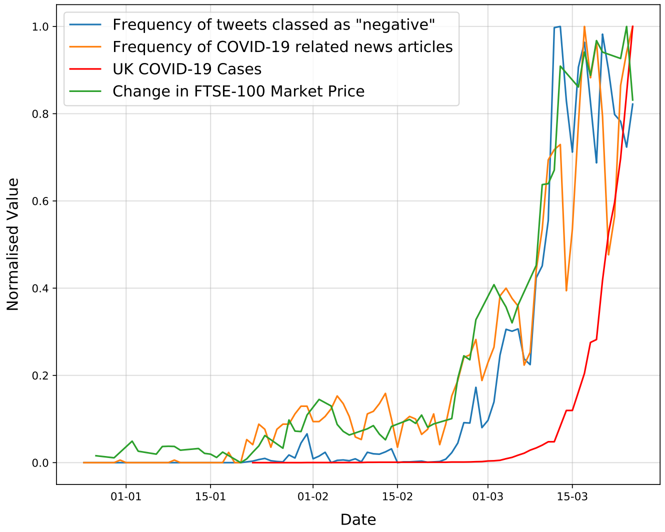


Fig. 6: A graph plotting the normalised values of four different variables relating to COVID-19 over the three month period. Specifically, it displays the correlation between the negative sentiment found in the Twitter dataset when compared to three other secondary sources which indicate sentiment.

Variable	Twitter	COVID-19	News	FTSE
Twitter	1.000	0.696	0.931	-0.964
COVID-19	0.696	1.000	0.729	-0.710
News	0.931	0.729	1.000	-0.983
FTSE	-0.964	-0.710	-0.983	1.000

TABLE III: A coefficient matrix displaying the Pearson correlation coefficients calculated between each variable. They include: tweets with sentiment classed as negative (Twitter), number of COVID-19 infections in the UK (COVID-19), number of UK-based COVID-19 related news articles (News), and the FTSE-100 market open price (FTSE)

Next, the correlation that is evident in Figure 6 between the variables is quantified using the Pearson correlation coefficient [2]. It is used as a method to determine the strength of the relationship between two variables. It indicates how a change in one variable affects another. For example, a strong positive correlation signifies that, with the increase in variable x, there is a tendency that variable y also increases. The Pearson coefficient, r, is a value in the range $[-1, 1]$, where -1 represents total negative correlation, 1 represents total positive correlation and 0 means there is no correlation between the two variables. It is assumed that the variables in question have a linear relationship. Within the 3 month period, there is an overall pattern of linear correlation seen in Figure 6. Although this relationship may change beyond this time frame, for the purposes of this project, it is assumed that they have a linear relationship.

The Pearson coefficient r for two variables x and y is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{8}$$

where n is the sample size, x_i, y_i are the individual values of the sample indexed with i and $\bar{x}, \bar{y}$ is the sample mean for each variable. The Pearson correlation can be applied to two variables of different units. Thus, it is not necessary to use the normalised values of the variables to calculate the coefficient.

The Pearson coefficient is calculated between the negative sentiment Twitter variable and each of the secondary sources to determine the model's effectiveness. Furthermore, the coefficient is also obtained between the secondary sources, to confirm their own validity as sentiment indicators. As mentioned in Section IV, the FTSE-100 data does not have sample points of market prices for weekends. This means its sample size is smaller than the other datasets. As such, any variable that is compared to the FTSE variable has its sample points for weekends omitted, in order to have the same sample size n for both variables.

The Pearson correlation coefficients calculated between the Twitter variable and the secondary variables show strong correlation in their relationships. There is particularly strong correlations found comparing the negative sentiment found on Twitter, with the frequency of news articles posted, and the FTSE-100 market price. They have correlation coefficients of 0.931 and -0.964 respectively. Note that all coefficients involving FTSE-100 are of a strong negative correlation. This is due to the inherent inverse nature of market price relating to confidence, as discussed earlier in this section. As a result, the strongly negative coefficient between the Twitter and FTSE variable shows that an increase in negative sentiment on Twitter points towards a decrease in the FTSE market price, and vice versa. Likewise, the FTSE coefficient is also negative for the other two secondary sources. This resembles the behaviour of markets such as FTSE-100 in reality.

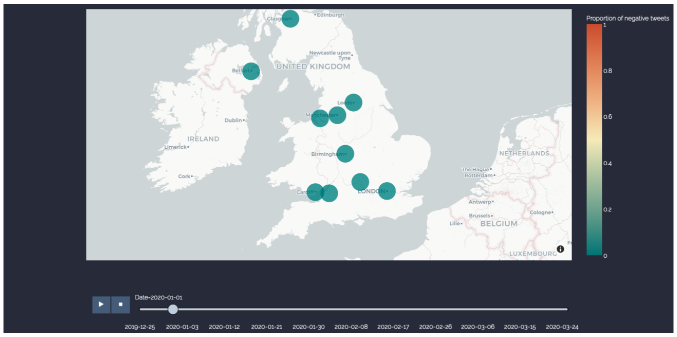


Fig. 7: A map showing the proportion of negative tweets in different UK cities on 01/01/2020.

Although the coefficient between the Twitter and COVID-19 variables is not as high, it is still significant at 0.696. The exponential increase, along with a long preliminary period in which cases were near to zero, may be reasons why the correlation is weaker.

Furthermore, relationships between secondary sources also show signs of strong correlation. Similar instances of the COVID-19 variable being having a slightly weaker correlation is found for the other two secondary sources. Nevertheless, these three independent secondary sources all have a strong to very strong correlation, justifying their credibility as a source of ground truth for negative sentiment related to COVID-19.

With these results, it can be said that the model trained and implemented to perform the task of sentiment analysis performs favourably. Its predictions strongly align with the gathered secondary sources of truth, and thus, verifies its capabilities of inferring sentiment from text. Note that this is not a fully conclusive view of the model, and that there are underlying assumptions made when assessing the model. For example, the quality of the tweets are not evaluated (e.g. bots, tweets unrepresentative of the population). Also, the assessment of the model is focused particularly on its classification of negative sentiment only. Thus, further investigation would be required in order to thoroughly assess the performance of the model.

C. Tweet Sentiment Based On Geographical Location

In order to analyse the sentiment of tweets based on geographical location, the data as shown in Figure 1 after being classified by the model is plotted on a map. A scatter map is used as they allow data points to be plotted over distinct geographical location. Mapbox's scatter map is used via the

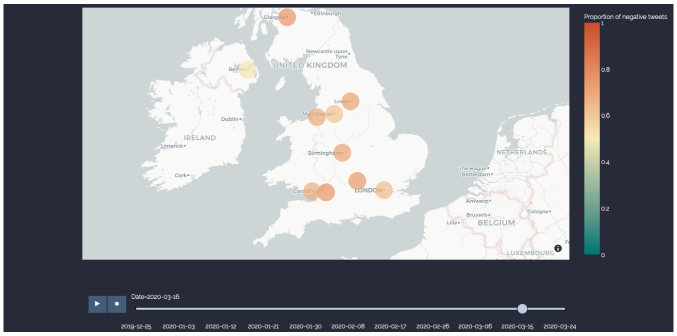


Fig. 8: A map showing the proportion of negative tweets in different UK cities on 16/03/2020.

`Plotly Express` library in `Python`. Additionally, a date slider is added onto the map. The date slider makes it possible to easily change the date of the data plotted on the map, enabling the visualisation of the changes in sentiment over time.

The results derived from the model is plotted as circles over the ten cities where the tweets originated from. The sentiment is visualised by the colour of the circle plotted over a city. A colour scale of green to red, where green denotes a low proportion of negative tweets and red denoting a high proportion of negative tweets, is chosen as it is intuitive to users, according to the Gestalt Law of Isomorphic Correspondence, hence aids in the ease of understanding of the map.

From the visualisation of the data, a few trends of the sentiment from the data collected can be observed. First of all, from Figure 7, showing the proportion of negative tweets on the 1st of January 2020 and Figure 8, likewise but using data from the 16th of March 2020, it is noticeable that the proportion of negative sentiment in tweets for all cities that are tracked has increased significantly over time. Moreover, from the 10th of March 2020 onwards, all cities recorded at least a 40% proportion of negative tweets for every subsequent day. This is contrary to the previous period where the sentiment fluctuates daily for most cities.

The second trend observed is that a spike of negative sentiment first occurs in the South of England and in Scotland, with Bristol being the first city to exhibit more than 50% of tweets being negative for the day, followed by Glasgow and London two days later. Northern Ireland, the midlands and the north of England did not record a significant spike in negative sentiment until a week later.

Third of all, there seems to be some correlation between geographical distance and the proportion of negative sentiment of tweets from the cities. When the proportion of negative tweets sporadically spikes, it seems cities in the south such as Bristol and London would observe similar patterns. Similarly up north, Leeds, Manchester and Liverpool would usually see a spike of proportion of negative tweets around the same time. However, there are outliers to this trend. Cardiff and Bristol, despite being geographically close, do not seem to record spikes of proportion of negative tweets at similar times. Similarly, Oxford and London do not seem to show the same correlation. It is important to note that although there are correlations, this does not suggest causation of any form. These patterns could merely be coincidences.

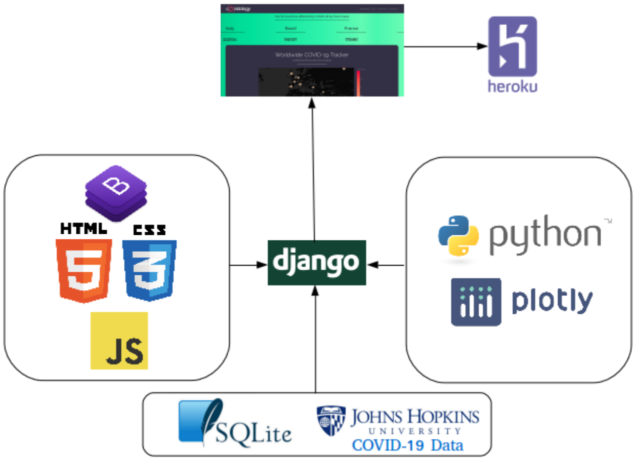


Fig. 9: Schematic diagram of our web application. The left-hand side represents the client-side technologies, the right-hand side our back-end technologies, and the bottom side the data source and data-storing tools utilised. Each of these interact with the Django web framework to then produce our Django web app. This is then deployed and hosted online using Heroku.

VII. Dashboard

Django makes effective use of an MVT (Model-View-Template) design pattern that allows us to build interactive implementations of our findings. Django Models are used to interface with the existing SQLite database. This integrates the database into the web application, and allows for much easier querying than compared to use of SQL syntax. To contextualise, our Tweet model is structured in a way that allows us to build functionality that can be rendered into a 'view'. The interaction views has with our processed logic and models are then rendered into a template This has ranged from filtered example tweets from our Tweet model collection, to our Plotly graphs. The rendered templates are then accessible and displayable on the site with the rest of the client-side data.

This is so we improve the possibility that we can garner a stronger understanding of the information we have at our disposal. The dashboard contains all the different graphical representations from our Tweet model produced by Plotly as well as our John Hopkins live map data. The upper half of the dashboard holds all map related data such as the sentiment heat map on Figure 8 and the COVID-19 World Map Tracker on Figure 10 whereas the lower half holds all graphical data generated by Plotly from our Twitter sentiment data. It is important to note that the aim is to provide a user to observe and interact with the data from different perspectives with each visualisation. This is so we improve the possibility that we can garner a stronger understanding of the information we have at our disposal.

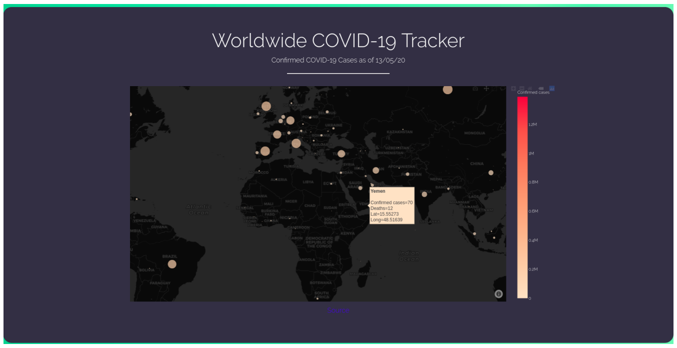
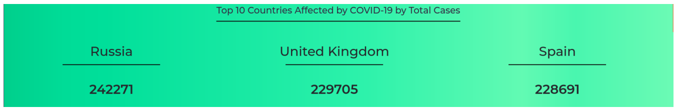


Fig. 10: World map showing the extent to which countries are affected by COVID-19. Hovering over a particular bubble allows the user to see more details about that particular country such as its confirmed cases and total deaths.



Top 10 Countries Affected by COVID-19 by Total Cases

Russia	United Kingdom	Spain
242271	229705	228691

Fig. 11: Auto-scroller that infinitely scrolls left to illustrate the top 10 countries affected by COVID-19 by the Total Number of Cases.

The main task for the dashboard is for it contextualise our findings in an intuitive and effective manner without appearing convoluted or misleading. One approach taken for this was the usage of an automatic scroll-bar. Whilst the World Tracker map illustrates the larger picture by providing a contrast across all countries by colour and bubble size, the scroll-bar directly picks a primary metric, the total number of cases, and illustrates the top 10 countries currently affected. This visualisation is beneficial in its simplicity as it directly outlines a characteristic from our John Hopkins data in a way a user would easily understand.

The dashboard is deployed using Heroku. It is routed through CloudFlare to enable HTTPS. The dashboard is publicly available to view at `http://www.covidology.uk`. It holds interactive versions of all the visualisations used as part of this report, along with the COVID-19 World Map Tracker described in this section.

VIII. Conclusion

The aim of the project was ultimately to gather relevant data, infer trends from the data and display the data clearly. In that respect the project has been a success as ultimately the sentiment of tweets based on their geographic location has been successfully analysed.

While there exists room for improvement, discussed in Section IX, this project provides a good example for where data analysis can be used to discover trends and model how behaviours from various data sources are correlated. Although the data collected from the FTSE-100 cannot be specifically linked to the coronavirus pandemic, using Twitter sentiment analysis to predict the behaviour of the stock market is an area which has been extensively researched [6].

Having collected data from four independent sources, trends in correlation are discovered. It is found that there exists a correlation between the number of COVID-19 cases, the FTSE-100 stock market prices, the number of news articles about COVID-19 and the negativity of public opinion found through the sentiment of tweets. This therefore validates the sentiment analysis model which is then used to look infer the sentiment of tweets around the UK.

The work presented here can be validated by the fact that the data used to train the models was sufficiently large and therefore reflects the general opinion of Twitter users. Furthermore, the model is transferable as the Word2Vec method was used meaning that the classifier is not trained on specific words but a vector representation of word meaning.

The methods used to classify the sentiment of the tweets were also tested on a labelled sub-set of the data: providing an empirical measure of accuracy. This measurement could be improved. Ideally, a labelled set of tweets regarding coronavirus would be used to train the data. If sufficiently large enough, using this data set would improve the accuracy of the classifier. It should be noted that even with this improvement, we would not expect an accuracy of 100%.

A point to consider when assessing the correlation of the variables is the accuracy of the data sets used. The FTSE-100 data is accurate as it portrays the true stock market prices stated by the London Stock Exchange. This is also true with the news articles and COVID-19 data as the data has been pooled from reliable sources and represent the number of negative articles and the number of cases respectively. However, when comparing the trends in the variables with the negative sentiment of the tweets, it must be considered that the negative sentiment of the tweets is not 100% accurate as this variable represents an estimation of sentiment and so correlation measures between these have uncertainty which must be considered.

As previously mentioned in Section IV, the FTSE-100 data obtained does not relay the true behaviour of the stock market for certain days within the time range considered. Therefore, the comparison of the market price with the other three data sets is not appropriate for certain dates as the validity of this comparison is compromised.

An additional point to note is that sentiment analysis is not conducted on the news articles. An assumption is made: all news articles have a negative sentiment. This leads to the assumption that the more news articles regarding COVID-19, the more negative sentiment. This assumption is invalid where there exist news articles which contain a positive sentiment. Sentiment analysis needs to be conducted on the news articles also to improve the model.

IX. Future Work

When analysing each of the classification techniques, it was found that the evaluation metrics found were relatively low. It is therefore necessary to find a method which can capture the sentiment of the tweets more accurately.

Another way in which this modelling could be improved is to have a larger set of manually labelled testing tweets. The models were evaluated using 300 tweets which were manually labelled. This is a relatively small number of tweets to use. A larger number of manually labelled tweets need to be used in order to assess how well the models can infer the sentiment of the tweets. Another issue with labelling the tweets in this way is that the sentiment of the tweet inferred through human inspection can differ.

This project has focused on Twitter as the social media platform to gather data about the public's opinions and views on COVID-19. Whilst a lot of information is able to be gathered using this platform, a bigger pool of information using other types of social media platforms could be necessary in order to infer a bigger picture on public opinion. However, with this comes a greater problem with privacy and ethics which would need to be considered.

Another point to consider in future work is the data collected from news articles. This information was sourced from the Guardian. Future work could consider a wider range of news sources as a wider range of opinion would then be captured in this data set.

The time period over which data was collected is considerably short. The point at which there is a significant rise in sentiment regarding COVID-19 is monitored and identified, data is then collected from this point. To improve, it would be advantageous to the accuracy of the project if data was collected over a longer time period- monitoring how the sentiment progresses as it becomes sustained.

References

[1] Al Jazeera. Timeline: How the new coronavirus spread. https://www.aljazeera.com/news/2020/01/timeline-china-coronavirus-spread-200126061554884.html.

[2] Benesty, J., Chen, J., Huang, Y., and Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[3] Forman, G. BNS feature scaling: an improved representation over tf-idf for svm text classification. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), pp. 263–270.

[4] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford 1*, 12 (2009), 2009.

[5] Johns Hopkins University Center for Systems Science and Engineering. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data. https://github.com/CSSEGISandData/COVID-19.

[6] Mittal, A., and Goel, A. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf) 15* (2012).

[7] ShareCast. Market Data: FTSE100. https://www.sharecast.com/index/FTSE_100.

[8] TWINT Project. TWINT. https://github.com/twintproject/twint.