A Reproducible Twitter-based Visual Analytics for Monitoring Public Opinions on Policing –

A Systematic Assessment of Pandemic Referencing

**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: Visual Analytics, Pandemic, Twitter, Text Mining, Policing.

Abstract (1 paragraph)

This this this...

1. Introduction (1 – 1 ½ pages)

We simply want to answer the question: ‘Has the pandemic-related tweets significantly impacted the overall public opinions on the police or policing?”, or put in a different way, “Has the pandemic-related tweets significantly increase or decrease the opinion (scores) on policing”. Hence, our null hypothesis is that there is no significant difference the opinion scores calculated from the tweets containing policing-keywords only and tweets containing ‘policing- and pandemic-‘ keywords.

2. Literature Review (2 – 3 pages)

**3. Reproducible Visual Analytics (RVA) for Monitoring Public Opinions**

Our RVA is designed to allow the exploration, analysis and visualization of public opinions based on Twitter data about policing across England and Wales. We developed this application as open-source in order to order to facilitate easy adaptation for other study areas. The application combines four research tools or methods, namely; Twitter-API, spatial analytics, sentiment analysis (text mining), and visualization. These tools or methods are all implemented as open-source with source codes made publicly available for reproducibility. In the following sub-sections, we provide a detailed description of each component.

3.1. Twitter API and Spatial Analytics

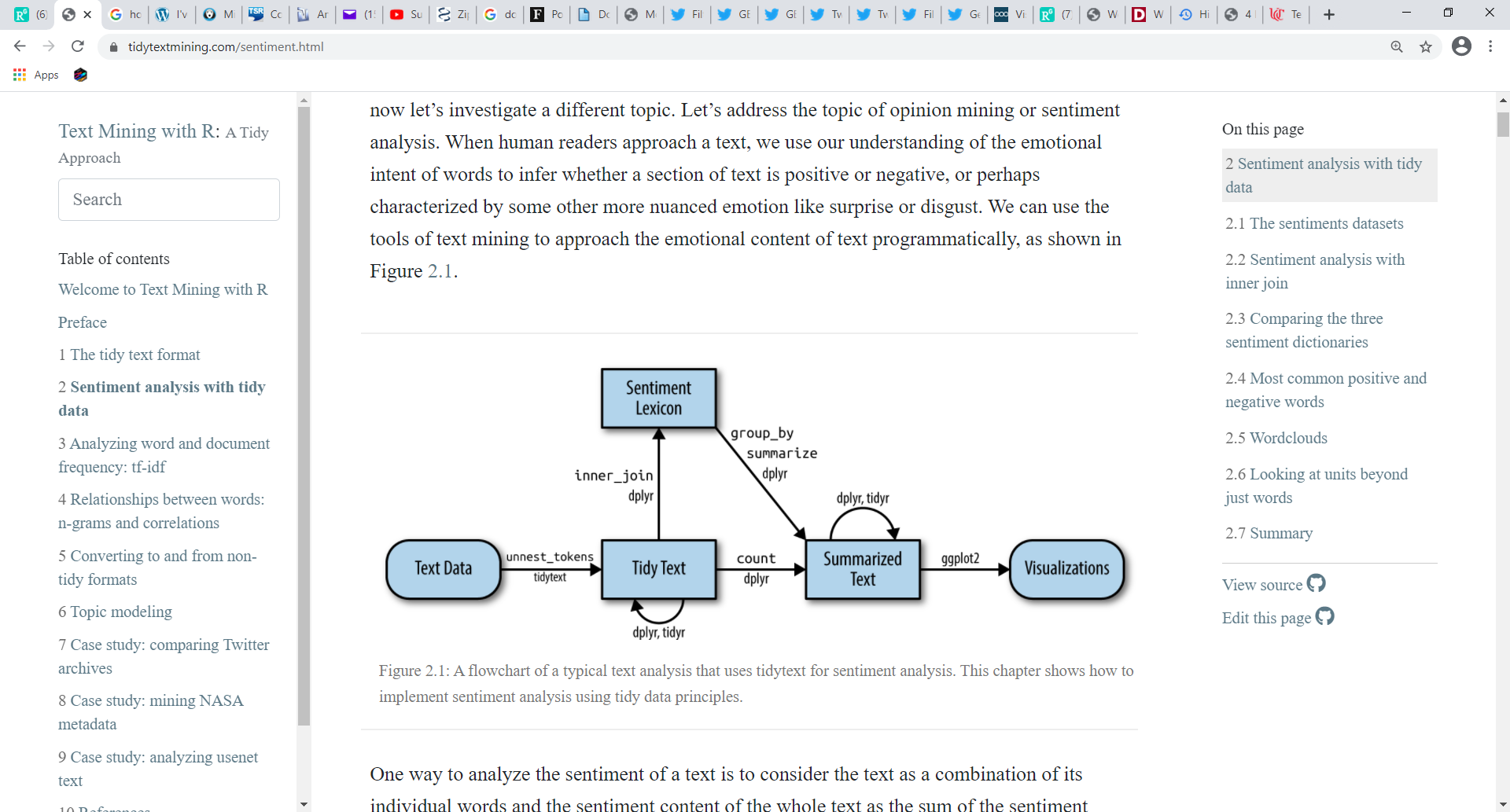
The Twitter API and spatial analytics are combined in order to procure the publicly available tweets for our research. The role of spatial analytics is to allow the integration of geographical specification for effective data download. The Twitter API is a programmable tool that provides access to the 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. The integration of geographical specification restricts the download 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 achieved by using the ‘*search\_tweets*()’ function of the ‘*rtweet*’ package in R.

The API searched for tweets that contain any of the specified keywords (or their hashtags) relating to the police or policing. These keywords are: ‘police’, ‘policing’, and ‘law enforcement(s)’. After the download, we geocoded each tweet to its respective spatial unit of analysis based on the user’s profile location. In this study, we use the operational units of the police forces in the UK, namely; the Police Force Areas, henceforth referred to as PFAs. The geocoding is completed by creating a ‘PFA-location-lookup’ table, which allow each tweet to be assigned to its respective PFA. Our ‘PFA-location-lookup’ table contains names of all cities, towns and villages across England and Wales. Created based on the UK ONS location gazette (ONS 2020), the lookup table contains a total of 35,604 names of locations.

3.2. Sentiment Analysis

The sentiment analysis is a text mining technique for computationally classifying opinions from a piece of text data into positive, negative or neutral, or some other more nuanced emotion like surprise, fear or disgust. Figure 1 shows the workflow for completing a sentiment analysis, with the annotation of the R packages (in double quotes) as well as the functions (in single quotes) that we employed in our own study. At the core of a sentiment analysis is the sentiment lexicon - a dictionary of words, which also shows their respective sentiment classification. Many spoken words express sentiment and can be collated together as a lexicons depending on the classification schemes. Most common lexicon is the uni-gram ‘BING’, which categorises a word into a positive, or a negative sentiment. By uni-gram we imply that the lexicon is based on a single word classification. In our study, we employ the AFINN lexicon, which provide a more nuance polarity classification by assigning a score indicating the degree of positivity or negativity of a word. The scores range from 5 (extremely positive) to -5 (extremely negative). The AFINN lexicon is used instead of the ‘BING’ lexicon in order to add more context to the tweet document classification. We understand that many tweets can contain multiple positive and/or negative words, all of which contribute to the overall opinion that a tweet expresses. It is then appropriate to classify a tweet as expressing either a positive or a negative sentiment by using the net sentiments across all sentiment words.

Another classification step that we took in order to ensure a more accurate scoring is to consider bi-grams classification (i.e. two words) in the cases in which a sentiment word is preceded by a negation word, such as ‘not’, ‘never’, ‘no’, or ‘without’. The score of such sentiment word is value in the opposite direction of the original word. For example, if the word ‘good’ which is scored as +3 is preceded by the negation word, such as ‘not’ (as in ‘not good’), then the sentiment of the word ‘good’ (in this type of scenario) becomes -3. Any tweet with zero a score of zero or that contains no sentiment words are considered neutral and removed from the tweet document.



## Fig. 1. Sentiment Analysis workflow (re-draw this figure: \*)

1. **Opinion Scores**

We define the opinion score of an area unit *i* for a given time window as the difference between the sum of all weighted positive tweets and the sum of all weighted negative tweets (Kuhn, M., 2008). This is given by the Equation 1:

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

Where, is the weight assigned to the tweets (e.g. based on the level of re-tweets or favorites), is the sum of positive sentiments and is the sum of negative sentiments. We ignore the weight i.e. in this study, for the lack of clear weighting function. Therefore, the final opinion score effectively becomes the difference between the total number positive and the negative tweets, following the AFINN classification. Given an opinion score (OP) of a spatial unit area *i*, we can state whether the public has a positive or a negative opinion about the subject matter. The public opinion in an area is positive if OP is the area has (+) sign, or negative if it has a (-) sign. We eliminate the neutral tweets from the tweet document.

1. **Document-Sentiment probabilities**

To assess the impacts of the pandemic-related tweets on overall opinion on policing, there is a need to estimate the exclusive probabilities of sentiments for the tweets that relate to policing only, as oppose to tweets that relate to both policing and pandemic. The two-by-two matrix showing the probabilities of positive and negative sentiments in a given document is the ‘Document-sentiment probabilities. By assigning the probabilities from the tweet document relating to policing only to the document relating to policing and pandemic, we can derive an expected sentiment document, that assumes zero impacts of the tweets that contain pandemic.

Fig. 2. Deriving the Expected-Sentiment Document

The expected sentiment document is crucial for calculating the statistical significance of the computed OP across different PFAs. We simply compare the OP score derived from the ‘observed-sentiment document (OSD) (i.e. the original datasets) with the expected-sentiment document (ESD)’ based on Figure 2. We identify the ‘policing-pandemic-related’ tweets from the original document by looking for tweets that contain the keywords or hashtags of words that references the covid-19 pandemic. These keywords include ‘*lockdown*’, ‘*corona*’, ‘*coronavirus*’, ‘*covid*’, ‘*covid-*19’, ‘*virus*’, and ‘*quarantine*’, ‘*infect*’, ‘*infection*’, etc.

**(c) Computing the P-Value of Opinion Scores**

The statistical significance (P-Value) of an observed OP score is required to make a judgment as to whether the observed OP score is beyond a change occurrence. To compute this 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 any area with scores higher than the observed scores *f*(O)?”. For each PFA, the randomization testing involves generating a large number of ESD, otherwise referred to as “replicas”, , and generate a distribution of expected opinion score . Given the *f*\*(E) distribution for a given PFA, the respective *p*-value can be 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 two critical regions + α (i.e. two-tailed), allowing us to concluded whether the pandemic-related tweets 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. However, since the run time is proportional to the number of replicas, a lower value (e.g. 99) are often recommended.

3.3 Visualization

We selected appropriate visualization tools based on the need to represent the two key aspects of our results, namely the temporal and the geospatial aspects (Kucher et al. 2018). The temporal aspects speak to the consecutive time periods at which we perform our analysis and how the opinions may change over time. Monitoring the changes could allow an analyst to better identify the factor that may be responsible for the observed change in public opinion. The geospatial aspect assumes that there are inherent spatial association between PFAs, particularly those found within the same policing region. For example, PFAs belonging to the same region may adopt similar operational policing tactics and therefore shape public opinion about policing similarly. These associations may manifest themselves in terms of similar patterns of opinion. Maps are best tools for revealing this type of associations (\*).

We selected line graphs, radar charts and spatial maps in order to fulfil our visualization objectives. The tools are simple and efficient for representing complex multi-dimensional information. The line graphs will allow us to visualize the changes in opinion from one time period to another. Both radar charts and geospatial maps are capable of representing multiple object attributes simultaneously, while also allowing any potential association to be visualized. Additionally,

3.4 Reproducibility of Research

The source code of our research has been included as supplemental material, including the set up of Twitter API for data procurement. Our analysis source code (in the R language) is available as supporting information and online at <https://github.com/MAnalytics/..> . The source code and required data are both open source, so the analysis presented here can be repeated in its entirety using the source file.

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.

Fig. 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. We carried out the data download twice a day (morning and night), extracting the tweets containing the specified police-related keywords within the last 7 days across the area. 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.

Fig. 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 divide the tweet document into three time period in order to monitor changes in opinion over time. We then perform the sentiment analysis and the associated statistical testing to identify the PFAs whose observation are unlikely to be due to chance. The time periods as defined as follow:

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

For each time period, we generate the OSDs for each PFA in order to allow the estimation of opinion scores across the study area. Thus, for each PFA, we have a time series (i.e. consisting of three data points) of observation. At the same time, we generate the ESDs for each PFA in order to allow the computation of corresponding statistical significance values. We use 999 replications in the randomization testing. Therefore, 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, we adopt the convention of 5% level, meaning each side of distribution is cut at 2.5% corresponding to a p-value of 0.025.

**5. Results**

ugugh

Fig. 5: Observed sentiment scores of PFAs for time period 1 to 3

[Explanation of results here]

Fig. 6. Breakdown of categories of tweets and their sentiment per PFA

Fig. 6: A bar of pie chart comparing sentiments percentages of sentiment classifications withrepresenting the sentiments classification of OSD, with inserted

It is obvious that the overall . Not only are thee negative more overall but .. in .



Fig. 6: Change in significance of observation

[Explanation of results here]

Fig. 7: Patterning and change in significance

[Explanation of results here]

**5. Discussion**

**6. Conclusion**

Acknowledgements

Supporting information

References

COMPUTER GRAPHICS forum

Volume 37 (2018), number 1 pp. 71–96

The State of the Art in Sentiment Visualization

COMPUTER GRAPHICS forum

Volume 37 (2018), number 1 pp. 71–96

The State of the Art in Sentiment Visualization

**Appendix**