An Analytical Framework for Measuring Inequality in the Public Opinions on Policing –

Assessing the impacts of COVID-19 Pandemic using Twitter Data

Adepeju Monsuru[1], Fatai Jimoh[2]

[1]Crime and Well-Being Big Data Centre, Manchester Metropolitan University, United Kingdom
[2]School of Science, Engineering and Environment, University of Salford, United Kingdom.
Email: m.adepeju@mmu.ac.uk

Keywords: Inequality, Policing, COVID-19 Pandemic, Sentiment Analysis, Visualization.

Abstract

As the COVID-19 pandemic sweeps across the globe, police forces are charged with new roles as they engage and enforce new policies and laws governing societal behaviours. However, how the police exercise these powers are an important factor in shaping public opinion and confidence concerning their activities across space and time. This research developed an analytical framework for measuring the inequality in the public opinion towards policing efforts using Twitter data. We demonstrate the utility of our framework using 3-months of tweets across 42 police force areas (PFAs) of England and Wales (UK). The results reveal that public opinions on policing is overwhelmingly negative across space and time, and that these opinions have been most exacerbated by the COVID-19 pandemic in three specific PFAs, namely Staffordshire, Thames Valley, and North Wales. We provided the link to the open-source script by which this research could be replicated and adapted to other study areas. This research has the potential to help law enforcement understand the dynamics of public confidence and trust in policing and facilitate action towards improved police services.

1. Introduction

For decades, the process of measuring outcomes of policing efforts – how those efforts have impacted public trust and confidence in the police - have depended largely on traditional data acquisition techniques, such as surveys and interviews [1] [2] [3]. However, the recent advent of social media systems, such as the Twitter, has not only heralded enormous data opportunities, but also new advances in the opinion mining of natural language texts. Because a key function of social media is to allow people to share their views and sentiments more widely, opinion mining is right at the centre of research and the application of social media itself [4]. Opinion mining is the technique of extracting sentiment from social media data using computational methods. The technique has gained growing interest across a wide range of application domains, including law enforcement [5] [6]. The technique mainly focusses on sentiments that express or imply positive or negative views. In this study, we introduce an analytical framework, based on an opinion mining technique, which allows the inequality in public opinions concerning policing to be measured and monitored systematically during the COVID-19 pandemic.

Through the analysis of publicly available Twitter data, it is possible to identify issues of greatest concern to the public. Since the start of 2020, the COVID-19 pandemic is the most consequential issue to the general public, as well as to many organisations, including law enforcement. Police forces are having to respond to and assist in a public health crisis, enforcing new regulations and by-laws in order to help manage the spread of the pandemic [7]. Although only a small proportion of citizens have direct face-to-face contact with a police officer [1], many citizens, may have gained certain opinions concerning police activities during the pandemic. Social media systems such as Twitter serve as platforms by which such opinions can be made known to the public, often with a specific

hashtag to indicate the context of the post [8] [9]. Through the analysis of this information, it is possible to measure the impact of the context on the subject matter. Yet, no studies have examined how the COVID-19 pandemic may have exacerbated or decelerated the orientation of public opinions concerning the police and/or policing in space and time. Addressing this research gap is the first major contribution of our study.

To date, most studies focussing on the analysis of public opinions on policing have examined the study area as a whole, rather than different local subdivisions of an area. To many police forces, understanding how different local areas perceive police operations is crucial for evaluation purposes. Previous attempts to remedy this research gap used geo-tagged tweets[1] [10] [11] in order to identify different local areas in which the tweets originate. However, the percentage of geo-tagged tweets within a stream of tweets is estimated to be around 1-2% [12] [13]. This has raised concerns regarding the adequacy and robustness of geo-tagged tweets for any meaningful analysis. We addressed this research challenge in our own study by extracting the location information from the user's profile to geocode the tweets accordingly. We achieved a 92% geocoding accuracy based on this approach, a significant improvement over the 'geo-tag' information. This approach creates a unique opportunity to analyse inequality in public opinions across space using Twitter data.

As public opinion varies geographically, so does it vary temporally [14]. To the best of our knowledge, no studies have examined both the spatial and temporal inequalities in public opinion on policing with respect to the pandemic using the Twitter data. People's opinions on policing is not static, but changes over time. These changes can be measured and monitored across space and time. In this study, we utilize the police force area (PFA) which represents the operational units of police forces in England and Wales as our spatial unit and a monthly time bin as the temporal unit of analysis. Thus, the analysis of public opinions on policing in relation to the pandemic, simultaneously in space and time, is the second major contribution of our study.

An important aspect of opinion analysis is the representation of the results. [15] provides an overview of a wide range of visualization methods that have been employed in previous research. These range from basic tools such as pie or bar charts (used to represent a simple summary for the proportion of positive/negative sentiment) to advance groups involving self-organizing term association maps (used for representing complex multi-dimension geospatial sentiment information). Mostly, the choice of visualization tool depends on the actual aspects of the measured opinion being represented. For example, a basic line graph is effective for time series plot, while sequential geospatial maps are effective for revealing spatial patterning and clustering of opinion across the space. In this study, we employ simple graphical tools, such as the radar charts and sequential geospatial maps.

The structure of this paper is as follows: Firstly, we provide a brief overview of related work, focussing on the opinion analysis, henceforth referred to as 'sentiment analysis', as well as its applications in two relevant fields – law enforcement and the pandemic. We discuss the development of our systematic framework for measuring the inequality in public opinion towards policing, spatially and temporally. We then present the case study, results and discussion sections. We conclude by explaining the significance of our study and plans for future research.

1.1 Aim and Research Questions

The primary aim of this study is to assess the impacts of COVID-19 pandemic (tweets) on the orientation of public opinion concerning policing across England and Wales, over a period of three months. Our research strategy is to develop an analytical framework that will allow the collection of tweets relating to policing, from which the subset

[1] Geo-tagged tweets are tweets in which the user enables the locations information (in form of coordinates) at the instance of the post

on COVID-19 pandemic can be isolated for assessment. Specifically, we plan to answer the following research question:

Q1: What are the orientations of public opinion concerning policing efforts across space over time?

Q2: How has the COVID-19 pandemic impacted the orientations of public opinions in Q1? Are there spatial and temporal patterning and/or clustering to the policing-COVID-19-pandemic interactions in Q2?

2. Related Work

We provide a brief overview of related work in the following section.

2.1 Applications of Sentiment Analysis in Policing and Pandemic

Sentiment analysis is a natural language processing task, which involves the detection of opinion and classification of attitudes in texts [16] [17]. The sentiment analysis of Twitter data has gained widespread interest across a variety of domains. However, some of the most recent applications can be seen in the study of COVID-19 pandemic. For example, [18] showed in their work that tweets regarding COVID-19 could produce a misleading outcome. This is evident in their results, where on one hand the largest proportion of retweets analysed between January 2019 and March 2020 were either neutral or negative, while on the other hand, those analysed between December 2019 and May 2020 showed larger proportion of positive opinions. Other related studies include [19] [20].

In law enforcement, only one paper has examine the COVID-19-crime association, using Twitter data [21]. In their study, [21] employed the qualitative approach called thematic analysis [22] [23] [24], rather than using the sentiment analysis. They showed that most of the law enforcement tweets were not crime-focused, but centred instead on encouraging the public to comply with government guidance about behaviour during the pandemic or concerned general policing. However, their study does not focus specifically on the subject of policing in relation to the pandemic. Therefore, to the best of our knowledge, no study has used Twitter data to examine the policing-COVID-19-pandemic association during the pandemic. In particular, there has not been any studies that examine how the COVID-19 pandemic may have exacerbated or decelerated the orientations of public opinions towards policing based on sentiment analysis. Furthermore, the majority of existing studies have focussed solely on the analysis of the textual components of the tweets, and paid little attention to how sentiments or opinions may vary across smaller regions within a wider study area, over time. In the remainder of this article, we lay out the strategy to fill this research gap in the form of an analytical framework and provide a case study demonstration to highlight the utility of our solution.

3. Developing the Context-based Spatial and Temporal Framework

Figure 1 is the schematic of our analytical framework for measuring and monitoring public opinions concerning policing in relation to the COVID-19 pandemic. The framework consists of three components, namely; the Data Acquisition, the Opinion (or Sentiment) Analytics, and the Sequential Visualization. In the following sub-sections, we give a detailed description of each of these components.

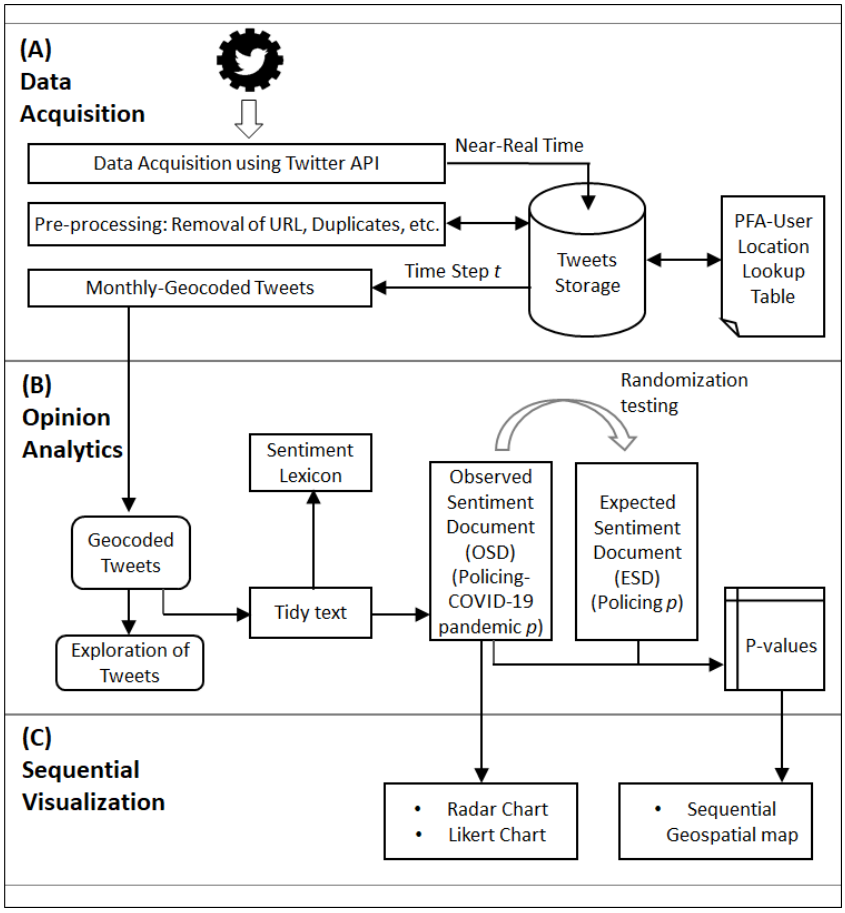

Figure 1. Systematic Framework for measuring public opinion spatially and temporally

3.1. Data Acquisition

(a) Data Download

The Twitter API is utilized to download the publicly available tweets for this study. The API is a programmable tool that provides public access to Twitter data that users have chosen to share with the world. However, the APIs pulls data (tweets) randomly from different locations around the world, leading to a spurious database. We disrupt this default process by restricting the API to a narrow geography. Essentially, we define geographical coverage in the form of a circle from which tweets must originate. This is process is achieved by using the 'search_tweets()' function of the 'rtweet' package in R language [25]. The API is customised to search for tweets that contain any of the specified keywords or the hashtags relating to the police or policing. These keywords include 'police', 'policing', and 'law enforcement(s)'.

(b) Geocoding

Following the data download, we geocoded each tweet to its respective spatial unit of analysis using the user's profile location. The chosen spatial unit of analysis is the actual operational units of police forces in the UK, called the Police Force Areas, henceforth referred to as 'PFAs'. For the geocoding, we created a 'PFA-location-lookup' table, which allow each tweet to be assigned to its respective PFA. The 'PFA-location-lookup' table contains names of all cities, towns and villages across England and Wales. We created this table based on UK Office of National Statistics location gazette [26]. In total, there are 35,604 unique location names in our 'PFA-location-lookup' table.

3.2. Sentiment Analysis

Sentiment analysis is a text mining technique for computationally classifying opinions from a piece of text data into positive or negative sentiments, or some other more nuanced emotion like surprise, fear or disgust. In order to aid easy transfer of data across different data science R packages used, we transformed each tweet document into a tidy format [27]. In our study, we employ the AFINN lexicon [28] which provide a more nuance positive/negative classification by assigning a sentiment score indicating the degree of the sentiment orientation. The scores range from 5 (extremely positive) to -5 (extremely negative). The AFINN lexicon is used as oppose to 'BING' lexicon [29], which gives an outright positive/negative classification, because the nuances provided by the former add more context to the classification. The final opinion classification (i.e. as a negative or positive sentiment) for a tweet is calculated by adding up all the sentiment scores from the tweet. Also, in order to add more context to our classification, we consider bi-grams (i.e. scoring of two consecutive words) classification in cases where a sentiment word is preceded by a negation word, such as 'not', 'never', 'no', or 'without'. The score of such a sentiment word is the score in the opposite direction of the original word. For example, if the word 'good' which is scored as +3 based on AFINN lexicon is preceded by a negation word, such as 'not' (as in 'not good'), then the sentiment score becomes -3. Those tweets with a net zero score or that contain no sentiment words are considered neutral (non-subjective) and therefore removed from the documents.

(a) Observed Opinion Scores

We define the opinion score (OP) of a geographical unit i as the difference between the sum of all weighted positive tweets and the sum of all weighted negative tweets within the area. This is expressed in Equation 1 as:

$$OP = \frac{\sum w_i \,.\, PS_i - \sum w_i \,.\, NS_i}{\sum w_i} \qquad (1)$$

Where, w_i is the weight assigned to each tweet, e.g. based on the level of re-tweets or favorites, PS_i and NS_i represents positive and negative tweets, respectively. In this study, we ignore the weight i.e. $w_i = 1\ \forall_i$ in order to allow a simplified opinion score. In other words, the final opinion score (OP) of a PFA then becomes the difference between the total number positive and the total number negative tweets. Therefore, the opinion score of a geographical unit is positive if OP has (+) sign, or negative if it has a (-) sign. In our study, the OP therefore represents the measure of public opinion concerning policing at a given time period.

(a) Expected Sentiment Document (ESD)

In order to assess the impacts of any given issue (e.g. the COVID-19 pandemic) on the observed public opinion, there is a need to isolate the effects of that issue from the computed OP score. We develop the idea of 'Expected Sentiment Document (ESD)' for this purpose. Essentially, the ESD replaces the sentiment probability of the words relating to the issue with the corresponding sentiment probabilities derived from the main subject matter i.e. the policing. In doing so, the effects or the contribution of the keywords relating to the issue can be eliminated from OP score. This gives us the 'Expected Sentiment Document (ESD)'. This idea is illustrated in Figure 2. For simplicity, we will refer to the tweets that relate only to policing i.e. contain only the policing keywords) as 'type 1' tweets while the tweets that relate to both policing and the chosen issue, i.e. the COVID-19 pandemic, as 'type 2' tweets.

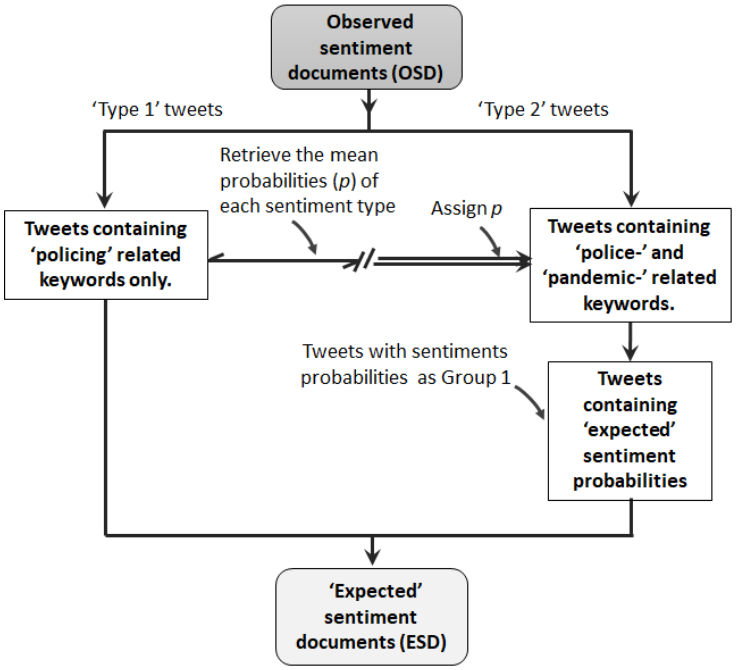

Figure 2. Developing the Expected-Sentiment Document (ESD)

The ESD represents the expectation assuming the pandemic has no impact on the OP score. Any OP score computed based on ESD can be referred to an expected opinion score (i.e. f(E)), while that computed from the OSD can be referred to as an observed score f(O). By comparing f(E) and f(O) the statistical significance of OP can be computed. In order to identify tweets relating to the COVID-19 pandemic, we search for keywords, such as 'pandemic', 'COVID-19', 'Coronavirus', and their variations, within the tweets. Any tweets that include one or a combination of these keywords belong to the type 2 tweets.

(c) Randomization Testing

As illustrated in Figure 2, we derive the statistical significance (p-values) for the observed OP scores. The P-value is required to assess whether an observed OP score is unlikely to be due to chance occurrence. To compute the p-value, we propose a non-parametric strategy based on randomization testing [30] [31]. We ask the question, "If expected opinion scores (i.e. f(E)) were generated under the null hypothesis (H0), how likely would we be to find

a score with scores higher than the observed scores f(O)?". At each PFA, the randomization testing involves generating a large number of ESD, referred here to as "replicas", S, and derive a distribution of expected opinion score $f^*(E)$. A replica is generated by randomizing the sentiment assignment of 'Type 2' tweets based on probabilities derived from 'Type 1' tweets. Given the $f^*(E)$ of a given PFA, the p-value is computed as $p = (S_{beat} + 1)/S_{total} + 1$, where S_{total} is the total number of replicas created, S_{beat} is number of replicas with f* value greater than f(O). As f(O) can be either be greater or less than f(E), we constructed a two-tailed distribution, allowing us to make the judgement as to whether type 2 document have significantly impacted the observed public opinion in either direction. For the randomization testing, the more replicas generated, the more precise the p-value; a typical value would be S = 999. Based on 999 replicas, if, for example, seven of the 999 replicas have higher scores than the f(O), then the p-value of the O is $(7 + 1)/(999 + 1)$ = 0.008. Since the run time is proportional to the number of replicas, a lower number of replications, such as 99, may be recommended.

3.3 Sequential Visualization

In order to select the visualization tools to represent our results, we consider how the spatial and the temporal information will be represented in a clear fashion. Therefore, we chose a sequential visualization strategy, meaning that the results of each time step is visualized separately. We observed that representing a geospatial map, for example, separately for each time step, produces a clearer and easy-to-read information, compared to using complex representation, such as a 3D map. That said, in visualizing the observed opinion scores concerning only policing, we combined radar charts across multiple time steps in order to aid the comparison. We employ the sequential visualization approach to produce the likert charts and geospatial maps that show the relationship between policing and the COVID-19 pandemic.

3.4 Reproducibility of Research

The entire source codes used to perform this analysis have been provided as a supplementary material to this article. The source code is in R language and is also available online as an Rmarkdown file in https://github.com/MAnalytics/JGIS_Policing_COVID-19

4. A Case Study of PFAs in England and Wales

We present the case study of PFAs in England and Wales aimed at demonstrating the utility of our analytical framework. We complete our demonstration under the following headings, (a) Study area and Data exploration, (b) Data analysis and (c) Results. We now provide details as follows:

4.1 Study area and Data exploration

Our study area is the geographical areas of 'England and Wales' - a legal jurisdiction covering two of the four constituent countries of the United Kingdom. 'England and Wales' comprises nine arbitrary policing regions, further subdivided into 43 police force areas (PFAs). The map in Figure 3 shows the spatial locations of the PFAs within their respective regions, shown in different colours. In our study, we consider 42 PFAs having merged 'City of London' and 'London Metropolitan' PFAs together due to their overlapping boundaries. It can be observed that the number of PFAs vary across the policing regions, with the 'North East' having the lowest number (three PFAs), while both 'Eastern' and the 'South East' regions have the highest number of PFAs, at six each. According to the Crime and Disorder Act of 1998 the PFAs are expected to work together to develop and implement strategies to protect their respective local communities.

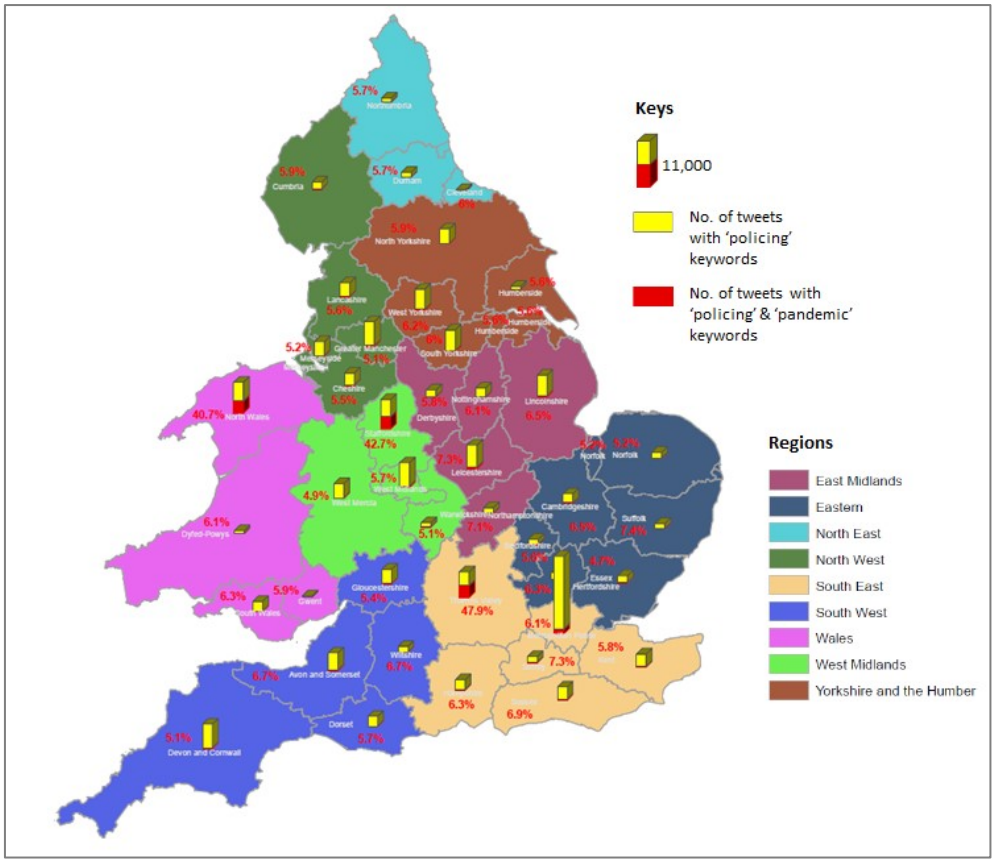

Figure 3. Map showing boundaries of policing regions and police force areas (PFAs) across England and Wales. The bars show the relative volume of tweets (after cleaning) for each PFA over our study period (i.e. from October 20, 2020 to January 20, 2020 – 3 months).

For this study, we downloaded the publicly available tweets relating to the police or policing from October 20, 2020 to January 20, 2020 (3 months). This time period covers the second and the third national COVID-19 lockdowns across the UK, and therefore, police had increased tasks during the study period. We carried out the data download twice a day (morning and night). Each time, the Twitter API retrieves tweets from the past 7 days to the current time (real-time). We focus only on tweets containing the specified police-related hashtags and/or keywords. This task is followed by data cleaning in which all duplicates and spurious texts, including the punctuations, hashtags, emojis and stop words, are eliminated. We also removed re-tweets, but retained the 'replies' (that contain the keywords). We then geocoded the tweets using our PFA-location lookup table, to achieve a geocoding accuracy of 92%. The stacked histograms in Figure 3 show the total volume of the tweets downloaded per PFA, with the red sub-bar and the percentage values (in red) showing the proportion of tweets containing pandemic-related hashtags or keywords. It is clear that the majority of PFAs have between 5–8% of tweets that focus on policing with respect to the COVID-19 pandemic. Dramatically different from these values are the proportions obtained from Staffordshire, Thames Valley, and North Wales PFAs with 42%, 47.4% and 40%, respectively. From the data exploration, it is unclear what factors are responsible for this sharp differences from the rest of the PFAs.

4.2 Data Analysis

The tweet document were divided based on the selected time steps (bins) for our analysis. The time steps are reiterated below:

- Time Step 1: October 20, 2020 to November 19, 2020 (1 month),

- Time Step 2: November 20, 2020 to December 19, 2020 (1 month), and;
- Time Step 3: December 20, 2020 to January 19, 2021 (1 month).

For each time step, we performed the sentiment analysis using the tweet document to derive the OSD and subsequently the observed opinions (using equation 1) for each PFA. We then performed the statistical testing using the approach described in section 3.2c. We perform 999 replications of each OSD documents for each PFA and for each time step. In all, a total of 42 PFAs x 3 time steps x 999 replicas = 125,874 data simulations were conducted. In order to determine whether an observation is considered significant for a two-tail test, we adopt the convention of 5% level, meaning each side of expected distribution is cut at 2.5% corresponding to a p-value of 0.025.

5. Results

We now explain the results in relation to the set research questions in section 1.1.

Q1: What are the orientations of public opinion concerning policing efforts across space over time?

Figure 5 shows the percentage OP score of each PFA within their respective policing regions using the radar chart. The result of the three time steps are represented using different colours, with light green, green and deep blue, representing the observations at time steps 1, 2 and 3, respectively. The OP score is represented in a way that the values increase outwardly from the center in the positive direction. In other word, the outermost circle represents the maximum opinion score while the innermost circle represents the lowest opinion score in each chart. Given that the opinion scores are all negative across the board, the observations closer to the outer circle are ‘less’ negative compared to the observations closer to the inner circle.

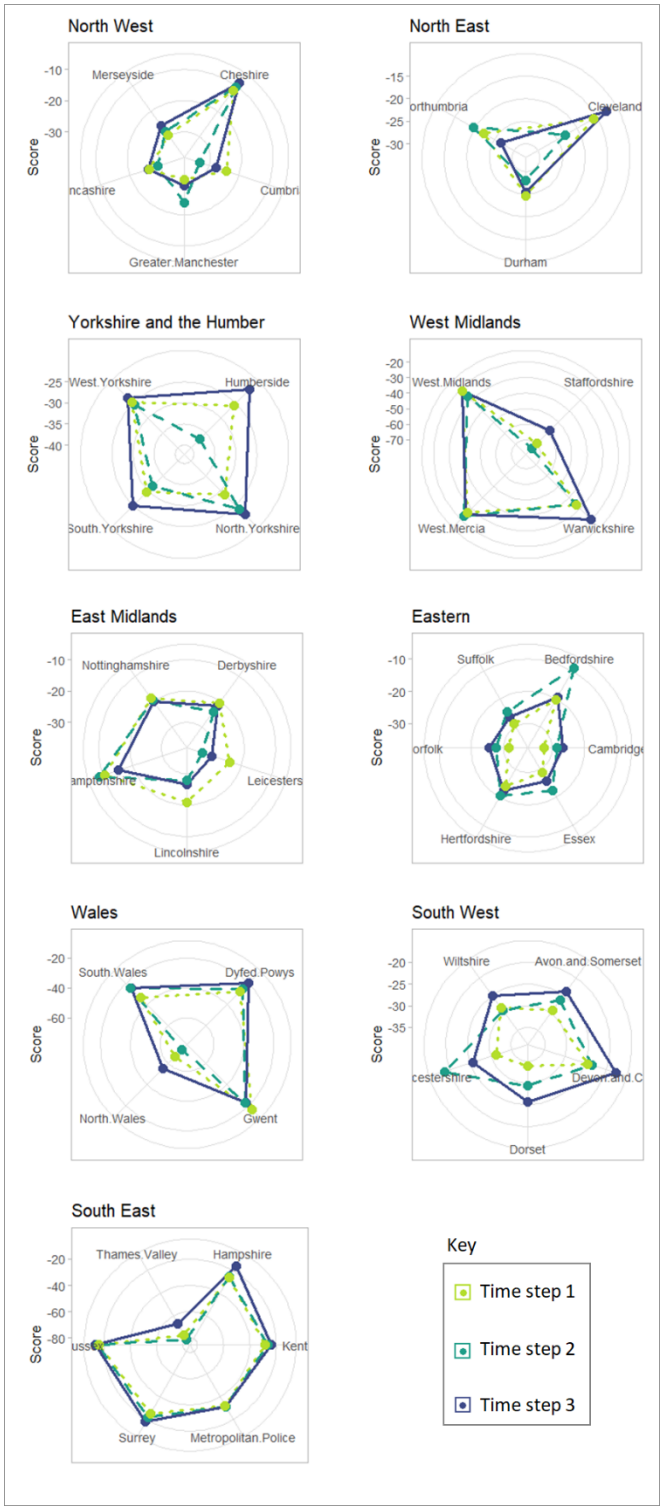

Figure 5: Orientations of public opinions by Regions, PFAs and Time steps

In general, Figure 5 reveal that there is a negative view of policing efforts in England and Wales, across all regio ns and time steps. The regions can be divided into two broad groups according to whether or not the region conta ins an outlier PFA. The region with outlier OP scores are the West Midland, Wales and South East regions, and t he outlier observations are Staffordshire, North Wales, and Thames Valley, respectively. These PFAs are identifi ed as the same three PFAs in Figure 3 with a significantly high volume of tweets with COVID-19 pandemic hash

tags. The outlier effect is also observed to be consistent in these PFAs over the three time steps. These findings i ndicate that the COVID-19 pandemic resulted in a higher negative opinion concerning policing. The second grou p with no outlier provide a clearer indication that the opinions could fluctuate dramatically from one time step to another. For example, the Humberside PFA in the Yorkshire and the Humber policing region shows a moderate negative opinion in time step 1, which rose by approximately 80% in time step 2, which then dropped to the low est negative opinion in time step 3 by 40%. The peak exhibited in time step 2, which covers most part of Decem ber period and coincided with the second lockdown may be indicative of reactions to policing activities during th is time period. However, a similar level of fluctuation observed in Gloucestershire PFA (South West region), but with time step 2 showing the lowest negative opinions, may be a positive reaction to policing activities during th e same period.

Q2: How has the COVID-19 pandemic impacted the orientations of public opinions in Q1? Are there spatial and temporal patterning and/or clustering to the policing-COVID-19 pandemic interactions in Q2?

We produce Figure 6 and Figure 7 in order to answer Q2. In Figure 6, we rank PFAs in the order of decreasing percentage proportion of type 2 tweets, so as to allow the assessment of the relationship between the sentiments and the COVID-19 pandemic. Starting with the outlier PFAs (top 3 bars) previously identified in the answers to Q1, we can see clearly that the opinions of the type 2 portions of the bars are overwhelmingly negative (around 95%) at each time step. The combination of type 1 and type 2 tweets together produce a much higher negative opinion compared with only type 1 tweets. For example, the opinion score of Thames Valley PFA at time step 1 is estimated as -76.5 combining both type 1 and type 2 tweets, an increase of 102% when compared with the score estimated using only type 1 tweets. Similar results are also obtained in Staffordshire and North Wales PFAs. The COVID-19 pandemic appears to have heightened the level of negative opinion in these three PFAs.

The remaining 39 PFAs have a relatively lower proportion of type 2 tweets. The proportions are slightly higher in time step 3 across all PFAs with around 8-10% compared to time step 1 and 2, which are around 5-8%. The potential impacts of type 2 tweets in these cases may not be readily apparent compared to the three outlier PFA above. However, the statistical testing proposed should tell us whether or not the impact is statistically significant.

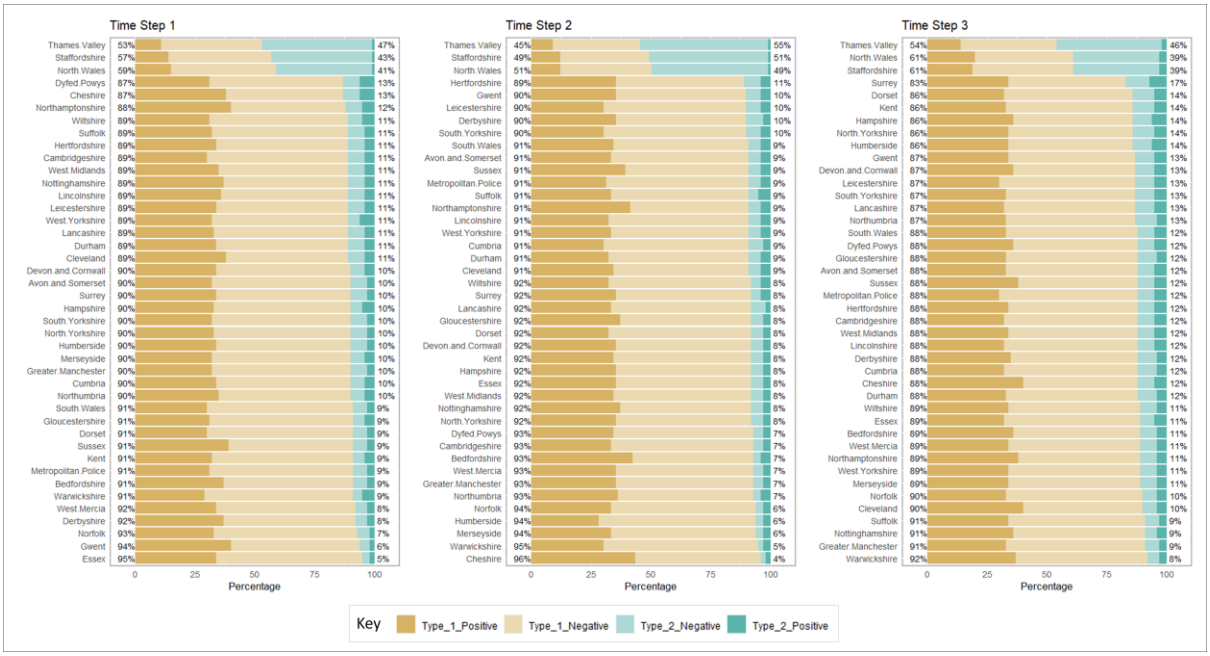

Figure 6. Proportion of tweet types and sentiments per PFA. The brown and light brown sub-bars, represent type 1 tweets with positive and negative sentiments, respectively, while the green and light green sub-bars represent type 2 tweets with positive and negative sentiments, respectively.

Figure 7 shows the results of significance testing represented spatially. We attempt to answer the question; "Given type 1 tweets as the expectation, how likely would we be to find an OP score higher than the ones derived from the type 2 tweets?". In other words, if the public opinion solely about policing is considered to be the expected opinion, how different statistically is the opinion expressed in relation to the 'COVID-19-pandemic? The red and the light red shades (in Figure 7) represent the significant 'lower-than-expectation' OP scores at p-value ≤ 0.001 and at p-value ≤ 0.025, respectively (note that 'lower-than-expectation' of a negative opinion means a higher negative score). On the other hand, the blue and light blue shades represent 'higher-than-expectation' OP scores at the corresponding p-values, respectively. Transparent polygons represent non-significant OP scores. In the supplementary materials, we provide tables showing the numerical representation of the results of the analysis. These tables include the 'Observation' tables, showing the computed OP scores across PFAs and time steps, the 'P-value' tables, showing the statistical significant values based on 999 replications, and the 'Position' table with value 'TRUE' if the observed OP score is greater than the mean expectation, and "0" if less or equal to the mean expectation. These three tables are combined in order to produce Figure 7 (see details in the source code).

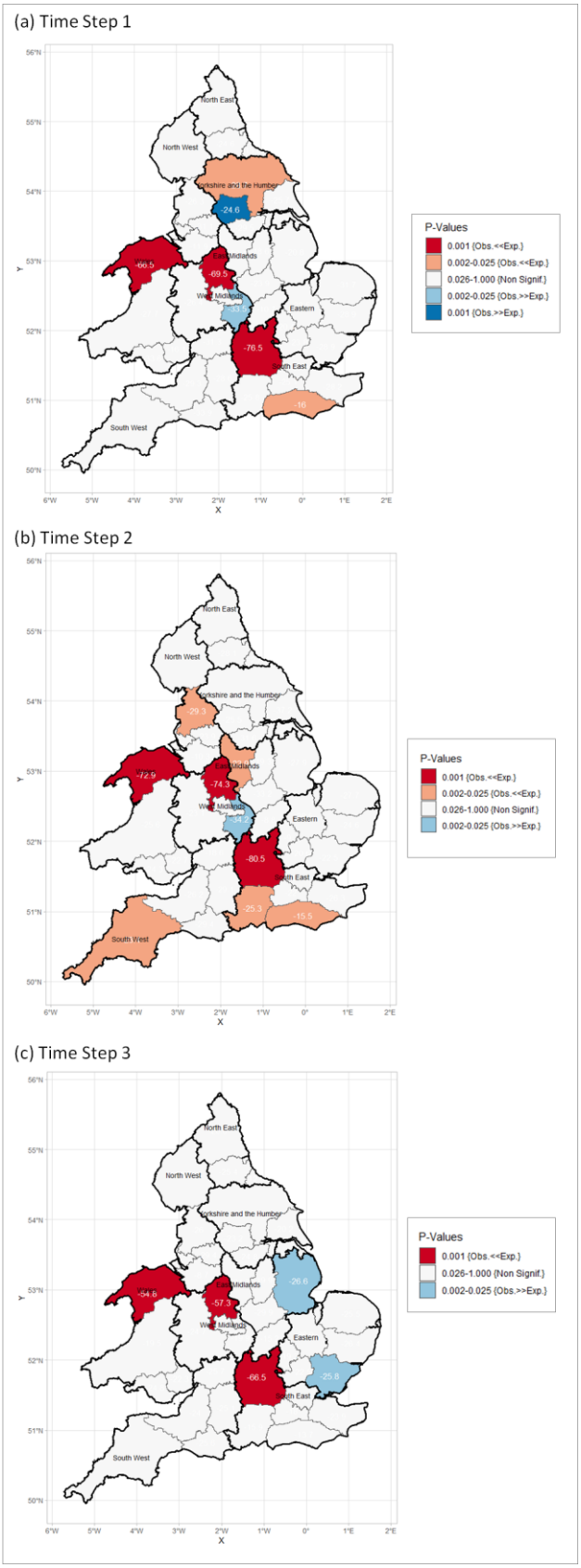

Figure 7. Spatial representation of opinion significance. The regular and the bold lines represent the boundary of PFAs and policing regions, respectively. The value labels within each PFA are the observed OP scores.

It can be observed that whilst the majority of PFAs show non-significant impacts of the COVID-19 pandemic

(tweets), there are a number of PFAs that show statistically significant impacts, with varying levels of stability over time. Again, we can identify the three PFAs, Staffordshire, Thames Valley, and North Wales, which exhibit 'lower-than-the-expected' OP scores at each time step, with statistical significance level of p-value ≤ 0.001. Here, the observed levels of significance are attributable to the high proportion of the pandemic-related tweets (> 40%) which carry more than 85% negative sentiment (see Figure 6). Spatially, Staffordshire, Thames Valley, and North Wales are located in three adjacent policing regions, but the PFAs themselves are not contiguous to each other. Therefore, the result is unlikely to be attributable to spatial autocorrelation effects.

The other significant PFAs exhibit non-stable significance over time. In other word, the PFAs only show significant OP at only one or two time steps. So, a PFA may show significant opinion at a certain time, but become non-significant at another time step. An example of this is the 'West Yorkshire' PFA located within 'Yorkshire and the Humber' region, which shows 'higher-than-expectation' OP score (p-value ≤ 0.001) at time step 1, then became non-significance in time step 2. Spatially, it can be observed that the PFAs in the Midland region tend to exhibit some clustering compared to other parts of the study area. The spatial clustering is more apparent in time step 2 with multiple contiguous PFAs, which run from the Southern regions up to the Midlands areas. There are only few cases of significant contiguous PFAs which also belong to the same policing regions. These categories of PFAs may be useful operationally when implementing interventions to address negative public opinions.

6. Discussion

This study addressed research challenges relating to Twitter data usage, methodology and the application of opinion or sentiment analysis. Firstly, the data usage challenge involved the lack of adequate geographical information (geotags) in openly available Twitter data, which would allow accurate geo-referencing of tweets to their respective local geographical area. To date, existing studies have relied on the use of geo-tagged tweets, which is estimated to be around 1-2% of the total downloadable volume at a given point in time. We argue that this sample size may not be robust enough to achieve reasonable accuracy, especially in real life applications. We addressed this challenge by proposing the use of users' profile location, linked to a database of location names created specially for this study. Based on this database (lookup table), we were able to accurately geocode 92% of tweets. Although there are often slight differences between a user's profile location and the geo-tag location, these differences are generally minimized when the spatial unit of analysis is sufficiently large, such as the geographical extent of the PFA spatial unit that we employed in this study. In England and Wales, the jurisdiction of a PFA generally extends across multiple town and cities.

Secondly, we developed a systematic approach by which the impact of an underlying issue can be assessed on a subject matter. That is, given a subject of interest, say A, how can we test whether another subject (or issue), say B, has impacted the observed opinion concerning A in a (statistically) significant fashion. This idea has never been implemented in previous studies using sentiment analysis. Further, in order to determine the statistical significance of such impact, we proposed a method of randomization testing through which we computed the p-values of an opinion score calculated for each geographical unit. These solutions are integrated as an analytical framework for measuring and monitoring the inequality in public opinion across a geographical area. Based on this framework, it is possible to assess the impacts of any other underlying subject, say C, D, and so on, on the subject matter A. Fellow researchers or practitioners only need to identify the hashtags or keywords relating to a new underlying subject, and follow the steps incorporated in the framework. As a demonstration, we employed 'Policing' as the subject matter, A, and 'COVID-19 pandemic' as the underlying subject.

Using the new analytical framework, we presented a case study in which Twitter data collected over a three month period, across 42 police force areas (PFA) in England and Wales, was used to assess the impact of the COVID-

19 pandemic on the public opinion towards policing. The results reveal that the public opinion towards policing is overwhelmingly negative across space and time, and that these negative opinions have been exacerbated significantly by the COVID-19 pandemic, particularly in three specific PFAs, namely Staffordshire, Thames Valley, and North Wales. These PFAs show statistically significant 'lower-than-expected' public opinions over time, the results of which are attributable to a large volume of highly negative COVID-19 tweets. We see evidence of spatial and temporal inequality of these significant public opinions across the study area. We observed a narrow spatial clustering of significant PFAs. However, we did not attribute this patterning to spatial autocorrelation due to the co-adjacency of PFAs that exhibit 'higher-than-expected' and 'lower-than-expected' significant opinion scores. In summary, it appears that public tweets about the COVID-19 pandemic have resulted in a more or less negative opinion in regions across England and Wales.

To the best of our knowledge, this study is the first to apply sentiment analysis to the examination of the policing-COVID-19 pandemic association using Twitter data. Only one study has examined the COVID-19-crime association using Twitter data [21]. However, they employ thematic analysis, rather than sentiment analysis. Hence, the combination of both policing and the COVID-19 pandemic using sentiment analysis of Twitter data is the application challenge that we addressed in this study.

In order to facilitate the uptake of our analytical framework in a real operational environment, we provide the open-source code that will allow any potential user to reproduce the analysis in its entirety. We provide the link to the source codes. The open-source code also create the opportunity for others to adapt this study to other areas.

Lastly, whilst sentiment analysis is increasingly being used across a wide range of domains, researchers have criticized the technique for missing the deeper context or meaning of communications [32]. Although we address this criticism partly by calculating the net opinion score of a tweet using the varying degrees of sentiments inherent in the tweet, as well as ensuring that 'negated' sentiments were captured effectively, the efficacy of these innovations will be assessed in our future work. Furthermore, the assumption that tweets that explicitly reference the COVID-19 pandemic are distinct in context from those that do not may be imperfect. To capture the intention behind a post, even with a manual classification, remains a challenge.

7. Conclusion

The aim of this study is to assess the impacts of COVID-19 pandemic (tweets) on the orientation of public opinion concerning policing across space and time. We achieve this aim by developing an analytical framework that deploys sentiment analysis technique for the purpose of extracting expressed opinions from Twitter data and allows a systematic assessment of impacts of subject matters within the tweets on one another. We demonstrated the utility of the analytical framework by assessing how COVID-19 pandemic (tweets) have exacerbated and/or decelerated public opinions towards policing across England and Wales.

A reliable and accessible analytical framework for measuring public opinions concerning policing could aid in the overall assessment of the quality and effectiveness of public policing, as it could help to understand community's needs and priorities, as well as the performance of the police in satisfying those needs. Such an analytical framework would enable: a police analyst to evaluate the police service; police decision makers to be better able to guide their organizations in order to provide maximum value to citizens; and police officers to know what is expected of them [33]. Our proposed analytical framework serves these operational goals.

This research focuses on one area of police effort assessments – public opinion. We were able deploy Twitter data to analyze public opinions concerning policing and their inequalities across space and time. It has been argued that measuring citizens' opinion of police performance is very important and is not a straightforward task [34]. In

as much as police work is complex and multi-dimensional, so is its assessment in the eyes of the public. In our own opinion, new innovations, such as the social media system, e.g. Twitter, serve as an interesting alternative to traditional approaches such as surveys and interviews. While valuable, the general notion of "positive," or "negative" opinion provide nothing more than a general sense of the public's confidence and trust in the police. More specific questions need to be asked in order to understand what it is that the citizens are satisfied or dissatisfied about when it comes to the police service. Answering more complex questions using Twitter data is an area that we intend to focus on in the future. Meanwhile, there is a large-scale annual survey of public opinion carried out to assess current perceptions of the police across PFAs in England and Wales. Our future research plan includes a comparative study between this survey approach and the use of the social media system.

Acknowledgements

We gratefully acknowledge the Economic and Social Research Council (ESRC), who funded the Understanding Inequalities project (Grant Reference ES/P009301/1) through which this research was conducted.

References

[1] Langan, P., Greenfeld, L., Smith, S., Durose, M. and Levin, D. (2001) Contacts Between Police and the Public: Findings from the 1999 National Survey. Bureau of Justice Statistics, Washington, DC., 2001.

[2] Mastrofski, S. (1981) Surveying clients to assess police performance: focussing on the police-citizen encounter. Evaluation Review, 5, 397-408.

[3] Mestre, J. (1992) Community feedback program: twelve years later. Law and Order, 40, 57-60.

[4] Liu, B. (2012) Sentiment analysis and opinion mining, Synthesis lectures on human language technologies, 5, 1-167.

[5] Istia, S. and Purnomo, H. (2018) Sentiment analysis of law enforcement performance using support vector machine and K-nearest neighbor, in In 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), IEEE 2018, Indonesia, 2018.

[6] Hand, L. C. and Ching, B. D. (2019) Maintaining neutrality: A sentiment analysis of police agency Facebook pages before and after a fatal officer-involved shooting of a citizen. Government Information Quarterly, 37, 101420

[7] Laufs, J. and Waseem, Z. (2020) Policing in pandemics: a systematic review and best practices for police response to COVID-19. International Journal of Disaster Risk Reduction, 51, p.101812.

[8] Chukwusa, E., Johnson, H. and Gao, W. (2020) An exploratory analysis of public opinion and sentiments towards COVID-19 pandemic using Twitter data. Research Square.

[9] Xue, J., Chen, J., Chen, C., Hu, R. and Zhu, T. (2020) The Hidden Pandemic of Family Violence During COVID-19: Unsupervised Learning of Tweets," J Med Internet Res, 22, e24361

[10] Jiang, Y. Li, Z. and Ye, X. (2019) Understanding demographic and socioeconomic biases of geotagged twitter users at the county level. Cartography Geographic Inf. Sci., 46, 228-242.

[11] Paul, D., Li, F. , Teja, M., Yu, X. and Frost, R. (2017) Compass: Spatio temporal sentiment analysis of US election what twitter says! In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017.

[12] Malik, M., Lamba, H., Nakos, C. and Pfeffer, J. (2015) Population bias in geotagged tweets. In Proceedings of the International AAAI Conference on Web and Social Media, 2015.

[13] Pavalanathan, U. and Eisenstein, J. (2015) Confounds and consequences in geotagged Twitter data. 2015.

[14] Kelman, H. C. (1961) Processes of Opinion Change. The Public Opinion Quarterly. 1, 57-78.

[15] Paradis, K. K., C. and Kerren, A. (2018) The State of the Art in Sentiment Visualization. Computer Graphics forum, 37, 71-96.

[16] Balahur, A., Mihalcea, R. and Montoyo, A. (2014) Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. Computer Speech & Language, 28, 1-6.

[17] Liu. B. (2015) Sentiment analysis: mining opinions, sentiments, and emotions., Cambridge: Cambridge University Press, 2015.

[18] Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R. and Hassanien, A. E. (2020) Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-A study to show how popularity is affecting accuracy in social media. Appl Soft Comput, 97, 106754.

[19] Xue, J., Chen, J., Chen, C., Zheng, C., Li, S. and Zhu, T. (2020) Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE, 15, e0239441.

[20] Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E. and Samuel, Y. (2020) COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information, 11.

[21] Nikolovska, M., Johnson, S. and Ekblom, P. (2020) "Show this thread": policing, disruption and mobilisation through Twitter. An analysis of UK law enforcement tweeting practices during the Covid-19 pandemic. Crime Science, 9, 20.

[22] Heverin, T. and Zach, L. (2010) Twitter for city police department information sharing. In Proceedings of the American Society for Information Science and Technology, 47, 1-7.

[23] Crump, J. (2011) What are the police doing on Twitter? Social media, the police and the public. Policy & Internet, 3, 1–27.

[24] Lieberman, J. D., Koetzle, D. and Sakiyama, M. (2013) Police departments' use of Facebook: patterns and policy issues. Police quarterly, 16, 438–462.

[25] Kearney, M. W. (2019) rtweet: Collecting and analyzing Twitter data. Journal of Open Source Software, 4, 1829.

[26] Office of National Statistics (2015) Major Towns and Cities (December 2015) Names and Codes in England and Wales. London, 2015.

[27] Silge, J. and Robinson, D. (2016) tidytext: Text mining and analysis using tidy data principles in R. Journal of Open Source Software, 1, 37.

[28] Nielsen, F. (2011) A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In Proc. ESWC-11, 2011.

[29] Hu, M. and Liu, B. (2004) Mining and summarizing customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, 2004.

[30] Fisher, R. A. (1935) The Design of Experiments, New York: Hafner, 1935.

[31] Good, P. (2006) Resampling Methods, 3rd ed., Birkhauser, 2006.

[32] Walsh, J. P. (2019) Social media and border security: Twitter use by migration policing agencies. Policing and Society, 30, 1138-1156.

[33] Moore, M. H., Thacher, D., Dodge, A. and Moore, T. (2002) Recognizing Value in Policing: the Challenge of Measuring Police Performance. Police Executive Research Forum, 2002.

[34] Maslov, A. (2016) Measuring the Performance of the Police: The Perspective of the Public. 2016.