An Analytical Framework for Measuring Variations in Public Opinions on Policing in Space and Time –

An assessment of COVID-19 Pandemic using Twitter Data

**Adepeju Monsuru1, Fatai Jimoh2**

1Crime and Well-Being Big Data Centre, Manchester Metropolitan University, United Kingdom

2Department of \*\*\*, University of Salford, United Kingdom.

Corresponding email: m.adepeju@mmu.ac.uk,

Keywords: Policing, COVID-19 Pandemic, Space-Time Sentiment Analysis, Twitter, 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 important factor in shaping public opinions and confidence concerning their activities, across the space and time. This research developed an analytical framework for measuring the variations in public opinions on policing efforts using Twitter data. We demonstrate the utility of our framework using a 3-month tweets across the 42 police force areas (PFAs) of England and Wales (UK). The results reveal that the public opinions on policing is overwhelmingly negative across space and time, and that these opinions have been exacerbated by the COVID-19 pandemic in three specific PFAs, namely Staffordshire, Thames Valley, and North Wales over time. We provided the link to the open-source code by which this research could be replicated and adapted to other study areas. This research has the potential to help the law enforcement understand the dynamics in public confidence and trust in policing and facilitate actions 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 the traditional data acquisition techniques, such as surveys and interviews (Bondurant, 1991; Langan et al., 2001; Mastrofski, 1981; Mestre, 1992). However, the recent advent of the 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, the opinion mining is right at the centre of research and application of social media itself (Liu, 2012). Opinion mining is the technique for extracting sentiment from social media data using computational methods. The technique has gained growing interest across a wide range of application domains, including the law enforcement (Istia et al. 2018; Istia and Purnomo, 2018; Hand and Ching, 2020). The technique mainly focusses on sentiments that express or imply positive or negative views. In this study, we introduce an analytical framework, based on opinion mining technique, which allows the variations in the public opinions concerning policing to be measure and monitored systematically during the COVID-19 pandemic.

Through the analysis of publicly available Twitter data, it is often possible to begin to identify those issues of greatest concern to the public. Since the start of year 2020, the COVID-19 pandemic is perhaps the most consequential issue to the general public as well as to many organisations, including the law enforcements. The 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 (Laufs and Waseem, 2020). Although only a small proportion of citizens has direct face-to-face contact with a police officer (Langan et al., 2001), many citizens however, may have gained certain opinions concerning police activities during the pandemic. Social media system such as Twitter, serve as platforms by which such opinion can be made known to the public, often with a specific hashtag to indicate the context of the post (Chukwusa et al. 2020; Xue et al. 2020). Through the analysis, it is possible to measure the impacts of the context on the subject matter (\*). Yet, not 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 attempt to remedied this research gap is using geo-tagged tweets[[1]](#footnote-1) (Jiang et al. 2020; Paul et al. 2017) 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% (Malik et al. 2015; Pavalanathan, U. and Eisenstein, 2015). 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 and use them to geocode the tweets accordingly. We achieved a 92% geocoding accuracy based on this approach, a significant improvement over the use ‘geo-tag’ information. This approach create a unique opportunity to analyse the variations in public opinions across the space using Twitter data.

As public opinion vary geographically, so does it vary temporally (Kelman, 1961). To the best of our knowledge, no studies have examined both the spatial and temporal variations in public opinion on policing with respect to the pandemic, using the Twitter data. People opinions on policing is not static, but change over time. These changes can be measured and monitored across the 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 result representation. Kucher (2018) 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 a visualization tool often depends on the actual aspects of the measured opinion to be represented. For example, the basic line graph is effective for time series plot, while sequential geospatial map is effective for revealing spatial patterning and clustering of opinion across the space. In this study, we employ simple graphical tools, such as the radarcharts and sequential geospatial map.

An over-arching aim of our work is to facilitate reproducibility and further adaptation of our research. Hence, we provide link to the open source codes that have been used to perform our analysis in its entirety. Our goal is to allow other academic researchers and police analysts to replicate our work, and customise our code to suit their respective aim and objective. We discuss how our analytical framework could be tailored towards studying the public opinions on policing with respect to any other context. This is the third major contribution of our work.

The structure of this paper is as follows: Firstly, we provide a brief overview of related work, focussing on the opinion analysis, otherwise referred to as the sentiment analysis, its applications in law enforcement and pandemic. We discuss the development of our systematic framework for measuring the spatial and temporal variations in public opinion and its associated statistical testing. Then, we presented the case study, results and discussion sections. We concluded 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 COVI9-19 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 on COVID-19 can be isolated. Specifically, we plan to answer the following research question:

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

Q2: How have the COVID-19 pandemic (references) 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 under the following sections.

2.1 Sentiment Analysis

Sentiment analysis is the natural language processing task, which involves the detection of opinion and classification of attitudes in texts (Balahur et al. 2014). Sentiment analysis has become very popular in social media applications for social science research (Pang and Lee L. 2008; Pak and Paroubek, 2010), especially using the Twitter data (Wang et al. 2011; Zhang et al. 2011; Agarwal et al. 2011; Kouloumpis et al. 2011). In these studies, opinions are classified into positive/negative or positive/negative/neutral. Whilst most studies have employed positive/negative classification (Taboada et al. 2011; Pang & Lee 2004; Guha et al. 2015; Hu and Liu, 2004), the classification can also be any number of point scales depending on the complexity of a task (Jurek et al. 2015; Whitelaw et al. 2005; Koto and Adriani, 2015; Taboada et al. 2011). In order to implement different sentiment classification, various sentiment algorithms were developed (Medhat et al. 2014; Serrano-Guerrero et al. 2015). Medhat et al. (2014) grouped the sentiment analysis into two categories: machine learning and lexicon-based approaches. Generally, machine-learning methods are used to automatically learn opinions or emotions of given texts or features. A variety of machine learning algorithms have been developed (Ye et al. 2009; Rushdi Saleh et al. 2011). Example usage for classifying Twitter data can be found in Wu et al. (2011), Pak et al. (2010), and Xia et al. (2011).

On the other hand, lexicon-based approaches focus on measuring subjectivity in texts using semantic orientation (Osgood et al. 1957). They capture the orientation of opinions and the degrees of the orientation (Taboada et al. 2011). At the core of lexicon-based approaches is the *sentiment lexicons* or *dictionary* which contain list of words with the associated sentiment classification label. The three most commonly used lexicons are the BING (Hu and Liu, 2004; ), AFINN (Nielsen, 2011) and NRC (Mohammad and Turney et al. 2013) lexicons.  The Bing lexicon uses a binary (i.e. polarity) categorization model that sorts words into positive or negative opinions. The AFINN lexicon grades words between -5 and 5, representing the most negative and the most positive sentiments, respectively. On the other hand, the NRC lexicon classify sentiment words into eight categories of emotions, namely; positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust. These lexicons are unigram lexicons, meaning that they are based on a single word classification. Example usage of these lexicon in Twitter data applications can be found in a wide range of studies (\*). In general, lexicon-based approaches have been shown to be less effective than machine learning models (Pang et al., 2002). However, opting for machine learning and ignoring the lexical knowledge in lieu of training data may not be optimal (Dhaoui et al. 2017).

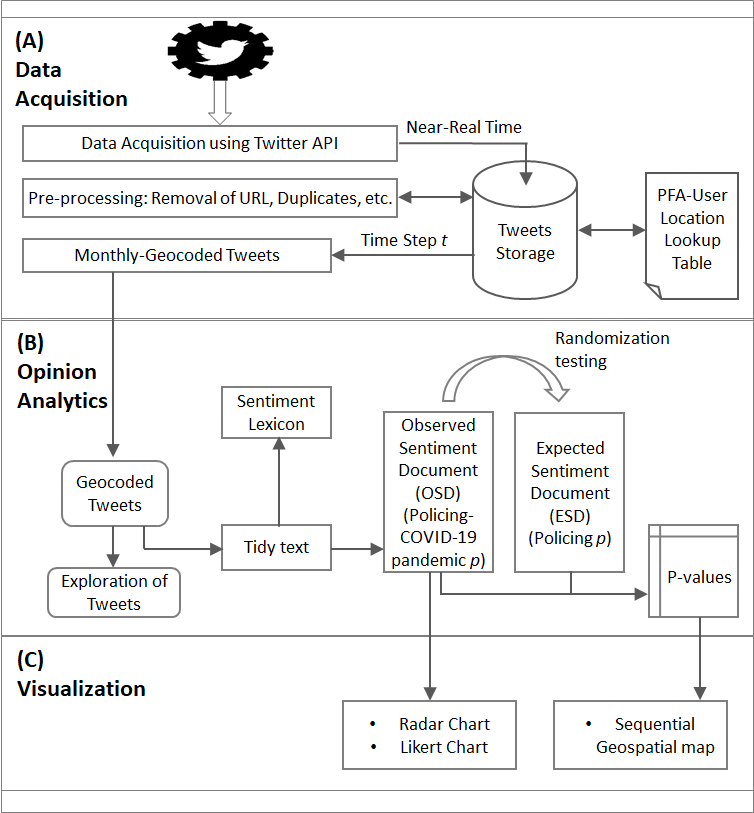
2.2 Applications in Policing and pandemic

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, Chakraborty et al. (2020) showed in their work that tweets regarding covid-19 could produce a misleading outcome. This is evident in their results, where on one side 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 Chakraborty et al., 2020; Xue et al., 2020; Samuel et al., 2020; Kruspe et al., 2020.

In the law enforcement, only one paper has examine the COVID-19-crime association, using Twitter data (Nikolovska et al., 2020). In their study, Nikolovska and colleagues showed that most of the 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 focussed 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 on sentiment analysis that examine how COVID-19 pandemic may have exacerbated or decelerated the orientations of public opinions on the policing. Furthermore, the majority of the existing studies have focussed solely on the analysis of the textual components of the tweets (\*), while paid little attention to how sentiments or opinions may have change 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 Sentiment (or Opinion) Analytics, and the 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 in order to download the publicly available tweets for this study. The API is a programmable tool that provides access to the public 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 spurious database. We disrupt this default process by restrict the API to a narrow geography, leading to a robust database. Essentially, we define a 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. 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 the police forces in the UK, called namely; 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 (ONS 2020). In total, there are 35,604 unique location names in our ‘PFA-location-lookup’ table.

3.2. Sentiment Analysis

The 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 (Silge et al. 2016). In our study, we employ the AFINN lexicon, which provide a more nuance positive/negative classification by assigning a sentiment score indicating the degree of the sentiment. The scores range from 5 (extremely positive) to -5 (extremely negative). The AFINN lexicon is used as oppose to ‘BING’ lexicon, 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 the 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 contains no sentiment words are considered neutral (non-subjective) and therefore 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 (Kuhn, M., 2008). This is expressed in Equation 1 as:

|  |  |
| --- | --- |
|  | (1) |

Where, is the weight assigned to each tweets, e.g. based on the level of re-tweets or favorites, and represents positive and negative tweets, respectively. In this study, we ignore the weight i.e. 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. Different variants of opinion scores can be found in Kuhn, M. (2008). 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.

1. **Expected Sentiment Document (ESD)**

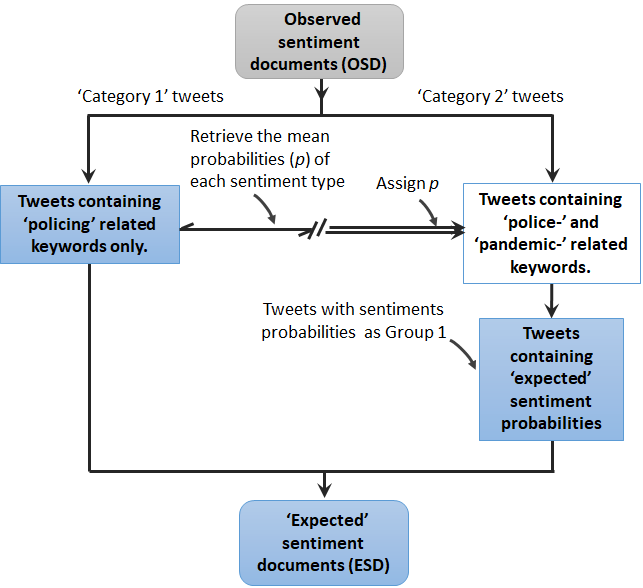
In order to assess the impacts of any given issue (e.g. the COVID-19 pandemic) on the observed public opinion, these is a need to isolate the effects of that issue from the computed OP score. We develop the idea of ‘Expected Sentiment Document (ESD)’ for this purpose. Essentially, the ESD replaces the sentiment probability of the words relating to the issue with the corresponding sentiment probabilities derived from the main subject matter i.e the policing. By so doing, 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 illiustrated in Figure 2. For simplicity, we will refer to the tweets that relate to only policing i.e. contains only the policing-keywords) as ‘category 1’ tweets while the tweets that relate to both policing and the chosen issue, i.e. the COVID-19 pandemic, as ‘category 2’ tweets.

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

The ESD represents the expectation assuming the pandemic has no impacts on the OP score. Any OP score computed based on ESD can be referred to an expected opinion scores (i.e. *f*(E)), while that computed from OP score as observed scores *f*(O). By comparing *f*(E) and *f*(O), we can derive 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 category 2 tweets.

**(c) Randomization Testing**

As illustrate 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 (\*). We simply 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”, , and derive a distribution of expected opinion score . Given the of a given PFA, the *p*-value is computed as , where is the total number of replicas created, 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 category 2 document have significantly impacted the observed public opinion on the either direction. For the randomization testing, the more replicas generated, the more precise the *p*-value; a typical value would be *S* = 999. However, since the run time is proportional to the number of replicas, a lower value (e.g. 99) are often recommended.

3.3 Sequential Visualization

In order to select visualization tools, we consider how the spatial and the temporal information are required to be clearly represented. Therefore, we chose a sequential visualization of results of each time step. This approach will help to visualize and monitor changes in the outcomes more effectively over time. For example, representing a geospatial map for each time slice often produce clearer visualization compared to using a more complex 3D view. Therefore, we will represent our output using three tools, namely; the radar charts, likert chart and geospatial maps for each time step.

3.4 Reproducibility of Research

The entire source codes used to complete this analysis have been provided as a supplementary material of this article. We provided the source code in R language and is also available online at https://github.com/MAnalytics/...

4 Analytical strategy

The analytical strategy used in this study involves three steps, namely (i) Data exploration, (ii) Data analysis and Results visualization. These are described in details in the following subsections.

4.1 Data Exploration

Here, we provide an overview of our study area, the data and its characteristics.

(a) Study Area

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. The ‘England and Wales’ comprises nine policing regions, further subdivided into 43 police force areas (PFAs). The map in Figure 3 shows the policing regions in different colours, with grey outlines showing the boundaries of the PFAs. In this study, we derived 42 PFAs for the area because we merged the ‘City of London’ and ‘London Metropolitan’ PFAs due to overlapping boundary issues. The ‘North East’ region has the lowest number of PFAs with three PFAs, while both ‘Eastern’ and the ‘South East’ regions have the highest number PFAs of six each. According to the Crime and Disorder Act of 1998the PFAs work together to develop and implement strategies to protect their respective local communities. Based on the data download strategy employed in this study, a specified circle defined by a centroid {latitude '53.805, longitude -4.242} and a radius 350miles, covers the entire areas of England and Wales.

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

(b) Data and its Characteristics

We downloaded the publicly available tweets relating to the police or policing from October 20, 2020 to January 20, 2020 (3 months) for our study area. This time period covers the second and the third national COVID-19 lockdowns across the UK. We carried out the data download twice a day (morning and night), downloading the 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. Following the downloads, we cleaned the data by eliminating all duplicates and spurious texts, including the punctuations, hashtags, emojis and stop words. We also removed re-tweets, but retain the replies (that contain the keywords) in addition to the organic tweets on the subjects. Then, we geocoded the tweets using the PFA-location lookup table. Our geocoding strategy is able achieves a 92% geocode of the tweets.

The inserted stacked histograms in Figure 3 show the total volume of the downloaded tweets downloaded per PFA, with the red sub-bar and the percentage values (in red) showing the proportion of tweets containing pandemic-related keywords. These values show that the majority of the PFAs has between 5–8% tweets speaking about policing in relation to the pandemic. The exceptions to these figures are the Staffordshire, Thames Valley, and North Wales PFAs with 42%, 47.4% and 40% police-pandemic tweet volume, respectively. The factors responsible for the sharp difference between these percentage values and those of the remaining PFAs are not readily apparent from the contents of the tweets.

A comprehensive descriptive summary of the data can be found in the supplementary material. The descriptive summary shows the skewness and kurtosis of the number of Twitter users (who generated the tweets). The skewness and kurtosis analysis allow us to examine the sampling of the tweets, and the results confirms absence of outliers, meaning that they are generated by a large number of users and not by a few prolific Twitter users. Then, we examine whether the tweet document follow the Zipf’s distribution (Zipf, 1936, 1946) – the famous frequency distribution expected of a natural language document. By Zipf’s distribution, we expect the frequency of words contained in the document to be inversely proportional to its rank in a frequency table. The distribution is most easily observed by plotting the data on a log-log graph, with the axes being log(rank order) and log(term frequency). Figure 4 is the distribution for our time period 1 dataset (i.e. from October 20, 2020 to November 19, 2020). The distribution plots of the remaining two time periods as well as the plots based on PFAs can be found in the supplementary material.

Figure 4: Zipf’s distribution of our tweet document by policing regions (for Period 1).

Figure 4 shows that the tweet document of each policing region is close to the classic version of Zipf’s law (\*). We see that all the nine regions have word frequency distribution similar to each other, and that the relationship between rank and frequency does have negative slope. The slope of the relationship is however not quite constant, and can be viewed in terms of a broken power law with three sections: the upper, the middle and the lower sections. By fitting a regression line, we can see what the exponent of the power law is for the middle section of the rank range. The deviations we see at high rank are common for many kinds of languages (\*), because a corpus of language often contains fewer rare words than predicted by a single power law. However, the deviations at low rank are very unusual compared with a typical natural language documents. This is a result of the existence of many non-conventional or made-up words in tweets, therefore, affecting the trajectory of the lower section of the distribution. In summary, the tweet documents does not follow the Zipf’s law perfectly, but not close enough to state that the law approximately holds within our document of text.

4.2 Data Analysis

We divided the tweets document based on the selected time steps (bins). The time steps are as follow:

* Time Step 1: October 20, 2020 to November 19, 2020,
* Time Step 2: November 20, 2020 to December 19, 2020, and;
* Time Step 3: December 20, 2020 to January 19, 2021.

For each time step, we performed the sentiment analysis on each document to derive the observed opinions (using equation 1) across the PFAs. This is followed by the statistical testing (as described in section \*) in order to determine if the impacts of the COVID-19 pandemic tweets on the observed opinion scores are statistically significant. This involves generating 999 999 replicas OSD documents for each PFA. In all, a total of 42 PFAs x 3 time steps x 999 replicas = 125,876 data simulation was carried out. 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 = 0.008. In order to determine whether an observation is considered significance 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 discuss the results of our analysis in relation to the set research questions.

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

In Figure 5, we represent the percentage observed opinion of PFAs within the same policing region, using the radar chart. The observed opinions at different 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. In each chart, the observed opinion score increase in value outwardly from the center. In other word, the outermost circle represent the maximum opinion score in the positive direction for a chart. Given that the opinion score 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.

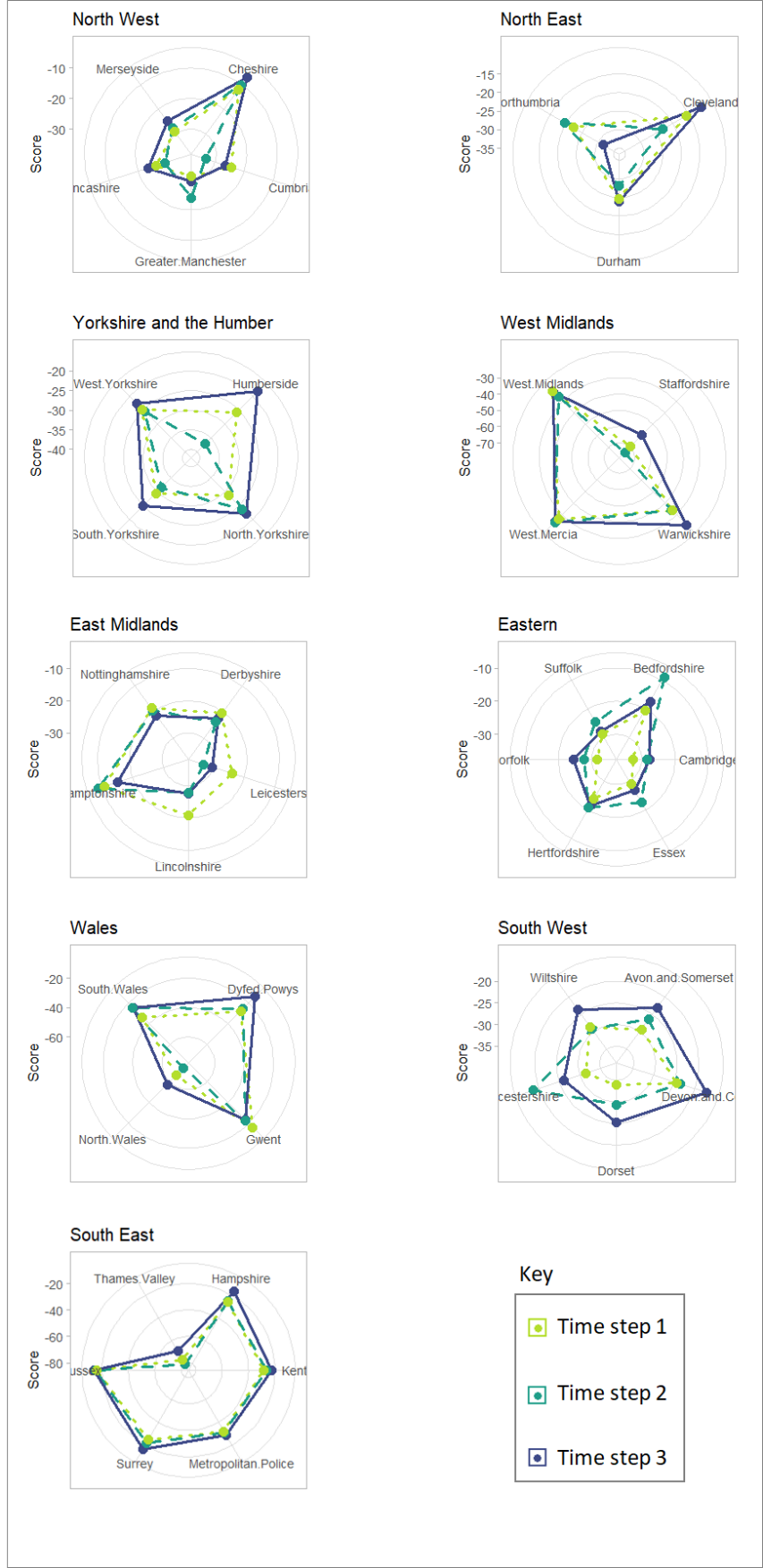


Figure 5: Orientations of public opinions by Regions, PFAs and Time steps

In general, there is a negative view of policing efforts across England and Wales as shown by the negative scores across all regions, all PFAs, and all time steps. We identify two broad group of regions based on the relative orientation of the observed opinion score amongst the PFAs belonging to the same region. These are the ‘similar’ variation group, and the ‘dissimilar’ variation groups. The ‘similar’ variation group include regions in which the relative change in the opinion scores are relatively similar both in size and distribution over time. We can identify three regions belonging to this group, and they are: West Midland, Wales and South East regions. Notably, one PFA in each of these regions is markedly distinct with maximum opinion score (in the negative direction) within the region. These are the Staffordshire, North Wales, and Thames Valley, for the West Midlands, Wales and South East policing regions, respectively. The ‘dissimilar’ variation group consists of the remaining six regions, namely the North West, North East, Yorkshire and Humber, East Midland, Eastern, and South West regions. These regions appear to exhibit no distinct patterns in the relative orientation of the opinion scores amongst the PFAs. However, there are a number interesting individual observations, in terms of their potential relationship with the time period. For example, the Humberside PFA in the Yorkshire and the Humber policing region shows a moderate negative opinion in time step 1, which rose in time step 2 by approximately 80% in time step 2, which then dropped to the lowest negative opinion in time step 3 by 40%. The peak exhibited in time step 2, which covers most part of December period may be indicative of reactions to certain policing activities during the this time period. A similar level of ‘dissimilarity’ can also be observed in Gloucestershire of South West region, but with time step 2 showing the lowest negative opinions.

Q2: How have the COVID-19 pandemic (references) impacted the orientations of public opinions in Q1? Are there spatial and temporal patterning to the observed impacts?

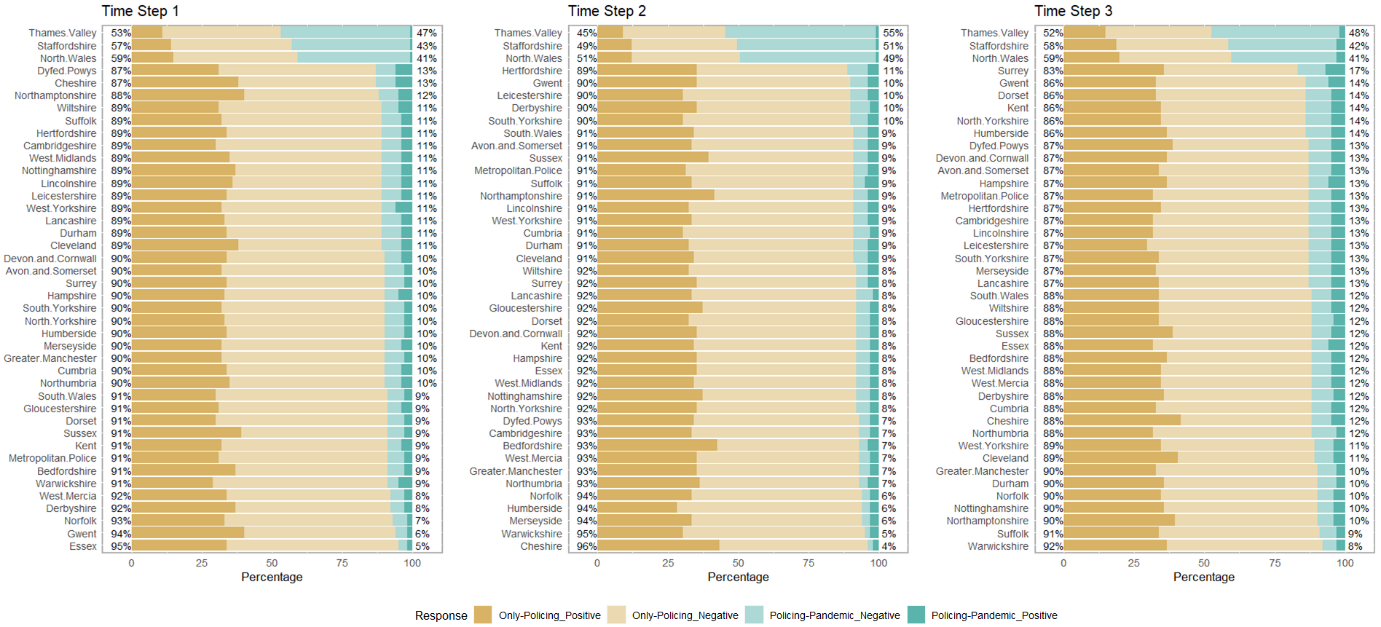
The common characteristic of the distinct PFAs in the ‘similar’ regions described above is that they contain a significantly large amount of tweets with COVID-19 pandemic references. These category of tweets comprise more than 40% of the total volume of tweets in these those PFAs. In Figure 6, we visualize the proportion of the two categories of tweets based on the keyword contents: the ‘category 1 tweets’, containing ‘policing-keywords’ only, and ‘category 2 tweets’ containing ‘both policing and COVID-19 pandemic keywords’. The PFAs are arranged from the top in the order of decreasing size of category 2. The distinct PFAs constitutes the three top bars of the Figure 6. In each of these PFA, category 2 tweets is shown to contain more than 95% negative opinions (indicated by the light green bars) in both time step 1 and 2, and a slightly lower percentage (around 85%) in time step 3.

Figure 6. Proportion of tweet categories per PFA

There is a significantly lower proportion of category 2 tweets in the remaining 39 PFAs. It can be seen that less than 8% of tweets belong to this category. The relative proportion of the negative and positive tweets in both categories appear similar in most cases, except in time step 2 in which category 2 tweets are slightly less than category 1 tweets in terms of the proportional size, as well as the proportion of the negative to positive ratio, across most PFAs. For example, the Humberside PFA only has 6% of category 2 tweets in time step 2, but has 10% and 14% of the same category in time step 1 and time 3, respectively. The negative to positive ratio is estimated as be 1:1 in time step 2, but are estimate as 3:1 and 3:2 in time step 2 and 3, respectively. This implies that amount of conversation that people have as well as their opinion concerning policing may vary from time to time.

Figure 7 shows the spatial representation of the results of significance testing carried out on the tweet documents. With the significance testing, we attempt to answer the question; “*If the opinion scores derived from the category 1 tweets are assumed to be the expectation, how likely would we be to find any areas with opinion scores higher than the ones derived from the category 2 tweets*?”. We formulated a two-tailed test because the observed opinion scores may fall on either side of the mean expectation, and therefore, we should be able to determine if the observation is statistically and significantly higher or lower than the expectation. In Figure 7, the red and the light red shades represent a scenario in which the observed opinion scores is statistically and significantly lower than the expectation at {p-value 0.001} and {p-value }, respectively. On the other hand, the blue and light blue is used if an observed opinion score is statistically and significantly higher, at the same p-value ranges, respectively. Transparent polygons are used to represent the Non-significant observations at {p-value }. The inserted values is the measured opinion score. In the supplementary materials, we provide tables showing the numerical representation of the derived results of the analysis. These tables include the ‘Observation’ tables, showing the observed opinion scores across PFAs and time steps, the ‘P-value’ tables, showing the statistical significant values based on 999 replications, and lastly, the ‘Position’ table that describe the position of an observed score in relation to the mean expectation on the number line. These three tables are combined in order to produce the spatial representation of the opinion significance in Figure 7 (see details in the source code).

In Figure 5, we observed that the public opinions concerning policing across all PFAs and time steps are negative. Figure 6 shows that tweet document for each PFA contain a combination of category 1 (sentiment about policing only) and category 2 (sentiment about policing-pandemic) tweets, with varying ratio of negative sentiment to positive sentiment. Now, Figure 7 tells us whether category 2 tweets have resulted in a statistically significant opinion scores at any given PFA and time step.

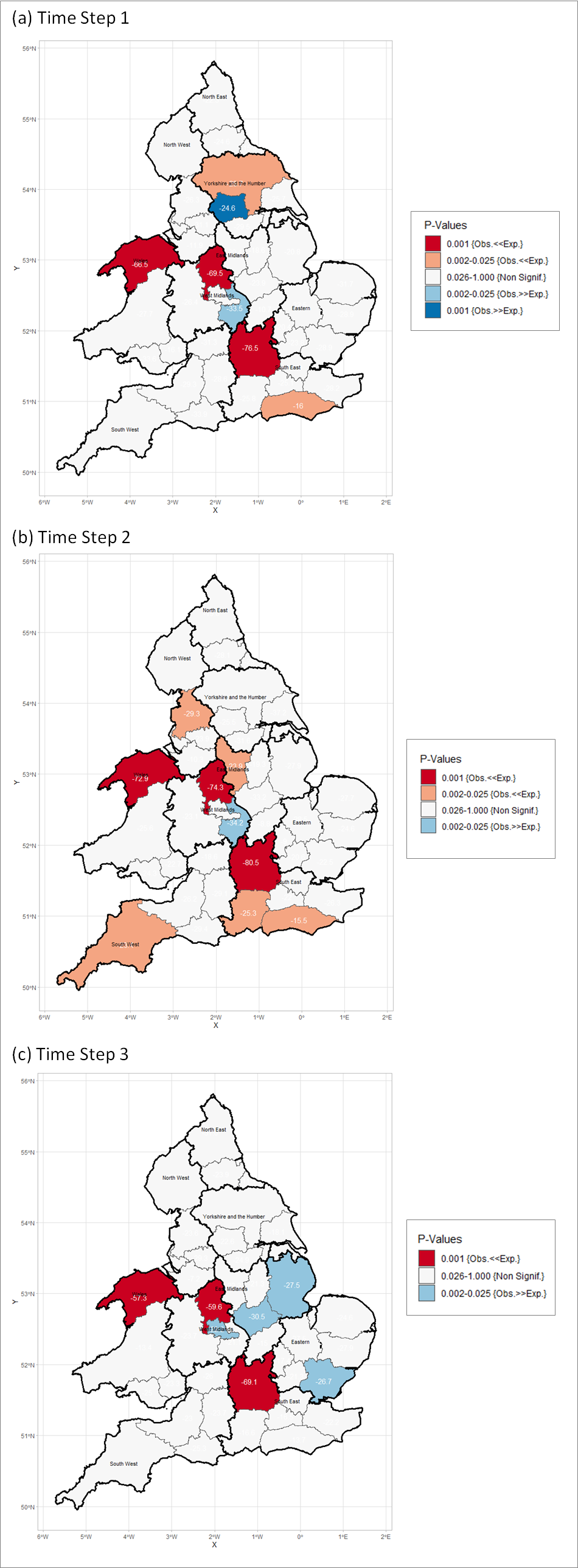


Figure 7: Spatial representation of opinion significance. The regular and the bold lines represent the boundary of PFAs and policing regions, respectively. The value labels within each PFA are the observed opinion scores.

It can be observed that the majority of PFAs show non-significant impacts of the pandemic on the public opinions concerning policing. On the other hand, a number of PFAs show statistically significant impacts, with varying degrees from significantly low impacts to significantly high impacts. We can identify two categories of these PFAs based on the stability of their significance over time. The first category comprises the three distinct PFAs previously identified in the answers to Q1, i.e. the Staffordshire, Thames Valley, and North Wales. These PFAs show ‘lower-than-the-expected’ opinion scores which statistically significant at p-value 0.001 over the three time steps. For example, the Staffordshire has an opinion score of –69.5 as compared with the mean expected score of –32.8 in time step 1. These level of significance can be explained by the high proportion of the pandemic-related tweets (> 40%) exhibited by these PFA, in which more than 85% of them carry a negative sentiment (see Figure 6). Spatially, the Staffordshire, Thames Valley, and North Wales are located in three different policing regions. Although, the regions are contiguous to one another, these PFAs are not geographically adjacent to each other (Figure 7). Therefore, the observed opinions could not have been a result of spatial autocorrelation between them.

The second category of PFAs exhibit unstable significance over time. In other word, the opinions are only significant at only one or two time steps. These category include PFAs that have ‘higher-than-the-expected’ and the ‘lower-than-the-expected’ opinion scores. In time step 1 and 2, there are two and five cases, respectively, of significant ‘lower-than-the-expected’ opinions scores. Amongst these PFAs, only Sussex PFA is significant (at p-value = 0.025) at both time steps. On the other hand, the ‘higher-than-expected’ significant opinion score can be found in the three time steps, with two, one and three cases, for time step 1, 2 and 3, respectively. Also, in this case, only one PFA, i.e. show significant opinion at two consecutive time steps 1 and 2.

Spatially, it can be observed that the PFAs in the Midland region tend to exhibit some forms of clustering compared to any other parts of the study area. The spatial clustering is more apparent in time step 2 in which there are multiple contiguous PFAs which run from the Southern regions up to the Midlands areas. In terms of policing regions, there few cases of contiguous PFAs belonging to the same policing regions having significant opinion scores. The most prominent example could be found in the South East regions in time step 2 in which three of the six PFAs are adjacent to each other and have significant opinion scores.

**6. Discussion**

Although, the factor(s) responsible for disproportional amount of pandemic-related tweets in these areas is not readily apparent from few sample tweets that we studied, sampling and studying few tweets, we can only conclude that a skew distribution is highly likely to produce a significant impacts on the general opinion score.

The is cgvgg

**6. Conclusion**

This implies that in these areas the Twitter conversation about police or policing in relation to the COVID-19 pandemic tend to be more negative than the Twitter conversation that focus on police or policing only

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

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

Istia, S.S. and Purnomo, H.D., 2018, November. Sentiment analysis of law enforcement performance using support vector machine and K-nearest neighbor. In *2018 3rd International Conference on* Information Technology, Information System and Electrical Engineering (ICITISEE) (pp. 84-89). IEEE.

Paul, D., Li, F., Teja, M.K., Yu, X. and Frost, R., 2017, August. 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 (pp. 1585-1594).

Laufs J., Waseem Z. (2020). Policing in pandemics: a systematic review and best practices for police response to COVID-19. Int. J. Disaster Risk Reduct. 51:101812. 10.1016/j.ijdrr.2020.101812 Processes of Opinion change .

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.

Bondurant, E. (1991), “Citizen response questionnaire: a valuable evaluation tool”, The Police Chief, pp. 74-6, November

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

Mastrofski, S. (1981), “Surveying clients to assess police performance: focusing on the police-citizen encounter”, Evaluation Review, Vol. 5 No. 3, pp. 397-408.

Mestre, J. (1992), “Community feedback program: twelve years later”, Law and Order, Vol. 40 No. 10, pp. 57-60.

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(1):101420

S. S. Istia and H. D. Purnomo, "Sentiment Analysis of Law Enforcement Performance Using Support Vector Machine and K-Nearest Neighbor," *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2018, pp. 84-89, doi: 10.1109/ICITISEE.2018.8720969.

Dende, K., 2014. Sentimental Analysis in crime detection: A case study of Kenya law enforcement agencies (Doctoral dissertation, University of Nairobi).

Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T (2020) Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE 15(9): e0239441. <https://doi.org/10.1371/journal.pone.0239441>

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.DOI:   
10.21203/rs.3.rs-33616/v1 (preprint)

Y. Jiang, Z. Li, and X. Ye, (2020) ``Understanding demographic and socioeconomic biases of geotagged twitter users at the county level,'' Cartography Geographic Inf. Sci., vol. 46, no. 3, pp. 228\_242, 2019.

Malik, M., Lamba, H., Nakos, C. and Pfeffer, J., 2015, April. Population bias in geotagged tweets. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 9, No. 1).

Pavalanathan, U. and Eisenstein, J., 2015. Confounds and consequences in geotagged Twitter data. arXiv preprint arXiv:1506.02275.

Kelman, Herbert C. "Processes of Opinion Change." The Public Opinion Quarterly 25, no. 1 (1961): 57-78. Accessed January 9, 2021. <http://www.jstor.org/stable/2746461>.

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

Fong, E., & Chang, L. Y. (2011). Community under stress: Trust, reciprocity, and community collective efcacy during SARS outbreak. Journal of community health, 36(5), 797–810

Pang, B., Lee, L. and Vaithyanathan, S. (2002), “Thumbs up? Sentiment classification using machine learning techniques”, in Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, Stroudsburg, PA, pp. 79-86

Balahur, A., Mihalcea, R., & Montoyo, A. 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. Computer Speech & Language, 28: 1–6.

Pak, A. and Paroubek, P. (2010) "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, 2010, vol. 10, no. 2010, pp. 1320-1326.

Pang, B. and Lee, L. (2008) "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1-135, 2008.

[28] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Language in Social Media (LSM 2011), 2011, pp. 30-38.

[29] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in Fifth International AAAI conference on weblogs and social media, 2011.

[30] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011: ACM, pp. 1031- 1040.

[31] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," HP Laboratories, Technical Report HPL-2011, vol. 89, 2011.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, 37: 267–307.

Pang, B., & Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, 2004. Association for Computational Linguistics, p. 271.

Whitelaw, C., Garg, N., & Argamon, S. 2005. Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM international conference on Information and knowledge management, 2005. ACM, pp. 625–631.

Medhat, W., Hassan, A., & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5: 1093–1113.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. 2015. Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311: 18–38.

Medhat, W., Hassan, A., & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5: 1093–1113.

Ye, Q., Zhang, Z., & Law, R. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, 36: 6527–6535.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. 2011. Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38: 14799–14804.

Osgood, C. E., Suci, G., & Tannenbaum, P. 1957. The measurement of meaning, 81. Paltoglou, G., & Thelwall, M. 2012. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology, 3.

Paltoglou, G., & Thelwall, M. 2012. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology, 3.

Khan, F. H., Bashir, S., & Qamar, U. 2014. TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 57: 245–257.

Jurek, A., Mulvenna, M.D. & Bi, Y (2015). Improved lexicon-based sentiment analysis for social media analytics. Secur Inform 4, 9.

Koto F., Adriani M. (2015) A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?. In: Biemann C., Handschuh S., Freitas A., Meziane F., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2015. Lecture Notes in Computer Science, vol 9103. Springer, Cham

Guha, S., Joshi, A. and Varma, V., 2015, June. Sentibase: Sentiment analysis in twitter on a budget. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 590-594).

[14] Y. Wu and F. Ren, “Learning sentimental influence in twitter,” in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

[15] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in Proceedings of LREC, vol. 2010, 2010.

[16] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[Dhaoui, C.](https://www.emerald.com/insight/search?q=Chedia%20Dhaoui), [Webster, C.M.](https://www.emerald.com/insight/search?q=Cynthia%20M.%20Webster) and [Tan, L.P.](https://www.emerald.com/insight/search?q=Lay%20Peng%20Tan) (2017), "Social media sentiment analysis: lexicon versus machine learning", [Journal of Consumer Marketing](https://www.emerald.com/insight/publication/issn/0736-3761), Vol. 34 No. 6, pp. 480-488.

Kruspe, A., Häberle, M., Kuhn, I. & Zhu, X. X. (2020). Cross-language sentiment analysis of European Twitter messages duringthe COVID-19 pandemic. arXiv preprint arXiv:2008.12172.

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

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R. & 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

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

Ahmed, W., Bath, P. A., Sbaf, L., & Demartini, G. (2019). Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data. Health Information & Libraries Journal, 36(1), 60–72.

Nikolovska, M., Johnson, S.D. & Ekblom, P. “Show this thread”: policing, disruption and mobilisation through Twitter. An analysis of UK law enforcement tweeting practices during the Covid-19 pandemic. Crime Sci 9, 20 (2020).

Minqing Hu and Bing Liu, ``Mining and summarizing customer reviews.'', Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, Aug 22-25, 2004.

Nielsen, F.. “A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs.” *#MSM* (2011).

Silge et al., (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. Journal of Open Source Software, 1(3), 37,

Supporting information

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